# DATA QUALITY STRATEGY: A STEP-BY-STEP APPROACH

(Practice-oriented Paper)

**Frank Dravis**
Vice President of Information Quality, Firstlogic, Inc.

## INTRODUCTION: WHY HAVE A DATA QUALITY STRATEGY?

Coursing through the electronic "veins" of organizations around the globe are critical pieces of information – whether they be about customers, products, inventories and transactions. While the vast majority of enterprises spend months and even years determining which computer hardware, networking, and enterprise software solutions will help them grow their business, few pay attention to the data that will support their investments in these systems. In fact, Gartner, Inc. contends, "By 2005, Fortune 1000 enterprises will lose more money in operational inefficiency due to data quality issues than they will spend on data warehouse and CRM initiatives (0.9 probability)." (Gartner Inc. T. Friedman April 2004).

*DM Review* in their 2002 readership survey conducted by the Gantry Group LLC, asked:

> "What are the three biggest challenges of implementing a BI/DW project within your organization?"

Of the 688 people who responded, the number one answer (35 percent of the respondents), was budget constraints. Tied with budget constraints, the *other* number one answer was data quality. An equal number of respondents (35 percent) cited data quality as more important than budget constraints.

Put simply, to realize the full benefits of their investments in enterprise computing systems, organizations must have a detailed understanding of the quality of their data — how to clean it, and how to *keep* it clean. And those organizations that approach this issue strategically are those that will be successful. But what goes into a data quality strategy? That is the purpose of this paper, to explore strategy in the context of data quality.

## WHAT IS STRATEGY?

Many definitions of strategy can be found in management literature. Most fall into one of four categories centered on planning, positioning, evolution, and viewpoint. There are even different schools of thought on how to categorize strategy, where one model describes corporate strategies, competitive strategies, and growth strategies. Rather than pick any one in particular, claiming it to be the right one, we will avoid the debate of which definition is best, and pick the one that fits the management of data. This is not to say other definitions such as, "A business strategy is a set of dynamic, integrated decisions which you must make in order to position your business in a complex environment," do not fit data. However, the definition this paper will use is, "Strategy is the implementation of a series of tactical steps." More specifically, the definition used in this paper is:

> "Strategy is a cluster of decisions centered on goals that determine what actions to take and how to apply resources."

Certainly a cluster of decisions – in this case concerning six specific factors – will need to be made to effectively improve the data. Corporate goals will determine how the data is used and the level of quality needed. Actions are the processes improved and invoked to manage the data. Resources are the people, systems, financing, and data itself. We can therefore apply the selected definition in the context of data, and arrive at the definition of data quality strategy:

> "A cluster of decisions centered on organizational data quality goals that determine the data processes to improve, solutions to implement, and people to engage."

## EXECUTIVE SUMMARY

This paper will discuss:

- Goals that drive a data quality strategy
- Six factors that should be considered when building a strategy
- Decisions within each factor
- Actions stemming from those decisions
- Resources affected by the decisions and needed to support the actions

You will see how these six factors — when added together in different combinations — provide the answer as to how people, process and technology are the integral and fundamental elements of information quality.

To help business or IT managers develop a data quality strategy, we provide as an appendix an outline of a strategy. We conclude the paper with a discussion on the transition from data quality strategy development to implementation via data quality project management.

## GOALS OF DATA QUALITY

Goals drive strategy. Data quality goals must support on-going functional operations, data management processes, or other initiatives such as the implementation of a new data warehouse (DW), CRM application, or loan processing system. Contained within these initiatives are specific *operational* goals, for example:

- Reduce time to process quarterly customer updates
- Cleanse and combine 295 source systems into one master customer information file
- Comply with USA Patriot Act and other governmental or regulatory requirements to identify customers
- Determine if a vendor data file is fit for loading into an ERP system

An enterprise-level initiative in itself is driven by strategic goals of the organization. For example, a strategic goal to increase revenue by 5 percent through cross-selling and up-selling to current customers would drive the initiative to cleanse and combine 295 source systems into one master customer information file. The link between the goal and the initiative is a single view of the customer, versus 295 separate views. This would allow you to have a complete profile of the customer and identify opportunities otherwise unseen. At first inspection, strategic goals can be so high level, and seem to provide little immediate support for data quality. Eventually, however, strategic goals are achieved by enterprise initiatives that create demands on information in the form data quality goals.

For example, a non-profit organization established the objective of supporting a larger number of orphaned children; and to do so, needed to increase donations. This would be considered a strategic goal for the charity. The charity, in this example, determined that to increase donations they needed to identify their top donors. A look at the donor files caused immediate concern. There were numerous duplicates, first names were missing, addresses were incomplete, and a less-than rigorous segmentation between donor and prospect files led to overlap between the two groups. In short, the organization could not reliably identify who their top donors were. At this point, the data quality goals became apparent: a) cleanse and standardize both donor and prospect files, b) find all duplicates in both files and consolidate the duplicates into "best-of" records, and c) find all duplicates across the donor and prospect files, and move prospects to the prospects file, and donors to the donors file.

As this example illustrates, every strategic goal of an organization is eventually supported by data. The ability of an organization to attain those goals will, in part, be determined by the level of quality of the data it collects, stores, and manages on a daily basis.

## THE SIX FACTORS OF DATA QUALITY

When creating a data quality strategy there are six factors, or aspects of an organization's operations that must be considered. Those six factors include:

- Context — the type of data being cleansed and the purposes for which it is used
- Storage — where the data resides
- Data Flow — how the data enters and moves through the organization
- Work Flow — how work activities interact with and use the data
- Stewardship— people responsible for managing the data
- Continuous Monitoring — processes for regularly validating the data

Figure 1 depicts the six factors centered on the goals of a data quality initiative, and shows that each factor requires decisions to be made, actions that need to be carried, and resources to be allocated.

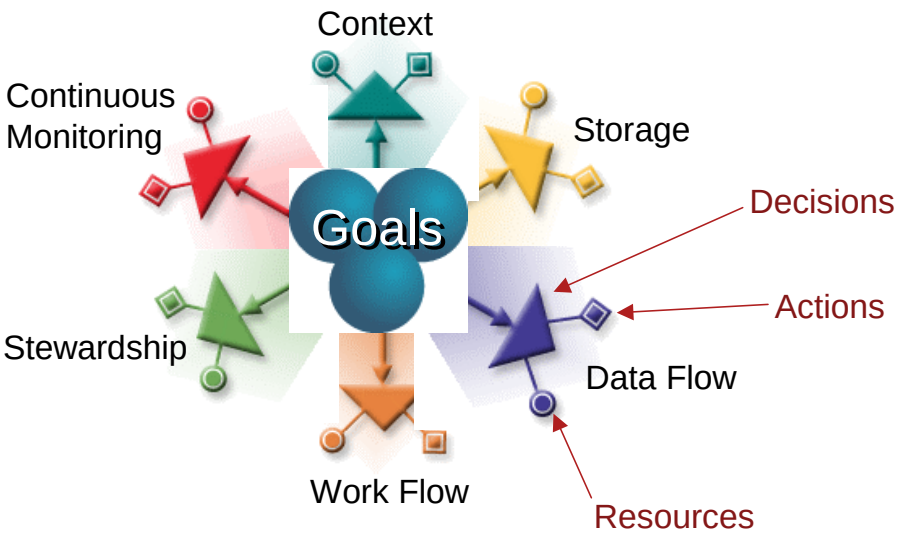



***Figure 1: Data quality factors***

Each factor is an element of the operational data environment. It can also be considered as a "view" or "perspective" of that environment. A factor, in this representation, is a collection of

decisions, actions, and resources centered on an element of the operational data environment. The arrows extending from the core goals of the initiative depict the connection between goals and factors, and illustrates that goals determine how each factor will be considered.

## *Context*

Context defines the type of data and how the data is used. Ultimately, the context of the data determines the necessary types of cleansing algorithms and functions needed to raise the level of quality. Examples of context and the types of data found in each context are:

- Customer data — names, addresses, phone numbers, social security numbers, etc.
- Financial data — dates, loan values, balances, titles, account numbers, types of account (revocable or joint trusts, etc.)
- Supply chain data — part numbers, descriptions, quantities, supplier codes, etc.
- Telemetry data — height, speed, direction, time, measurement type, etc.

Contexts can be matched against their appropriate types of cleansing algorithms. Take customer name as an example. A subset of a customer name is title. In the customer name column, embedded with the first name or last name or by itself are a variety of titles, VP, President, Pres, Gneral Manager, and Shoe Shiner. It takes a specialized data cleansing algorithm to "know" the complete domain set of values for title, then to be configurable for the valid domain range which will be a subset. A title cleansing function may need to correct Gneral Manager to General Manager, to standardize Pres to President, and to either eliminate Shoe Shiner or flag the entire record as out of domain depending on the business rules.

## *Storage*

Every data quality strategy must consider where the data physically resides. Considering storage as a data quality factor ensures the physical storage medium is included in the overall strategy. System architecture issues such as whether data is distributed or centralized, homogenous or heterogeneous are important. If the data resides in an enterprise application, the type (CRM, ERP, DW, etc.), vendor, and platform of the application will dictate connectivity options to the data. Connectivity options between the data and data quality function generally fall into the following three categories:

- Data extraction
- Embedded procedures
- Integrated functionality

**Data extraction** occurs when the data is copied from the host system. It is then cleansed, typically in a batch operation, and then reloaded back into the host. Extraction is used for a variety of reasons, not the least of which is native, direct access to the host system is either impractical or impossible. An example is how an IT project manager may attempt to cleanse data in VSAM files on an overloaded mainframe, where the approval process to load a new application – a cleansing application in this case – on the mainframe takes two months, if approved at all. Extraction of the data from the VSAM files to an intermediate location for cleansing in this case is the only viable option. Extraction is also a preferable method if the data is being moved as part of a one-time legacy migration or a regular load process to a data warehouse.

**Embedded procedures** are the opposite of extractions. Here, data quality functions are embedded, perhaps compiled, into the host system. Custom-coded, stored procedure programming calls invoke the data quality functions, typically in a transactional manner. Embedded procedures are used when the strategy dictates the utmost customization, control, and tightest integration into the operational environment. A homegrown CRM system is a likely candidate for this type of connectivity.

**Integrated functionality** lies between data extraction and embedded procedures. Through the use of specialized, vendor-supplied *links*, data quality functions are integrated into enterprise information systems. A link allows for a quick, standard integration with seamless operation, and can function in either a transactional or batch mode. Owners of CRM, ERP or other enterprise application software packages often choose this type of connectivity option. Links, which are a specific technology deployment option, are discussed in additional detail in the Work flow factor. Deployment options are the technological solutions and alternatives that facilitate a chosen connectivity strategy.

Data model analysis or schema design review falls under the Storage factor as well. The existing data model must be assessed for its ability to support the project. Is the model scalable and extensible? What adjustments to the model are needed? For instance, field overuse is one common problem encountered in a data quality initiative that requires a model change. This can happen with personal names, for example, where pre-names (Mr., Mrs.), titles (president, director), and certifications (CPA, PhD) may need to be separated from the name field into their own fields for better customer identification.

## *Data flows*

Each of the six strategy factors builds a different view of the operational data environment. With Context (type of data) and Storage (physical location) identified, the next step in developing a data quality strategy is to focus on Data Flow (the movement of data).

Data does not stay in one place. Even with a central data warehouse, data moves in and out just like any other form of inventory. The migration of data can present a moving target for a data quality strategy. Hitting that target is simplified by mapping the data flow. Once mapped, staging areas provide a "freeze frame" of the moving target. A data flow will indicate where the data is manipulated, and if the usage of the data changes context. Certainly the storage location will change, but knowing the locations in advance will make the strategy more effective as the best location can be chosen given the specific goals. Work evaluating data flow will provide iterative refinement of the results compiled in both the storage and context factors.

Data flow is important because it depicts access options to the data, and catalogs the locations in a networked environment where the data is staged and manipulated. Data Flow answers the question: Within operational constraints, what are the opportunities to cleanse the data? In general, these opportunities fall into the following categories:

- Transactional updates
- Operational feeds
- Purchased data
- Legacy migration
- Regular maintenance

Figure 2 shows where these opportunities can occur in an information supply chain. In this case, a marketing lead generation work flow is used with its accompanying data flow. The five cleansing opportunities are discussed in the subsequent sections.

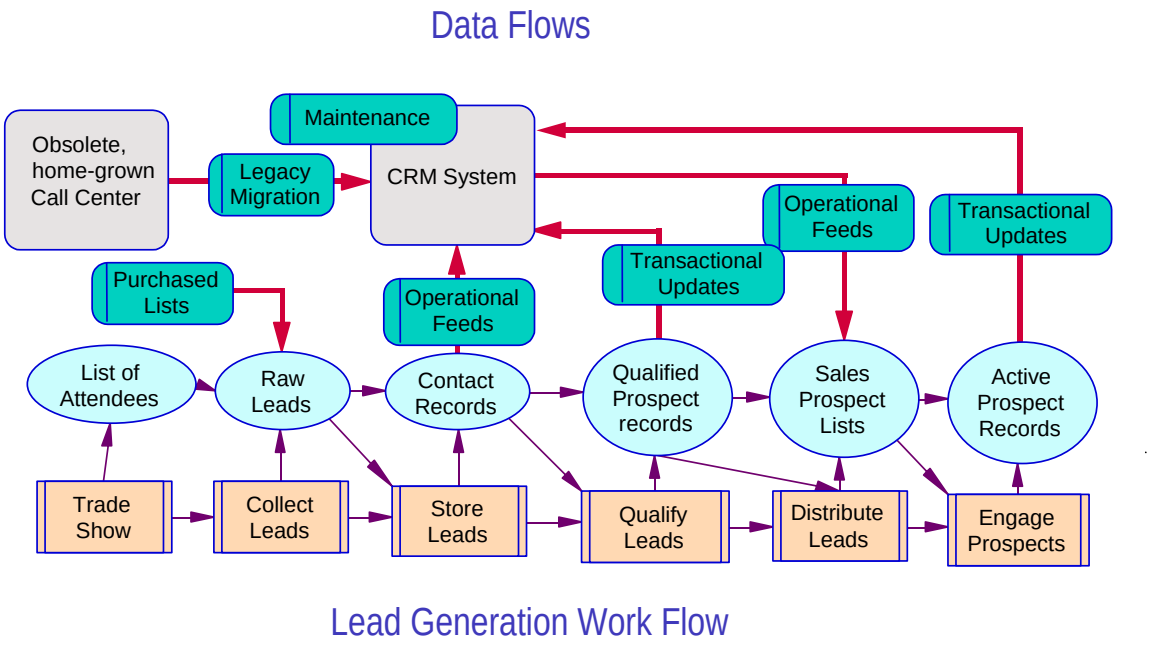



***Figure 2.***

**Transactional Updates**

An inherent value of the Data Flow factor is that it invites a proactive approach to data cleansing. The entry points — in this case, transactions — of information into the organization can be seen, depicting where the exposure to flawed data may occur. When a transaction is created or captured there is an opportunity to validate the individual data packet before it is saved to the operational data store (ODS). Transactional updates offer the chance to validate data as it is created or captured in a data packet, rich with contextual information. Any defects encountered can immediately be returned to the creator or originator for confirmation of change. This contextual setting is lost as the data moves further in the work flow and away from the point of entry.

The difference between a created and captured transaction is subtle, but important. A created transaction is one where the creator (owner of the data) directly enters the data into the electronic system as a transaction. A good example is a new subscriber to a magazine who logs onto the magazine's Web site and fills out an order for a subscription. The transaction is created, validated, and processed automatically without human intervention.

Alternatively, a captured transaction is where the bulk of data collection took place off-line and is later entered into the system by someone other than the owner of the data. A good example is a new car purchase where the buyer fills out multiple paper forms, and several downstream operators enter the information (i.e. registration, insurance, loan application, and vehicle configuration data) into separate systems. Created and captured data work flows are substantially different. The ability to correct the data with owner feedback is substantially easier and less complex at the point of creation, than in the steps removed during capture.

**Operational Feeds**

The second opportunity to cleanse data is operational feeds. These are regular, monthly, weekly, or nightly updates supplied from distributed sites to a central data store. A weekly upload from a subsidiary's CRM system to the corporation's data warehouse is an example. Regular operational feeds

collect the data into batches that allow implementation of scheduled batch-oriented data validation functions in the path of the data stream. Transactional updates, instead of being cleansed individually (which implies slower processing and wider implementation footprint), can be batched together if immediate feedback to the transaction originator is either not possible or not necessary. Transaction-oriented cleansing in this manner is implemented as an operational data feed. Essentially, transaction cleansing validates data entering an ODS, such as a back-end database for a Web site, whereas operational-feed validation cleanses data leaving an ODS, passing to the next system — typically a DW, ERP or CRM application.

**Purchased Data**

A third opportunity to cleanse is when the data is purchased. Purchased data is a special situation. Many organizations erroneously consider data to be clean when purchased. This is not necessarily the case. Data vendors suffer from the same aging, context-mismatch, field over use, and other issues from which all other organizations suffer. If a purchased list is not validated upon receipt, the purchasing organization is essentially abdicating their data quality standards to those of the vendor.

Validating purchased data extends beyond verifying that each column of data is correct. Validation must also match the purchased data against the existing data set. The merging of two clean data sets is ***not*** the equivalent of two clean rivers joining into one; rather it is like pouring a gallon of red paint into blue. In the case of a merge, 1+1 does not always equal 2, and may actually be "1.5" with the remainder being lost because of duplication. The merged data sets must be matched and consolidated as one new, entirely different set to ensure continuity. A hidden danger with purchased data is it enters the organization in an ad-hoc event, which implies no regular process exists to incorporate the data into the existing systems. The lack of established cleansing and matching processes written exclusively for the purchased data raises the possibility that cleansing will be overlooked.

**Legacy Migration**

A fourth opportunity to cleanse data is during a legacy migration. When data from an existing system is exported to a new system, old problems from the previous system can infect the new system unless robustly checked and validated. For example, a manufacturing company discovered during a data quality assessment that they had three types of addresses (site location, billing address, and corporate headquarters) but only one address record per account. In order to capture all three addresses their account staff was duplicating account records. To correct the problem, the account record structure model of the new target system was modified to hold three separate addresses, before the migration occurred. Account records duplicated because of different addresses could then be consolidated during the migration operation.

A question will often arise at this point: The account managers were well aware of what they were doing, but why was the duplication of accounts not taken into consideration during the early design of the target system? The answer lies in the people involved in the design of the new system – what users were interviewed, and how closely the existing work flow practices were observed. Both of these topics are covered in the Work Flow and Data Stewardship factors discussed later in this paper.

**Regular Maintenance**

The fifth and final opportunity to cleanse data is during regular maintenance. Even if a data set is perfectly defect-free today (highly unlikely), tomorrow it will be flawed. Data ages. For example, 17 percent of U.S. households move each year, and 60 percent of phone records change in some way each year. Moreover, every day we get married, divorced, have children, have birthdays, get new jobs, get promoted, and change titles. The companies we work for start up, go bankrupt, merge, acquire, rename, and spin-off. To account for this irrevocable aging process, organizations must implement regular data

cleansing processes – be it nightly, weekly or monthly. The longer the interval between regular cleansing activities, the lower the overall value of the data.

Regular maintenance planning is closely tied to the sixth strategy factor: Continuous Monitoring. Both require an organization to assess the volatility of its data, the frequency of user access, the schedule of operations that use the data, and the importance, and hence, minimum required level of quality for the data. Keeping all of this in mind, the organization can establish the periodicity of cleansing. The storage factor will have identified the location of the data and preferred connectivity option.

## *Work Flow*

Work flow is the sequence of physical tasks necessary to accomplish a given operation. In an auto factory, a work flow can be seen as a car moving along an assembly line, each workstation responsible for a specific set of assembly tasks. In an IT or business environment, the work flow is no less discrete, just less visually stimulating. When an account manager places a service call to a client, the account manager is performing a work flow task in the same process-oriented fashion as an engine bolted into a car.

Figure 3 shows a work flow for a lead generation function where a prospect visits a booth at a tradeshow and supplies their contact information to the booth personnel. From there the workflow takes over and collects, enters, qualifies, matches, consolidates, and distributes the lead to the appropriate sales person, who then adds new information back to the new account record.

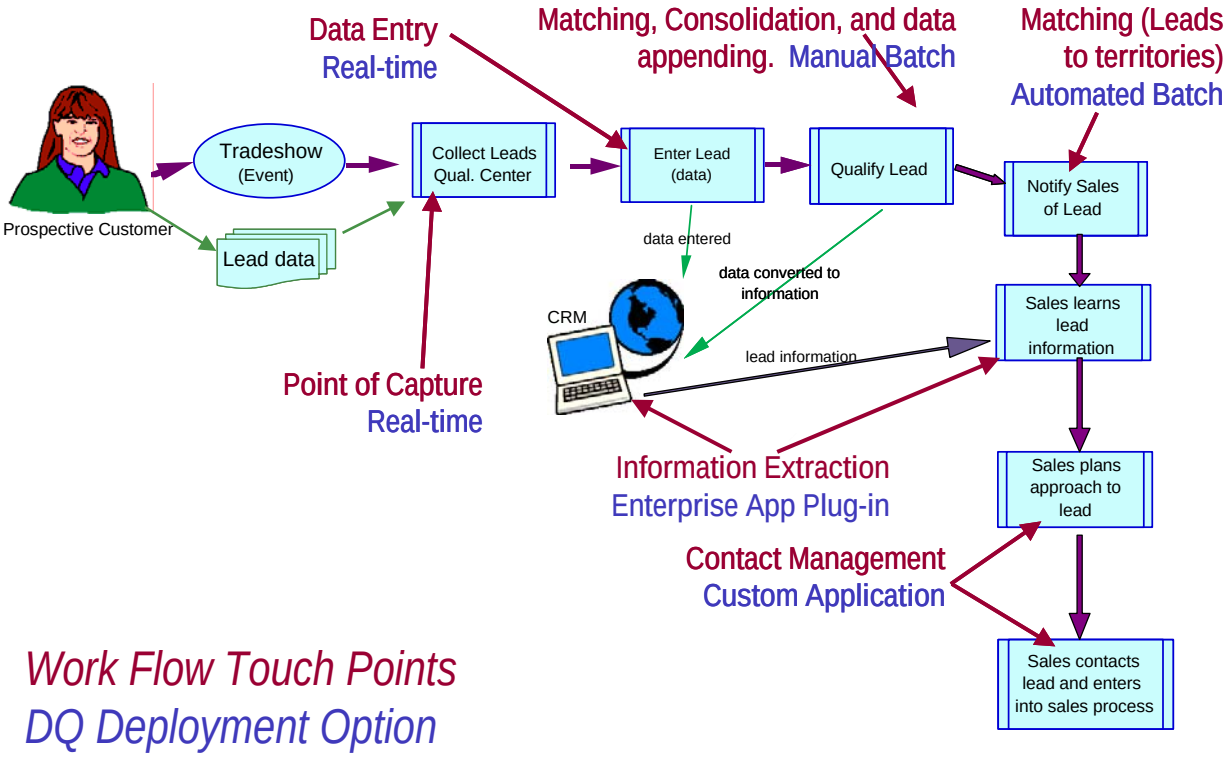



***Figure 3.***

In the diagram, two different concepts are indicated. Work flow touch points are shown in red. They are the locations in the workflow where data is manipulated. You can consider these as the locations where the workflow intersects the dataflow. Some of these locations, like “Point of Capture” actually spawn a

dataflow. Data quality deployment options are shown in purple. A deployment option is a specific type of software implementation that allows connectivity or use of data quality functionality at the point needed. In regards to work flow, data quality operations fall into the following areas…

- Front-office transaction — real-time cleansing
- Back-office transaction — staged cleansing
- Back-office batch cleansing
- Cross-office enterprise application cleansing
- Continuous monitoring and reporting

Each area broadly encompasses work activities that are either customer facing or not, or both, and the type of cleansing typically needed to support them. To facilitate these areas there are specific types of cleansing deployment options. Not to be confused with the connectivity options discussed in the Work flow factor, connectivity options are the three general methods for accessing the data (i.e. extraction, embedded procedures, and integrated functionality). Deployment options are forms of cleansing technology implementations that support a particular connectivity strategy. The deployment option list below identifies the types of options:

- Low-level Application Program Interface (API) software libraries — high-control custom applications
- High-level API software libraries — quick, low-control custom applications
- Web-enabled applications —real-time e-commerce operations
- Enterprise application plug-ins — ERP/CRM/ETL integrations
- GUI (graphical user interface) interactive applications — data profiling
- Batch applications — auto or manual start
- Web services/ASP connections — access to external or out-sourced functions

In each option, data quality functions can be incorporated that measure, analyze, identify, standardize, correct, enhance, match, and consolidate the data.

In a work flow, if a data touch point is not protected with validation functions, defective data will be captured, created, or propagated per to the nature of the touch point. An important action in the Workflow factor is listing the various touch points so locations where defective data can leak into your information stream are identified. The list can be superimposed on a work flow diagram, lending to planners the ability to visually map cleansing tactics, and logically cascade one data quality function to feed another.

If there is a “leaky” area in the information pipeline, the map will help to position redundant checks around the leak to contain the contamination. When building the list and map, concentrate on the data defined by the goals. A work flow may have numerous data touch points, but a subset will interact with specified data elements.

For example, a teleprospecting department needed to have all of the telephone area codes for their contact records updated because the account managers complained they were spending an increasing amount of time researching wrong phone numbers stemming from area code changes rather than making calls. The data touch points for just the area code data were far fewer than that of an entire contact record. By focusing on (in this case) the three touch points for area codes, the project manager was able to identify two sources of phone number data to be cleansed, and limit the project scope to just those touch points and data sources. With the project scope narrowly defined, operational impact and costs were reduced, and expectations of disruption were lowered. The net result was that it was easier to obtain approval for the project.

## *Stewardship*

No strategy is complete without the evaluation of the human factor and its effect on operations. Work flows- and data flows are initiated by people. Data itself has no value except to fulfill purposes set forth by people. The people who manage data processes are, in the current data warehouse vernacular, called data stewards. A plain, non-specialized steward is defined in the dictionary as, "One who manages another's property, finances, or other affairs." Extending that definition for our purposes, a ***data*** steward is a person who manages information and those activities that encompass data creation, capture, maintenance, decisions, reporting, distribution, and deletion. Therefore if a person performs any of these functions on a set of data, he or she is a data steward.

Much can be said about each of these activities, not to mention the principles of how to manage, provide incentives for, assign accountability, and structure responsibilities for data stewards. A discussion on organizational structures for data stewards could easily occupy a chapter in a book on data quality.

In the definition of steward, there is a caption to emphasize: "One who manages ***another's*** property …" Many times project managers have complained they can not move their project past a certain point because the stakeholders could not agree on who "owned" the data. This is dead center a stewardship issue. No steward owns the data. The data is owned by the organization just as surely as the organization owns its name, trademarks, cash, and purchased equipment. The debate on ownership is not really about ownership, but usually centers on who has the *authority* to approve a change to the data. The answer is the data stewardship team.

An action in the Stewardship factor is to identify the stakeholders (stewardship team) of the source data. Inform them of the plans, ask each one about their specific needs, and collect their feedback. If there are many stakeholders, selecting a representative from each user function is highly encouraged. To do less will surely result in one of three conditions: a) a change is made that alienates half of the users and the change is rolled back; b) half of the users are alienated and they quit using the system; c) half of the users are alienated, but are forced to use the system, and grumble and complain at every opportunity. Most would agree that any of these three outcomes are not good for future working relationships!

Some organizations have progressed to the point where a formal data stewardship team has been appointed. In this case, someone has already identified the stakeholders, and selected them as representatives on the team. This definitely makes strategy development a quicker process, as data stewards don't have to be located.

When evaluating the data stewardship factor for a new project the following tasks need to be performed:

- Answer such questions as — Who are the stakeholders of the data? Who are the predominant user groups, and can a representative of each be identified? Who is responsible for the creation, capture, maintenance, reporting, distribution, and deletion of the data? If one of these is missed — any one of them — their actions will fall out of sync as the project progresses, and one of those "You never told me you were going to do that!" moments will occur.
- Carefully organize requirements-collecting sessions with the stakeholders. Tell these representatives any plans that can be shared, assure them that nothing yet is final, and gather their input. Let these people know that they are critical stakeholders. If strong political divisions exist between stakeholders, meet with them separately and arbitrate the disagreements. Do not setup a situation where feuds could erupt.
- Once a near-final set of requirements and a preliminary project plan are ready reacquaint the stakeholders with the plan. Expect changes.

- Plan to provide training and education for any new processes, data model changes, and updated data definitions.
- Consider the impact of new processes or changed data sets on organizational structure. Usually a data quality project is focused on an existing system, and current personnel reporting structures can absorb the new processes or model changes. Occasionally, however, the existing system may need to be replaced or migrated to a new system, and large changes in information infrastructure are frequently accompanied by personnel shifts.
- Data quality projects usually involve some changes to existing processes. The goal of half of all data quality projects is, after all, work flow improvement. For example, a marketing department in one organization set a goal of reducing processing time of new leads from two weeks to one day. The existing process consisted of manually checking each new lead for duplications against its CRM system. The department decided to implement an automated match and consolidation operation. The resulting work flow improvement not only saved labor time and money, but also resulted in more accurate prospect data. With improvement comes change (sometimes major, sometimes minor) in the roles and responsibilities of the personnel involved. Know what those changes will be.

A plan to compile and advertise the benefits (return on investment) of a data quality project deserves strategic consideration. This falls in the Stewardship factor because it is the data stewards and project manager that will be tasked with justification. Their managers may deliver the justification to senior management, but it's often the data stewards who will be required to collect, measure, and assert the "payoff" for the organization. Once the message is crafted, do ***not*** underestimate the need for and value of repeatedly advertising how the improved data will *specifically* benefit the organization. Give the organization the details, which should be a component of an internal public or employee relations campaign. Success comes from continually reinforcing the benefits to the organization.. This builds inertia, while hopefully managing realistic expectations. This inertia will see the project through budget planning when the project is compared against other competing projects.

## *Continuous Monitoring*

The final factor in a data quality strategy is Continuous Monitoring. Adhering to the principals of Total Quality Management (TQM), continuous monitoring is measuring, analyzing, then improving a system in a continuous manner. Continuous monitoring is crucial for the effective use of data, as data will immediately age after capture, and future capture processes can generate errors.

Consider the volatility of data representing attributes of people. As stated earlier, in the U.S. 17 percent of the population moves annually. That means the addresses of 980,000 people change each week. A supplier of phone numbers reports that 7 percent of non-wireless U.S. phone numbers change each month, which equates to approximately 3.5 million phone numbers changing each week. In the U.S. 5.8 million people have a birthday each week, and an additional 77,000 are born each week. These sample statistics reflect the transience of data. Each week there are mergers and acquisitions that change the titles, salaries, and employment status of thousands of workers. The only way to effectively validate dynamic data for use in daily operations is to continuously monitor and evaluate using a set of quality measurements appropriate to the data.

A common question in this regard is, "How often should I profile my data?" Periodicity of monitoring is determined by four considerations:

1. How often the data is used — hourly, daily, weekly, monthly, etc.
2. The importance of the operation using the data — mission critical, life dependent, routine operations, end of month reporting, etc.

3. The cost of monitoring the data. After the initial expense of establishing the monitoring system and process, the primary costs are labor and CPU cycles. The better the monitoring technology, the lower the labor costs.
4. Operational impact of monitoring the data. There are two aspects to consider here. a) Impact of assessing operational (production) data during live operations and b) impact of the process on personnel. Is the assessment process highly manual, partially automatic, or fully automatic?

The weight of these four considerations will vary depending on their importance to the operation. The greater the importance, the less meaningful cost and operational impact of monitoring will be. The challenge comes when the operation is of moderate importance, and cost and operational impact are at the same level. Fortunately, data is supported by technology. While that same technology improves, it lowers the costs of monitoring, and lowers operational impacts.

Data is stored in electronic media, and even data stored in non-relational files can be accessed via sophisticated data profiling software. It is with this software that fully-automated and low-cost monitoring solutions can be implemented — thereby reducing the final consideration of continuous monitoring to "*how often"* it should be done. When purchased or built, a data profiling solution could be rationalized as "expensive," but when the cost of the solution is amortized over the trillions of measurements taken each year or perhaps each month, the cost per measurement quickly nears zero. Another factor that reduces the importance of cost is the ultimate value of continuous monitoring; finding and preventing defects from propagating, and therefore eliminating crisis events where the organization is impacted from those defects.

As the previous data-churn statistics show, data cleansing cannot be a one-time activity. If the data is cleansed today, tomorrow it will have aged. A continuous monitoring process allows an organization to measure and gauge the data deterioration so it can tailor the periodicity of cleansing. Monitoring is also the only way to detect spurious events such as corrupt data feeds — unexpected and insidious in nature. A complete continuous monitoring plan should address each of the following areas.

a) **Identify measurements and metrics to collect.** The place to start is with the project goals. The goals determine the context, the first data quality strategy factor. In Context, we determined what data support the goals. The measurements will be focused on this data. Various attributes (format, range, domain, etc.) of the data elements can be measured. The measurements can be rolled up or aggregated – each having its own weight – into metrics that are the combination of two or more measurements. A metric of many measurements can be used as a single data quality score at the divisional, business unit, or corporate level. A group of measurements and metrics can form a data quality dashboard for a CRM system, for instance. The number of defective addresses, invalid phone numbers, incorrectly formatted email addresses, and non-standard personnel titles can all be measured and rolled up into one metric that represents quality of just the contact data. Then, if the quality score of the contact data does not exceed a threshold defined by the organization, a decision is now possible to postpone a planned marketing campaign until cleansing operations raise the score above the threshold.

b) **Identify when and where to monitor.** The Storage, Data Flow, and Work Flow factors provide the information for this step. The Storage factor tells what data systems house the data that needs to be monitored. The Work Flow factor tells how often the data is used in a given operation and will provide an indication as to how often it should be monitored. The Data Flow factor tells how the data moves, and how it has been manipulated just prior to the proposed point of measure. A decision continuous monitoring will face is whether to measure the data before or after a given operation. Is continuous monitoring testing the validity of the operation, or testing the validity of the data to fuel the operation, or both?

One pragmatic approach is to put a monitoring process in place to evaluate a few core tables in the DW on a weekly basis. This will identify defects inserted by processes feeding the data warehouse, and defects caused by aging during the monitoring interval. It may not identify the source of the defects if multiple inputs are accepted. The monitoring operation would need to be moved further upstream or timed to occur after each specific update to isolate changes from multiple events.

Organizations should be aware that this simple approach will not optimally fit an organization's goals, but will suffice for an initial implementation. An enhancement to the simple plan is to also monitor the data at the upstream operational data store (ODS) or staging areas. Monitoring at the ODS will identify defects in isolation from the data warehouse, and capture them closer to the processes that caused them. The data in the ODS is more dynamic and therefore monitoring may need to be performed in greater frequency, i.e., nightly.

c) **Implement monitoring process.** This involves configuring a data profiling software solution to test specific data elements against specific criteria or business rules, and save the results of the analysis to a metadata repository. Once established, when to monitor and where implementing the process is relatively straightforward. Most data profiling packages can directly access relational data sources identified in the storage factor. More sophisticated solutions are available to monitor non-relational data sources such as mainframe data, and open systems flat files.

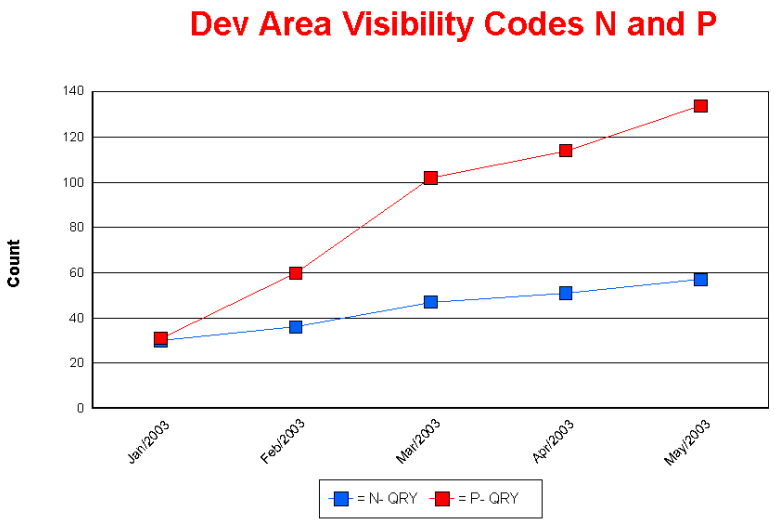
Configuring the data profiling software involves establishing specific business rules to test for. For example, a part number column may have two allowed formats: ###A-### and ###-###, where # is any valid numeric character, and A is any character in the set A, B, C, and E. The user would enter the two valid formats into the data profiling software where the rules are stored in a metadata repository. The user can then run the rules as ad-hoc queries or as tasks in a regularly scheduled, automated monitoring test set.

d) **Run a baseline assessment.** A baseline assessment is the first set of tests conducted to which subsequent assessments in the continuous monitoring program will be compared. Identifying the business rules and configuring the data profiling software for the first assessment is where the majority of work is required in a continuous monitoring program. Building the baseline assessment serves as a prototyping evolution for the continuous monitoring program. First iterations of tests or recorded business rules will need to be changed as they will not effectively evaluate criteria that are meaningful to the people reviewing the reports. Other rules and the data will change over time as more elements are added or the element attributes evolve. The initial setup work for a baseline assessment is leveraged when the final set of analysis tasks and business rules is run on a regular basis.

e) **Post monitoring reports.** A common failing of a continuous monitoring program is poor distribution or availability of the analysis results. A key purpose of the program is to provide both information and impetus to correct flawed data. Restricting access to the assessment results is counterproductive. Having a data profiling solution that can post daily, weekly or monthly reports automatically, after each run, to a corporate Intranet is an effective communication device and productivity tool. The reports should be carefully selected. The higher the level of manager reviewing the reports, the more aggregated (summarized) the report data should be.

The report example below in Figure 4 offers two different measurements superimposed on the same

chart. In this case, a previous business rule for the data stipulated there should be no NULL values. When numerous NULL values were indeed found another test was implemented to track how effective the organization was at changing the NULLs to the valid values of N or P.




(put label in here: Figure 4)

This level of reporting was appropriate for field level analysts and managers who had to cure a specific process problem, but is too low level for a senior manager. For a director level or higher position a single aggregate score of all quality measurements in a set of data would be more appropriate.

f) **Schedule regular data steward team meetings to review monitoring trends.** Review meetings can be large or small, but they should occur regularly. Theoretically, they could occur as often as the battery of monitoring tests. If the tests are run nightly, meeting daily as a team may be a burden. A single person could be assigned to review the test runs, and call the team together as test results warrant. However, a typical failing of continuous monitoring programs is follow-through. The information gained is not acted upon. While tremendous value can be derived from just knowing what data is defective and avoiding those defects, the greatest value comes from fixing the defects early in the trend. This cannot be done unless the stewardship team, either as individuals, or as a team, implements a remediation action to both cleanse the data and cure the process that caused the defects.

In summary, continuous monitoring alerts managers to deterioration in data quality early in the trend. It identifies which actions are or are not altering the data quality conditions. It quantifies the effectiveness of data improvement actions, allowing the actions to be tuned. Last, and most importantly, it continually reinforces the end users' confidence in the usability of the data.

The irony is many systems fall into disuse because of defective data, and stay unused even after strenuous exertions by IT to cleanse and enhance the data. The reason is perception. The system is perceived by the users, not IT, to still be suspect. A few, well-placed and ill-timed defects can destroy overnight the reliability of a data system. To regain the trust and confidence of users a steady stream of progress reports and data scores need to be published. These come from a continuous monitoring system that shows and convinces users over time the data is indeed improving.

## TYING IT ALL TOGETHER

In order for any strategy framework to be useful and effective it must be scalable. The strategy framework provided here is scalable from a simple one-field update such as validating gender codes of male and female, to an enterprise-wide initiative where 97 ERP systems need to be cleansed and consolidated into 1 system. To ensure the success of the strategy, and hence the project, each of the six factors must be evaluated. The size (number of records/rows) and scope (number databases, tables, and columns) determines the depth to which each factor is evaluated.

Taken all together or in smaller groups, the six factors act as operands in data quality strategy formulas.

- Context by itself = The type of cleansing algorithms needed
- Context + Storage + Data Flow + Work Flow = The types of cleansing and monitoring technology implementations needed
- Stewardship + Work Flow = Near-term personnel impacts
- Stewardship + Work Flow + Continuous Monitoring = Long-term personnel impacts
- Data Flow + Work Flow + Continuous Monitoring = Changes to processes

It is a result of using these formulas the people come to understand that information quality truly is the integration of people, process, and technology in the pursuit of deriving value from information assets.

## IMPLEMENTATION AND PROJECT MANAGEMENT

Where the data quality strategy formulation process ends, data quality project management takes over. In truth, much, if not all of the work resolving the six factors, can be considered data quality project planning. Strategy formulation often encompasses a greater scope than a single project and can support the goals of an entire enterprise, numerous programs, and many individual projects. Sooner or later, strategy must be implemented through a series of tactics and actions, which fall in the realm of project management. While the purpose of this paper is not to cover the deep subject of data quality project management, it does need to set the stage for a clear transition from strategy formulation to the detailed management of the tasks and actions that ensure its success.

Once a strategy document – big or small, comprehensive or narrowly focused – is created, it can be handed to the project manager and everything he or she needs to know to plan the project should be in that document. This is not to say all the work has been done. While the goals have been documented, and the data sets established, the project manager must build the project requirements from the goals. The project manager should adhere to the sound project management principals and concepts that apply to any project, such as task formulation, estimation, resource assignments, scheduling, risk analysis, mitigation, and project monitoring against critical success factors. Few of these tactical issues are covered in a strategy-level plan.

Another facet of a successful data quality strategy is consideration of the skills, abilities, and culture of the organization. If the concept of data quality is new to the organization, a simple strategy is best. Simple strategies fit pilot projects. A typical pilot project may involve one column of data (of phone numbers for example) in one table, impacting one or two users, and involved in one or two processes. A simple strategy for this project, encompassing all six factors, can fit on one page of paper.

However, the more challenging the goals of a data quality strategy, the greater the returns. An organization must accept that with greater returns come greater risks. Data quality project risks can be mitigated by a more comprehensive strategy. Be aware that the initial strategy is a first iteration. Strategy

plans are “living” work products. A complex project can be subdivided into mini-projects, or pilots. Each successful pilot builds inertia. And therein lies a strategy in itself: divide and conquer. Successful pilots will drive future initiatives. Thus an initial strategy planning process is part of a larger recurring cycle. True quality management is, after all, a repeatable process.

# 1. APPENDIX A: DATA QUALITY STRATEGY CHECKLIST

To help the practitioner employ the data quality strategy methodology, the core practices have been extracted from the factors and listed here.

a) A statement of the goals driving the project
b) A list of data sets and elements that support the goal
c) A list of data *types* and *categories* to be cleansed[(1)]
d) A catalog, schema or map of where the data resides[(2)]
e) A discussion of cleansing solutions per category of data[(3)]
f) Dataflow diagrams of applicable existing dataflows
g) Work flow diagrams of applicable existing work flows
h) A plan for when and where the data is accessed for cleansing[(4)]
i) A discussion of how the dataflow will change after project implementation
j) A discussion of how the workfflow will change after project implementation
k) A list of stakeholders affected by the project
l) A plan for educating stakeholders as to the benefits of the project
m) A plan for training operators and users
n) A list of data quality measurements and metrics to monitor
o) A plan for when and where to monitor[(5)]
p) A plan for initial and then regularly scheduled cleansing

(1) A type is text, datetime, double, etc., and a category is street address, part number, contact name, etc.
(2) This can include the name of the LAN, server, database, etc.
(3) This should include possible and desired deployment options for the cleansing solution. See *Section 4.4 Work flow* for specific deployment options
(4) This will cover the *when* – during what steps in the dataflow and work flow will the cleansing operation be inserted – and *where* – on what data systems will the cleansing operation be employed – of the cleansing portion of the project.
(5) This will include running a baseline assessment, and then selecting tests from the baseline to run on a regular basis. Reports from the recurring monitoring will need to be posted, and regular review of the reports scheduled for the data stewardship team.