Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management

Received (in revised form): 11th October, 2004

Richard Marsh

founded Datanomic in 2001. Datanomic specialises in end-to-end data quality management and information assurance. It delivers the world's only integrated data quality software system to combine data audit, cleaning, error prevention and compliance. It solves data quality and business process problems for its rapidly growing European blue-chip client base across all sectors including telecoms finance, utilities and engineering. Datanomic's clients include AMEC Oil and Gas, Powergen, Alliance & Leicester and COLT Telecom. Dr Marsh holds a degree in engineering, a diploma in computer science, and a doctorate in knowledge capture and information structure from Cambridge University.

Abstract Information technology (IT) has become all-pervasive. In business, IT systems collect and authenticate data, process payments, allow access and ensure the accurate and timely delivery of stock. Today, systems share and exchange data 24 hours per day, seven days per week.

In any system, there has always been 'dirty' — or erroneous — data. Today, however, the effects of 'wrong' data are much more visible and the consequences more serious. Data quality management, meanwhile, has historically been treated as a relatively low priority activity — one that is often adversely affected by budget cuts and looming deadlines.

There are, however, early signs that the traditional inertia to data quality management activities is starting to change. This paper sets out a new methodology for data quality management that encourages data to be seen and managed as a corporate resource. At the heart of the methodology is the premise that data quality must be an ongoing, active and preventative process — not just a retrospective corrective activity.

INTRODUCTION

In every walk of life, from business to leisure, IT has become all-pervasive. IT systems collect and authenticate data, process payments, allow access and ensure stock is delivered on time — in the right quantities to the right place at the right time. Until relatively recently, most IT systems were discrete 'silos' operating in relative isolation to address specific business problems. Today, systems share and exchange data to form a bigger interconnected IT landscape; one that operates 24 hours per day, seven days per week. The expectations of customers and IT users alike have increased and today there is a general expectation of up-to-date information in real time, all the time.

Alongside this rapid growth and raised expectations has been a growth in 'dirty' or erroneous data. Data has always been 'wrong', but now the effects of it are much more visible and the consequences more serious. Dirty data — inaccurate, incomplete and inconsistent data — can

Richard Marsh
Datanomic, Jeffreys Building, St John's Innovation Centre, Cowley Road, Cambridge, CB4 OWS, UK.

Tel: +44(0) 1223 421 630; e-mail: richard.marsh@datanomic.com website: www.datanomic.com

no longer be treated as a relatively benign issue. Dirty data can and will have a direct impact on stocking levels, sales orders, customer perceptions, loyalty and profitability. Business data — *business information* — are key business assets that need to be managed actively to ensure that they are accurate, current and fit for purpose.

Data quality management has historically been treated as a desirable, but relatively low priority activity. Consequently, when budgets and timescales become tight, as they invariably do, it is one of the first worthy activities to be left for another time. In addition, even when initiatives are undertaken — often in response to a crisis — they are commonly *ad hoc* in nature rather than part of a systematic solution.

There are early signs that the traditional inertia to data quality management activities is starting to change. Several leading companies have launched top-down, long-term initiatives for data quality management. In effect, they are recognising the value of the *data*, as well as the systems, to the business and putting in place systems and processes to protect their investment.

This paper presents a new approach to data quality management that allows organisations to model and manage data as a strategic corporate resource. The rationale is based on the premise that total data quality management demands a holistic, integrated approach based on four core elements:

— Audit
— Clean
— Prevention
— Compliance.

At the heart of the methodology is the premise that data quality must be an ongoing, active, preventative process and not just a retrospective corrective activity.

DATA DAMAGE: THE FACTS

Research and reports by industry experts, including Gartner Group,[1] PriceWaterhouseCoopers[2] and The Data Warehousing Institute[3] clearly identify a crisis in data quality management and a reluctance among senior decision makers to do enough about it. The facts speak for themselves.

— 88 per cent of all data integration projects either fail completely or significantly over-run their budgets
— 75 per cent of organisations have identified costs stemming from dirty data
— 33 per cent of organisations have delayed or cancelled new IT systems because of poor data
— $611bn per year is lost in the US in poorly targeted mailings and staff overheads alone
— According to Gartner, bad data is the number one cause of CRM system failure
— Less than 50 per cent of companies claim to be very confident in the quality of their data
— Business intelligence (BI) projects often fail due to dirty data, so it is imperative that BI-based business decisions are based on clean data
— Only 15 per cent of companies are very confident in the quality of external data supplied to them
— Customer data typically degenerates at 2 per cent per month or 25 per cent annually
— Organisations typically overestimate the quality of their data and underestimate the cost of errors
— Business processes, customer expectations, source systems and compliance rules are constantly

changing. Data quality management systems must reflect this
— Vast amounts of time and money are spent on custom coding and traditional methods — usually firefighting to dampen an immediate crisis rather than dealing with the long-term problem

In the information economy the true measure of value is the ability to manage knowledge, intelligence and data effectively across the enterprise. In every sector, data quality determines which organisations are best placed to service the needs of their customers and hence are fit to survive and prosper. Those with inadequate and poorly managed processes will never realise the potential of one of their most important assets — their data. This will impinge directly on their turnover, profit and growth. Ultimately, in a competitive, global economy, their existence may be threatened.

The situation is becoming more pronounced. Globalisation, mergers, acquisitions and the universal dependence on IT are fuelling an exponential growth and exchange of variant information across disparate hardware and software systems. This creates an environment in which data can fall below standards that are required by the business and its systems.

Large, high-profile IT project failures and cost or time over-runs are nothing new. They are as old as the IT business itself. However, as IT engineering develops and matures, the nature and cause of the over-runs has changed. Historically, over-runs were associated with the business of producing the hardware and software of the system itself. Today, they are most commonly associated with in-house systems replacement by standard, configurable packages, and the stumbling block for these projects is often the data migration and gradual or big-bang cut-over. According to Bloor Research,[4] many major projects fail simply because developers give insufficient time and effort to understanding fundamental data requirements.

In addition to the data quality challenges germane to interlinked systems, corporate competence and trust are under intense scrutiny by critical stakeholders, shareholders, customers and regulatory bodies. If the accuracy and security of a company's data are in doubt, stakeholders will lose confidence, trading will be affected and performance will suffer. If organisations don't have a firm grip on their own data, how can they have a grip on the business?

In the past, there has been a tendency to make the false assumption that data can be treated as fixed, static and finite — as if once recorded, nothing much changes. Nothing could be further from the truth. Whether the data records pertain to people, inventory or abstract constructs, most data entities are subject to change over time. Live data need constant care to maintain the highest levels of accuracy, consistency and completeness and thus to meet any company's strategic goals. Furthermore, it is commonly presumed that new data entered onto a system is error-free. This too is seldom the case, as users of a system work around the constraints it imposes, overload the meaning of data fields to encode additional meaning and simply enter poor or erroneous data in error or ignorance.

DIRTY DATA: THE PITFALLS

As one of an organisation's most valuable assets, all corporate data should be accurate, consistent, complete, up-to-date and readily accessible. Anything else — including incorrect, irrelevant, incomplete, ambiguous and inconsistent information — distorts the true picture

and may lead to bad and costly business decisions.

It is common practice in business to tolerate dirty data to a substantial degree rather than to manage or eliminate it. When this is the case, dirty data proliferates across systems and the discrepancies multiply. In the long term, this undermines the business from within. Unless additional business processes are implemented to accommodate the possibility of bad data, business decisions are taken under false premise. Common symptoms include:

— Cumulative increases in costs and a drain on the bottom line;
— Rising costs due to down time to reconcile data;
— Risk of using inaccurate data to inform policy and decisions;
— Diversion of resources from mission-critical areas;
— Erosion of trust and credibility with stakeholders;
— Delays in new system deployment;
— Inability to comply with industry and quality standards;
— Increasing focus on internal issues allows competitors to gain ground;
— Demoralised teams: if staff are hampered by poor data, they will lose impetus;
— Unproductive, frustrating environment will drive talent elsewhere;
— Reduced ability to respond will affect customer service and slow growth; and
— Poor performance will jeopardise reputation and damage the brand.

The problem is not merely one of recognising that the data quality is poor and taking appropriate remedial action to fix it. This in itself would be a tractable, if potentially costly, operation. Often the biggest single problem is that the business is not even aware that it has poor data. According to Bloor Research, a major problem is identifying what you don't know, particularly where surprises, which are often undocumented, lurk in legacy systems, the source code has been lost and the person who originally wrote it has long since retired.

THE CASE FOR CHANGE

Given the extent of the problems and recognition of the potential impacts on business, one must first ask why data quality management hasn't been taken more seriously in the past. Put simply, there are two main reasons: first, a lack of corporate commitment or political will to tackle a problem that is unquantified and most usually crosses business unit boundaries; and secondly, the absence of effective technology-based solutions to automate substantial portions of the remedial action that will be required.

Champions for change

Data quality should be at the heart of the board's agenda, but it has traditionally been regarded as the responsibility of the IT department. Following periods of rapid change, where companies with disparate IT systems, conventions and standards have merged, fragmented and regrouped, the subsequent need to manage and interpret a disparate mass of data can raise divisive cross-departmental political and cultural conflicts that only intervention by senior management can resolve.

All the evidence points to the need for data quality to be championed as a strategic, long-term initiative from the top of the business down. This is the only way to establish and maintain consistent standards across the business and the underlying data. Improving data quality does not simply mean fixing the erroneous data. It also involves exposing sloppy processes, changing user

behaviour, securing broad commitment to new practices and resourcing the whole programme with time, support, education and training.

A complete, intelligent solution

Historically, there has not been a simple, systematic technology solution for managing the complete data quality cycle. A comprehensive data quality programme required multiple tools plus considerable scripting, manual checking and, often, custom development. The few tools that were available were expensive, labour-intensive (and hence extremely error prone themselves) limited in what they could do and prescriptive about how it should be done. This meant that you could invest in a tailored solution to audit your data, for example, but would then need a different type of solution to fix the problems and clean the data that the audit had unearthed. Organisations generally had little choice but to pay a high price for custom-coding and tailored solutions employing a high degree of human intervention. This approach requires a large team, manual rekeying and extreme vigilance to achieve an adequate and appropriate successful data fix rate. Undertaken in this manner, the whole exercise only amounts to managing data quality in a piecemeal fashion. At best this route delivers a short-term and relatively limited benefit and is not functionally expandable across the enterprise to other systems and data.

Most data quality software solutions have their heritage in the direct mail marketing domain. Consequently, the commercial solutions available only deal with basic name- and address-type customer data, reflecting traditional mailing house requirements. The main driver in this market is, and has always been, to eliminate duplicate names and addresses and those that are no longer valid. The business imperative and case are simple to make: reduced mailing costs and fewer aggrieved customers. However, accurate data is vital in a wide variety of modern IT environments from customer relationship management (CRM), inventory, BI and entity resource planning (ERP) to data integration and migration tasks. Moreover, the nature of the data in such systems goes way beyond names and addresses to cover every type of real world and abstract entity imaginable. To support this new, wider and more significant challenge, a powerful generic solution designed for complete quality control of any type of data, from any source, in any format is needed. For customer data this includes account numbers, organisations, personal details etc. In the wider sphere it can extend to products, networks, catalogues, tariffs, operations, maintenance, product and raw materials inventory, asset registers and anything else that might be relevant.

REDEFINE YOUR DATA

Data quality management ensures that the total body of information needed to support the day-to-day and long-term business strategy is of the highest quality. It should embody the following key attributes:

— Accuracy: is the information correct, objective and can it be validated?
— Integrity: does it have a coherent, logical structure?
— Consistency: is the data consistent and easily understood?
— Completeness: does it provide all the information required?
— Validity: is it within the acceptable parameters needed by the business?
— Timeliness: is it available whenever required?

— Accessibility: can the data be easily accessed and exported to other applications?
— Compliance: does it comply with regulatory and industry standards?

THE FOUR-STAGE METHODOLOGY

The most effective way to ensure long-term data quality is via a continual, integrated process of proactive data management. Retrospective, corrective actions alone can never meet the above requirements of business data and its associated key attributes. The following methodology for data quality management is comprised of four separate but related elements — audit, clean, error prevention and compliance. The four-stage methodology is a systematic approach that makes it easy for any organisation to implement a total, integrated data quality management system at a pace that suits the structure and culture of the business. It provides a flexible framework to manage data quality as an enterprise-wide strategic resource and to address tactical, project-based data quality issues as they arise.

Step one — audit: the data healthcheck

The starting point for dealing with data quality is to gain a clear overview of the data and see where discrepancies lie. Data profiling is the initial step which paves the way for an in-depth data audit to identify common defects and create rules for fixing the data. The audit should be a systematic review to detect missing data, incorrect values, duplicate records, inconsistencies and any type of inaccuracy.

The results of a data audit can be alarming. It is not uncommon to find significant (double digit) percentages of fields missing, fields failing to meet standard tests of reasonableness, or potential duplicate fields. This only goes to reinforce the value in making it easy to benchmark data quality, good or otherwise. It is essential to help one understand the true state and nature of the data — possibly for the first time. Only once the scale and shape of the problem can be seen, can a strategy for putting it right and preventing the situation recurring be planned.

Initiating this first step demonstrates the early benefits of a proactive approach, which also helps to win over any sceptics so that the next stage can be tackled with confidence.

Deliverables

— Clear visibility of data to assess the nature and extent of problems.
— Create tailored reports for easy analysis and interpretation.
— Ensure, on an ongoing basis, that high quality data is available for key applications and processes.
— Remove uncertainty and project risk.
— Create cost–benefit roadmap of priorities.
— Immediate results help to gain buy-in from stakeholders.

Step two — clean: the data detox

Having identified the source and type of problems, the next logical step is to clean the data to remove errors, deal with inconsistencies and fix problems. Cleansing is also often concerned with intelligent matching of incorrect data across different sources.

With the latest generation of tools, the data cleanse process can take a very short time — days compared with months for manual or customised methods. Thereafter, data can be exported to any target application or data warehouse for storage in a pure, readily accessible form

for future use. Data cleansing can also routinely be applied between applications to ensure that stored data is maintained in prime condition.

Deliverables

— Correct and reformat any data to any specification.
— Output clean data in a standard, consistent format.
— Automated high throughput solution.
— Matching of disparate or incomplete data.
— Eliminate need for extensive custom-coding and piecemeal manual rekeying.

Step three — error prevention: keep it clean

Auditing and cleansing data is a massive step forward but, for enterprise applications, it is essential to prevent new defects entering the system. Real time error prevention is the most effective way of overcoming the stop–start reactive approach to data quality management.

In normal circumstances, customer data degenerates at a rate of approximately 2 per cent a month — almost 25 per cent a year. Furthermore, a major source of new defects is human error when new data are entered into the system. Another key problem area is corruption or incompatibility of new data imported into the corporate network.

The latest data quality solutions embed real time error prevention software and the data rules that have been determined from the audit and clean stages into operational processes via a data quality server.

Deliverables

— Guarantee a consistently high level of clean data at all times.
— Ensure confidence and user uptake.
— Apply consistent and updateable validation rules without coding.
— Combine with staff education and training to reduce input errors.

Step four — compliance: fit for the future

Like any regime, proactive data quality management yields maximum benefit as an integral part of day-to-day business processes. With compliance capability it is much easier to deal with problems in real time as they arise, rather than tackling issues *ad hoc* or once they've reached crisis point. Compliance with the business data rules that are required for the company's data can be monitored and reported on continuously, alerting one to potential problems in a timely manner and providing re-assurance via metrics of the state of the business data.

Establishing clear compliance criteria to continuously monitor, measure and manage data quality — whether in line with regulatory requirements, international standards, service level agreements, key performance indicators or key data elements — makes it easy to maintain the long-term benefits of high quality data.

Deliverables

— Ensure measurability against strategic data quality initiatives.
— Guarantee long-term compliance with industry standards.
— Continually monitor and measure data quality against agreed metrics.
— Ensure business processes run smoothly, efficiently and effectively.

BENEFITS OF THE FOUR-STAGE PROGRAMME

From the very first step, the benefits of proactive data quality management are

tangible and immediate. Left unchecked, dirty data will drive a downward spiral of decline, while the reverse is true when a coherent programme is initiated. The cumulative effect is a positive chain of benefits that impact across the enterprise, resulting in:

— Reductions in wastage, costs and unexpected expenditure;
— A clear measurable view of the value of data;
— Increased customer satisfaction;
— An ability to form a clear, objective analysis of the business at any time;
— Increased confidence that the business can deliver its strategic goals;
— An ability to focus resource on proactively building the business rather than firefighting;
— Increase credibility and trust among stakeholders, staff and customers;
— Accountability and compliance with standards;
— Optimum working conditions for a productive, motivated workforce;
— Added value to the company reputation, brand and business; and
— An opportunity to exploit the competitive advantage gained over rivals.

SUMMARY

Combining all elements of the four-stage methodology guarantees that every stage of the data quality lifecycle runs in tandem with the strategic needs of the organisation. By adopting the four-stage approach to data quality, organisations can start to reap the true benefits of their data as a valuable information asset and profitable resource that adds serious value to every aspect of the business.

The four stage methodology provides a framework for addressing all aspects of data quality management. There are data quality management products that can be used to automate large portions of the four-stage methodology, allowing one to take the output of one stage and to use it to drive the subsequent stages. Data quality management has matured beyond the narrow confines of achieving cost reductions in direct mail marketing. With modern software tools, getting to grips with data quality across the whole enterprise is easier today than it has ever been. Are you going to take the plunge or wait until you're up to your neck in it? Sink or swim — the choice is clear.

References

1 Gartner Dataquest, various papers on data quality by Ted Friedman, 2001 onwards.
2 PriceWaterhouseCoopers (2002) 'Global Data Management Survey', Citreon Wolf Communications, London.
3 Eckerson, W. W. (2002) 'Data Quality and the Bottom Line', The Data Warehousing Institute, 101 Communications LLC, Chatsworth, CA.
4 Howard, P. (2003) 'Bloor Research Report', Bloor Research, Wiltshire, UK.
5 *Ibid*.