Data quality in practice: experience from the frontline

IQ'96 Oct 25-26, 1996

Chris Firth, Citibank Singapore

Abstract

Information, stored in databases, is a key competitive advantage of many companies. However, this importance does not imply that managers will view data as a strategic resource or that they have the experience to interpret business processes in terms of data management.

One way to overcome this corporate inexperience is to establish a *data quality system* - a controlled and robust process that executes the measurement, analysis and improvement of data quality. Yet, often the most difficult problem in practice is to get a data quality system up and running, and accepted as worthwhile by senior management.

Based on my experience, practice and research into data quality, I recommmend four steps to initiate the implementation of a successful system: (i) establish data quality position, (ii) formulate a data quality policy, (iii) determine objectives and (iv) obtain management and employee commitment. For each of these steps I highlight ways to ensure the best chances to succesfully manage data quality.

1. Motivating Factors and Benefits

Businesses face many problems, some of which are caused, or at least, aggravated by poor data quality management. I briefly describe here the core motivating factors relating to data quality.

Reducing management failures: making the right decision

Since many management decisions are based on a quantitative analysis, from MIS, customer surveys, accounting systems or otherwise, it is indisputable that poor quality data can have a devastating consequence. The data could be incomplete, inconsistent or just plain wrong. For example, First Financial Management Corp.of Atlanta, US had to restate its earnings for the first nine months of 1991 after discovering that a subsidiary lost track of some records after changing its accounting systems.

Similarly, targetted marketing based on demographics can be inefficient or even counter-productive if the data underlying the strategy are faulty in some way. In one case, a large New York casino that routinely keeps in touch with its customers via direct mail, a process that depends on up-to-date, reliable mailing lists, estimated that only about 80% of the data in the customer list was accurate.

Thus, any improvement in the corporate MIS database will lead to more informed management decison making and strategy setting.

Raising product and service quality

Because product and service design, manufacturing and delivery are often based on an IT infrastructure of systems and data, a organisation's competitive position can be at risk if data quality is poor. Federal Express, for example, reportedly made significant improvements to the quality of data in its COSMOS database system, and gained a substantial market share and competitive advantage.

In 1995 the US Information Industry Association started development on a new 26-letter ID numbering system for stock exchange symbols. The scheme was an answer to the problem of multiple, sometimes inconsistent, identification codes from multiple international sources. In this example, incorrectly interpreting information from a stock exchange could effect the competitive position of an investor. The converse is also true: an exchange supplying uninterpretable or "error-inducing" data could contribute to impaired reputation of the exchange.

In another, more stark case, *Information Systems Management*, Spring, 1993, reported on an operational data entry error that led to a loss of $500 million for Salomon Brothers.

Less quantifiable, is the opportunity cost or loss of customer goodwill associated with every mistake or minor error due to poor-quality data. Examples of these "minor errors" are

- male customer addressed as "Mrs" because of an error in the gender code
- product delivered to an old address, because there is no data enrichment process

- wrong account credited because of a data input error, customer has cheque bounced because funds were insufficient
- cross-sell opportunity missed because trigger demograpgic data was wrong or missing

According to 1994 research by the Gartner Group, the justification for investing in integrated customer information lies in the improved data quality, customer service, and trend analysis. When launching a direct marketing campaign "companies utilizing a data warehouse to perform this task have reported an increase from a response rate of two to three percent to a rate of 20 to 30 percent." Restating from the data quality perspective, poor data quality could be responsible for your company under-achieving in direct marketing by a factor of ten. Note, however, that implementing a data warehouse in itself doesn't raise data quality[1].

Reducing risk

Companies can be sued or fined for wrong reporting or misconduct, the causes for which simply may be polluted databases. Organisations are also subject to specific data quality laws. In 1992, *Corporate Computing* cited inaccurate and incomplete data contained in consumer credit reports as one of the reasons in a lawsuit against TRW. The damage to corporate image or standing can also be affected by mistakes caused by data quality problems.

Many enterprises, particularly those extending credit, need to detect, assess and control fraud or credit exposure. Detecting fraud or bad credit is often a highly data-dependent activity: identifying spending patterns, totaling expourse to individuals at a customer level, identifying insolvent individuals or companies. All of these tasks are made harder with incomplete, inaccurate or missing data.

Reducing major clean-ups

It is not uncommon for corporations to lack a data quality policy or focus (although the authors know of organisations like Reuters, who at one time had a *Director of Data Quality Projects* and AT&T who, in various guises, had a dedicated data quality group). For many companies there is often a short period of intense activity on data clean-up or scrubbing, usually after several prominent quality issues have surfaced and impacted their business, to "fix" a particular problem. This activity is usually expensive and time consuming. Moreover, several months later a new ad-hoc project may be started to achieve something similar. This patchwork data quality approach, and the expense of rectifying the effects of data quality problems, can be substantial when viewed over a timeframe of 2-5 years.

2. A Data Quality System

It is a platitude to state that information is a key competitive advantage of companies. However, this self-evident truth does not imply that managers view data as a strategic resource or that they have the experience to interpret business processes in terms of data management.

This lack of experience can be diffused by the establishment of a *data quality system* - a controlled and robust process that executes the measurement, analysis and improvement of data quality. This kind of system or process is familiar to many managers of operations or manufacturing. A major benefit is that this approach tends to eliminate root causes, permanently fix problems and will be less expensive over the long run. The trade-offs are shown in Table 1.

	Ad-hoc approach	**Data Quality System**
Short term		
Cost	Medium	High
Improvement to data	High	Medium
Improvement to process	None	Medium
Long term		
Cost	High	Low
Improvement to data	Low	High
Improvement to process	None	High

Table 1: Costs of clean-ups v process change

[1] what it will possibly do is highlight what severe data quality problems you have

Another benefit of a data quality system is its transparency to senior management, in terms of the contribution it brings to an enterprise.

An important point to digest is that a data quality system is not a software tool, rather it is a management discipline. An effective system can be built without any expensive or specialized packages[2].

Steps to build a data quality system

I present here the basic steps needed to initiate a data quality system or to "get started". These represent the typical first steps, as well as the chief hurdles, to a successful data quality project, both from a cost-benefit angle and from an organizational inertia perspective. These steps are:

(i) establish data quality position
(ii) formulate a data quality policy
(iii) determine objectives
(iv) obtain management and employee commitment

There are different views on the execution sequence of the four steps. Many companies may decide to measure data quality levels first before investing time in any other activity, so at least there is some notion of what problems they face, how much they are costing and by how much they are likely to improve. Some corporations may take a top-down approach and first mandate a data quality (or perhaps, *information management*) policy. The latter would be more likely when information is the firm's product (as for a financial information feeder) or when senior management are already convinced that better data quality will lead to a major business benefit.

A logical sequence is first to understand the size and nature of the problem, then to draft a policy, set objectives and lastly present to senior managemnet for concurrence. Policy and objectives can be fine tuned after management feedback.

3. Establish current position

The single most important and revealing step is to measure exisiting quality levels and to quantify their associated costs. Because a company may only want an *indication* of the current position, sampling techniques could be allied with either a manual review, simple "health check" programs (written by the IT department) and some data collection on costs. Table 2 shows an example of a first cut indication.

Error Type	Level %
Overall	17.4
Duplicate records	3.4
Gender inconsistency	1.1
Name inconsistency	0.3
IC/passport error	1.7
Missing IC/passport number	4.9
DOB inconsistency	0.1
Missing DOB	7.8

Table 2: First Measure of Data Quality Levels by Type

This approach, that is, to defer the use of specialized data quality tools, has the advantage of giving quick and relatively inexpensive feedback to management.

During this step it may also be useful to review software quality processes, product quality assurance, data management procedures and other related company standards or procedures. This will give managers an idea of where they stand in terms of the maturity of data quality process management and how much they need to be improved.

Data collection and analysis

[2] in certain cases tools are a very good investment

Data collection involves the quantifying of the unit costs of quality failure, computation of the quality levels (failure rates) and for non-quantifiable costs, fixing a level of business criticality that the data quality aspect demands. With these numbers it will be possible to calculate an overall cost and impact of data quality problems. The data needed to calculate rates, costs etc will primarily belong to your company and hence should at the very least be available, even if they may be difficult to tease out or to convert to a useful unit of measure.

During the data analysis, it is best to use the concept of dimensions (problem types) discussed by many data quality researchers, so as to gain a top-level understanding of the nature of problems and likely root causes. A single gross error rate, or quality level, while perhaps useful to report to senior management will not be very helpful as a diagnostic tool.

At this stage it could well be worth doing some limited "benchmarking" of your data quality performance against other companies in your industry, so as to give a competitive context to your problems and allow you to understand the severity of your quality levels. Table 3 gives a basic benchmark (because the data is not industry specific it may not be so useful).

Inaccuracy	
US Average	4.3%
International Average	5.1%
Best	0.2%
Worst	38.0%
Duplicates	
US Average	6.6%
International Average	8.1%
Best	0.6%
Worst	22.0%

Table 3: Defect rates for data quality
Source: Innovative Systems, Inc.

Some advice on data collection is given:

- Many useful *rates, ratios* and *costs* may already known by the financial control or accounts department, for example, unit labour costs (useful in calculating cost of rework), unit computer processing costs, average cost of a sales call, avreage profit per customer, average revenue by product.
- External sources of information, such as vendor newsletters, user group conference proceedings, trade journals, may be useful in *comparing* data quality levels.
- Sources of qualitative information, such as employees, customers, industry analysts, consultants, distributors and agents may have a deep if not so precise knowledge of data quality *causes, effects,* and *relative importance.*

Identify current documented practices

The purpose of this step is to understand what is being done today in your organisation officially or unofficially. Much of this input will be illustrative of why there are data quality problems or will give ideas on where to make improvements. Even if a data quality system never existed in an integrated form before there will be some documentation relating to data quality, including

- software documentation (particularly on database schemas, data dictionaries, data entry systems)
- user procedures
- accounting rules and uses
- information security manuals amd policies
- complaint handling procedures

SWOT

As part of this exercise, management may decide to catalogue the SWOT (strength, weakness, opportunity and threat) to the business from the current data quality position. An example is shown below:

Strength • Service and Quality Culture • Innovative and receptive Management • Widespread usage of service indicators (akin to data quality measures)	**Weakness** • "Instant results" culture not supportive of data quality effort • Fragmented systems and databases work against high data quality • No data quality policy
Opportunity • Market differentation is increasingly service quality, heavily dependent on data quality • Sales and amrketing are increasingky data driven, hence data quality dependent • Existing inhouse expertise in data quality • Competition is assessed to be far behind	**Threat** • Short-termism rules out the upfront costs of data quality investment • Initiatives from competitors • Pressure to deliver new technology fast erodes software quality and hence data quality

Table 4: Example SWOT analysis

A qualitative analysis

You have read that informal sources of information, such as employees, may be a valuable aid in the establishment of your current position and practices. One way to structure this task is by means of a user survey or questionnaire. A survey can help to crystallise issues and also gives staff a channel for their feedback (the latter should raise data quality awareness and motivation among staff). Senior management is also likely to be galvanized by seeing in black and white some of the issues facing their staff.

What should be in a data quality survey? I suggest a template here:

Criticality	Rate your data elements according to their importance. For example: • Required: data is required for legal/other reasons • Important: data is needed to deliver product on time with superior service • Nice to have: data is never needed, but may be useful to exceed customer expectations • Not important: data does not increase customer satisfaction, is not needed for other reasons
Accuracy, Timeliness	Rate your data elements according to their accuracy or timeliness
When and Where	Consider situations to help you identify where and how data corruption occurs: (list common situations such as System Conversion/Integration, Inadequate testing, Inconsistent data across systems)
Data Enrichment	Rate the proposed new data elements in terms of usefulness (these are fields not currently captured)
Control Mechanisms	Give examples of the controls that can be put into place to improve data quality. Ask users to think about how better controls may reduce errors.
Reliability	Identify the system in which the "best" information (most reliable, most timely) is held.
General Issues	What is the major data quality problem you face today? Identify any issues related to specific processes, that may be related to the quality of data. What is the most important change you would like to see that could improve data quality?

Table 5: Template for a data quality survey

4. Formulate a data quality policy

A template data quality policy is given:

Management shall define for new or existing products, processes or services

- *data quality objectives, such as performance, costs and losses.*
- *data quality factors affecting market position*

Line managers shall implement, monitor, and manage controls over their data. These controls shall maintain appropriate data quality levels based upon the specific conformance criteria set by the business. A separate assurance unit should track the execution of the policy; the quality system should be internally audited and evaluated on a regular basis.

Such a data quality policy can be incorporated into an information security, data management, software engineering or service policy, or may be kept separate and hence more visible. Some of the aspects of a data quality policy that a company should consider are:

Organizations will need to identify or develop, and document data quality methods and procedures, which will allow line mangers to follow an agreed process. Effort needs to be placed on educating and training line staff, in addition to training quality "experts". Relevant international and national standards should be noted

In order to achieve accountability, a business must establish and document data ownership for all databases. Data owners would be responsible for the quality level of their database, among other administration tasks, such as approval of access to the data and data classification. Without ownership the statistics showing quality levels will become simply more wasted trees.

Firms may also consider appointing dedicated data quality champions, as *Computerworld* noted in Sept 4, 1995:

> Data stewards are becoming ever more important if not critical positions in the management of corporate data. Companies have come to recognize that data is a critical asset, resulting in the growth of sophisticated data access tools and data warehousing. Businesses also recognize that ensuring the quality of data is equally important. The data steward manages data assets to optimize their quality, utility, accessibility and reusability. A wide range of tasks is involved, including establishing naming standards, criteria for data quality and retention, security specifications and definitions of standard data entities and attributions. Ideally, a data steward is assigned to each major data subject area. A data steward should have a variety of technical, business and interpersonal skills.

5. Determine the objectives

Once the current position is established, a firm should set measurable data quality goals, either in terms of *percentages* levels or *monetary values*. For example, a firm could try to halve its overall data quality indicator from 5% to 3%, or it could try to reduce its data quality related losses from £2 million per annum to £500,000 per annum. Since it may be difficult to set a single data quality rate, one approach is to classify the criticality of data and set targets for each class. Table 5 demonstrates this concept.

Class of Data Elements	**Current Quality Level***	**Target Quality Level***
Critical (eg social security number)	97.6%	99.9%
Customer Service Related (eg address)	93.2%	98.0%
Marketing Related (eg residence type)	53.1%	80.0%

Table 5: Objectives by class of data
*defined as "complete and accurate"

Several factors will influence the setting of goals, particularly the overriding corporate priority assigned to objectives such as:

- *Cost reduction:* a prime goal may be to target quality problems that bring the greatest payback in terms of reduced cost. Since the frequency/level and cost of individual problem types/dimensions have been calculated, focussing on the key areas for improvement presents few problems.
- *Ease of improvement*: certain problems may be fixed quickly, for example, adding range checks on input screens. Other problems may require more involved and prolonged implementation, for example, the introduction of a data quality tool. Quick fixes have the advantage of helping to demonstrate to management the tangible value of the data quality project.
- *Competitive Factors:* another major goal will be to achieve corporate customer service or product quality goals, usually with reference to competitive position or industry norms. If the link between data quality and service/product quality is well understood, or at least can be estimated, quality levels can be set based on this relationship.
- *Regulation and risk:* if a certain class of problems is likely to expose the corporation to regulatory or serious financial risk, then obviously these data quality issues would be tackled first.
- *Differing business need:* an enterprise may run several business lines and these may be more or less sensitive to data quality levels. The business with higher dependence on accurate data would be the obvious consideration for constructing the objective.

During the implementation stages, these goals should be tracked closely, so that progress is known, and focus is not lost on the *rasion d'etre* of the data quality system.

6. Obtain management and employee commitment

The success of any project depends on the backing of senior business management and employee buy-in. Often senior managers will not have a ready appreciation of data quality, either because data quality is too "technical", too abstract or has not been publicized enough in business magazines and literature. However, because poor data quality can be translated into losses (monetary, goodwill, sales opportunity) it should be relatively easy to persuade business managers that some action should be taken.

Likewise, cast in terms of improved customer service, raised morale, and competitive advantage, it should be easy to sell data quality to employees. This buy-in is enhanced, and the data quality system much improved, if line staff are involved during the project.

The most persusive statistics are the outputs of the step to establish the current data quality position since these will quantify the effects of doing nothing. This can highlight, in tangible and easily understandable terms, the effect of data quality on that corporation.

7. Conclusions

Based on my experience, practice and research into data quality, I have recommmended four steps to initiate the implementation of a successful data quality effort: (i) establish data quality position, (ii) formulate a data quality policy, (iii) determine objectives and (iv) obtain management and employee commitment. For each of these steps I have highlighted some of the stumbling blocks and success factors that could make or break a data quality project. Establishing a *data quality system* - a process that executes the measurement, analysis and improvement of data quality - is a leap forwards for companies grappling with poor data management. The preparatory activites described here make that task easier.