

Paper 098-29

Data Quality Management

The Most Critical Initiative You Can Implement

Jonathan G. Geiger, Intelligent Solutions, Inc., Boulder, CO

ABSTRACT

SIX HUNDRED BILLION DOLLARS ANNUALLY – Got your attention? That is what poor data quality costs American businesses, according to the Data Warehousing Institute. Poor data is also the leading cause of many IT project failures. So, given that this is such a serious problem, why aren't more companies addressing it more aggressively? This session discusses these topics as well as those detailing how companies can improve their data quality using the proven architectural blueprint, the Corporate Information Factory, as a basis for mapping your data quality processes. This session will explain the importance of data quality management, quality expectations and techniques for setting them. Finally, the program ends with practical advice for getting started on your data quality management program. The specific topics covered include:

- What is Data Quality Management?
- Data Quality Management Challenges
- Data Quality Definition
- Four Pillars of Data Quality Management
- Getting Started

INTRODUCTION

Corporations have increasingly come to realize that data is an important corporate asset. Unlike tangible corporate assets, however, it is difficult to place a definitive value on it. In today's tough economic climate, the lack of a tangible return on investment makes it difficult to fund activities to manage data as a strategic asset.

This paper defines data quality and its role within a business intelligence environment, and explains the importance of data quality management and the major challenges facing companies trying to implement a data quality management program.

WHAT IS DATA QUALITY MANAGEMENT?

Simply put, data quality management entails the establishment and deployment of roles, responsibilities, policies, and procedures concerning the acquisition, maintenance, dissemination, and disposition of data. A partnership between the business and technology groups is essential for any data quality management effort to succeed. The business areas are responsible for establishing the business rules that govern the data and are ultimately responsible for verifying the data quality. The Information Technology (IT) group is responsible for establishing and managing the overall environment – architecture, technical facilities, systems, and databases – that acquire, maintain, disseminate, and dispose of the electronic data assets of the organization.

Organizations of all kinds make decisions and service customers based on the data they have at their disposal. A data warehouse is often used to examine business trends to establish a strategy for the future; within the scope of a customer relationship management (CRM) program, data about the customer is used to make appropriate decisions concerning that customer; and data in the financial systems is used to understand the profitability of past actions. The viability of the business decisions is contingent on good data, and good data is contingent on an effective approach to data quality management.

The initial emphasis of many new data quality management initiatives launched in recent years has been on customer data, and technology has stepped up to this challenge by automating solutions to many of the data quality problems associated with customer data. Business data consists of much more than just customer data and the technology to support it. For instance, standardizing part codes and names and combining product information that is stored differently in various systems poses some new data challenges. The technology that deals with this non-name and address data must include an engine that consistently learns and evolves with the new data types, enabling it to

clean, reconcile, and match any type of information.

ROLES AND RESPONSIBILITIES

Within a business intelligence environment, there are several roles that are involved in data quality management:

- Program Manager and Project Leader
- Organization Change Agent
- Business Analyst and Data Analyst
- Data Steward

The Program Manager and Project Leader are responsible for overseeing the business intelligence program or individual projects, and for managing day-to-day activities based on the scope, budget, and schedule constraints. These people set the tone with respect to data quality and interact with the business representatives to establish the data quality requirements.

The Organization Change Manager helps the organization understand the value and impact of the business intelligence environment, and helps the organization address the issues that arise. Often, data quality issues are unearthed during the business intelligence projects, and the organization change agent can play an instrumental role in helping the organization understand the importance of dealing with the issues.

The business analyst conveys the business requirements, and these include detailed data quality requirements. The data analyst reflects these requirements in the data model and in the requirements for the data acquisition and delivery processes. Together, they ensure that the quality requirements are defined, reflected in the design, and conveyed to the development team.

The data steward is ultimately responsible for managing data as a corporate asset. This role is defined further in subsequent sections of this paper.

REACTIVE AND PROACTIVE COMPONENTS

A successful data quality management program has both proactive and reactive components. The proactive component consists of establishing the overall governance, defining the roles and responsibilities, establishing the quality expectations and the supporting business practices, and deploying a technical environment that supports these business practices. Specialized tools are often needed in this technical environment.

The reactive component consists of dealing with problems that are inherent in the data in the existing databases. The quality of data in legacy systems that were developed without a data quality management program in place may be inadequate for meeting new business needs, as shown by the different representations of the same data in Figure 1. For example, the accuracy of customer information may be good enough to bill a customer, but not good enough to understand the profitability of the customer; parts information may be good enough for each manufacturing facility's needs, but not good enough for understanding inventory levels and moving to a virtual parts warehouse approach. Another data problem that requires reactive action results from mergers and acquisitions. Data from the two previously separate companies needs to be combined, and this is often a daunting task, particularly if it must be undertaken without adequate tool support.

<i>Mill Names</i>
Courtland Mill
Courtlan
Mill, Courtland

<i>Product Names</i>
Savvy Matte 87 60-100
courtland Savvy Matte 60-100 87
Courtland 60-100, 87, Savvy Matte

<i>Same Address</i>
1002 W Main St #1234
PO BOX 1234 1 002 West Main

Figure 1 – Information from different databases may appear different

DATA QUALITY MANAGEMENT IMPORTANCE

Both aspects of the data quality management program are important – the reactive components address problems that already exist, and the proactive components diminish the potential for new problems to arise. The partnership between the business and information technology groups is critical for a successful program. Once IT understands the business rules, it is in a position to deploy the technology needed to ensure that data is being managed as a corporate asset.

Unfortunately, many companies learn about the importance of data quality management the hard way. Only after several documented problems with the data do they recognize the need to improve its quality. The U. S. Government estimates that billions of dollars are lost annually due to data quality problems. The Data Warehousing Institute concluded that the cost of data quality problems exceeds \$600 Billion annually. Additional estimates have shown that 15-20% of the data in a typical organization is erroneous or otherwise unusable. The importance of data quality management should be evident – so why aren't companies addressing it more aggressively?

DATA QUALITY MANAGEMENT CHALLENGES

Deploying a data quality management program is not easy; there are significant challenges that must be overcome. Some of the most significant reasons companies do not pursue a formal data quality management initiative include:

- No business unit or department feels it is responsible for the problem.
- It requires cross-functional cooperation.
- It requires the organization to recognize that it has significant problems.
- It requires discipline.
- It requires an investment of financial and human resources.
- It is perceived to be extremely manpower-intensive.
- The return on investment is often difficult to quantify.

RESPONSIBILITY

One of the more significant challenges is that no single business unit is responsible for all of the data in an enterprise, and the inclusion of a responsibility for data quality in a job description is very unusual. Furthermore, once the data is in the computer, business units often wash their hands of the problem and blame it on IT. IT cannot create the business rules nor should it be held responsible to make business decisions concerning the data. IT can only ensure that electronic rules, based on business rules, operate correctly.

Effective data quality management requires organizations to adopt a data stewardship approach. Stewardship is different than ownership. A steward is a person who is expected to exercise responsible care over an asset that he or she does not own. The data is actually owned by the enterprise. The steward is responsible for caring for that asset.

Data stewardship is important, but establishing a data stewardship program is very difficult! One immediate challenge to a stewardship program is to identify the group or person responsible for a set of data. Some responsibilities are fairly easy to allocate: the Facilities Management Department may be responsible for data about real estate and property; the Human Resources Department may be responsible for data about employees; and the Finance Department may be responsible for the organization's financial data. But who is responsible for payroll data (its financial data about employees), who is responsible for customer data, and who is responsible for product data?

Another requirement is to get business people to focus on the data issues. The business leaders are concerned with their functional responsibilities. A Marketing Vice President understands customer segmentation and campaign management; data quality management is not his or her forte. The Marketing Vice President must recognize that unless his or her quality expectations are established for the data, it is unlikely that the condition of the data will be

sufficient to support his or her needs. Similarly, the Manufacturing Vice President is focused on producing the products. He or she must recognize that data about the products (e.g., specifications, substitutable raw materials, inventory levels at warehouses and customer sites) impacts the department's ability to profitably produce the products.

CROSS FUNCTIONALITY

The lack of clear responsibility for data leads to the second major challenge. Unlike most corporate functions, which are vertically aligned within an organizational unit, data quality management is horizontal in nature. That is, data quality management responsibilities cross organizational boundaries.

Effective data quality management provides organizations with the ability to share data and that requires consistent definitions. If three business units (e.g., Manufacturing, Product Development, and Sales) have responsibility for products, they must work together to arrive at common business rules and definitions for data about products and the way they may be sold. This means that each of the autonomous units needs to give up some of its control over the product, and that is often difficult to attain.

PROBLEM RECOGNITION

At the beginning of a project, we've had clients tell us that they have no data quality problems. We've never had a client tell us that after completing that project. This is because during the course of the project, we helped the client understand its data better, and in that process, defects in the existing data come to light.

One of the first steps to solving any problem is recognizing that the problem exists. Organizations are often in denial about their data quality problems, and it sometimes takes a major catastrophe to change that attitude. Absent such an event, organizations are not prone to spend money fixing something that they don't think is broken.

DISCIPLINE

The fourth major challenge is discipline. An effective data quality management program requires discipline. Responsibilities must be assigned and formal procedures must be created and followed – by everyone who handles data for the enterprise. Assigning responsibilities means that a field agent needs to understand the value of the data gathered about a product sale to a customer so that he or she captures it correctly, it means that IT understands the quality expectations so that it builds the systems to process the data correctly, and it means that the business analysts using the data understand its meaning so that they can make well-informed decisions. Further, it means that job descriptions for individuals in these positions reflect their data quality management responsibilities.

INVESTMENT

Philip B. Crosby wrote a book entitled "Quality is Free", and I firmly believe that this claim is true. The thrust of the book is that it isn't quality that is expensive; it's the cost of "unquality". Examples of the cost of "unquality" include the cost of sending duplicate promotional materials because customers are duplicated in the database, the opportunity cost of not sending materials to the right customers because the data used to segment customers is flawed, the opportunity cost of not shipping products to a customer because of inaccurate information about inventory levels, and the time spent finding and reconciling data needed to make effective decisions.

ON-GOING EFFORT

Organizations are constantly looking for ways to avoid increasing their staff size, or in many cases, to reduce the staff size. A data quality management program requires people, and it's very difficult to obtain authorization for an effort that will increase staff size. Implementing data quality management technology, however, automates manual data quality and integration projects and frees up resources for other projects. Because a data quality management software investment costs less than one employee (salary, benefits, rent, etc), Implementations actually support organizational goals of increased productivity or reduced staffing.

Business representatives will be needed to perform the governance related activities, and this is unavoidable. The effort that is avoidable is the staff required to identify and correct data problems on an on-going basis. This is an area in which technology plays an important role. Data quality management technology can directly support each of the four phases of the program described in the next major section. The technology can help gain an understanding of the data (data profiling), take steps to correct problems (data quality), automate some discrepancy resolution resulting from merging data from multiple sources (data integration), and add value to the data (data augmentation).

To be effective, the tools need to be customizable so that they impose rules established by the organization. In

addition, the tools need to learn from the routines they execute so that they can deploy rules based on learned data patterns.

RETURN ON INVESTMENT

Data quality management efforts are difficult to fund because the cost of “unquality” is not documented. The documentation of these costs requires recognition of the problem (as discussed earlier), and also requires managers to admit that they are wasting money or that they are not effectively utilizing resources at their disposals. Making these admissions, particularly in tough economic times, is risky. It is imperative that top management create an environment in which people are not unduly penalized for admitting to past problems.

People within most organizations are aware of data quality problems and have taken steps to work around them. Solicit information from people working with the data and canvass customer and supplier complaints and try to discern the ones that may have been caused by erroneous data. While this information may not provide a return on investment prediction in financial terms, you are likely to find enough examples of problems to justify addressing at least one area. When you do address an area, be sure to document both the costs and the resultant savings, and use that information to justify data quality management initiatives in other areas.

SUMMARY

Once an organization is committed to moving forward with data quality management, it needs a disciplined approach for tackling its data problems.

DATA QUALITY DEFINITION

Contrary to popular belief, quality is not necessarily zero defects. Quality is conformance to valid requirements. In defining quality, we must:

- Determine who sets the requirements
- Determine how the requirements are set
- Determine the degree of conformance that is needed

Ultimately, it is the business that needs to set the quality requirements. If a stewardship program is in place, then the person responsible for a specific set of data is known; if there is no stewardship program in place, then the person responsible for the data in question needs to be identified and that person’s authority to make decisions concerning the quality requirements needs to be recognized. The IT organization contributes to the decision. In the absence of other information, there will be a tendency to define quality as perfection. The IT organization needs to ensure that the business person making the decision is aware of the existing data quality deficiencies and the practicality and cost of overcoming them. Sometimes, changes in the business processes and compensation policies will be needed to address data quality problems. These factors must enter the decision process.

DEGREE OF CONFORMANCE

The degree of conformance must then be set. The degree of conformance represents the tolerance level for errors. Data quality is then established by knowing whether or not the target has been achieved.

DATA WAREHOUSE ROLE

The data warehouse environment is interesting in that it is the source of information used by the business to make strategic decisions, but it does not actually create any data. That means that data quality problems in the data warehouse originate in the source system environment, are created because of faulty data acquisition and delivery processes, or are due to interpretation problems.

Data quality problems in the source systems need to be recognized, and the data warehouse team is responsible to either address these problems or gain business concurrence that the problems are acceptable. The data warehouse team must then ensure that the data warehouse users are aware of these data quality deficiencies.

Faulty data acquisition and delivery processes are fully within the control of the data warehouse team, and it is responsible to ensure that these don’t happen.

Interpretation problems also need to be addressed by the data warehouse team. The responsibilities of the business analyst and data analyst include establishing clear definitions for each data element and ensuring that the data

acquisition and delivery specifications instruct the programmers on how to properly populate the data element.

STRIKING A BALANCE

Ultimately, data quality is a balance. Using two of the more common metrics of data quality – completeness and accuracy – we see that to have data that is both 100% accurate and 100% complete can be very expensive, and is not always achievable. We often find ourselves sacrificing one or the other. For example, when there is a data error, we need to decide whether completeness is more important, in which case we may include the record with an error, or whether accuracy is more important, in which case we may omit the record.

DEALING WITH ERRORS

When data quality problems are encountered importing data into the data warehouse, there are four viable actions that can be taken. We can:

- Reject the error
- Accept the error
- Correct the error
- Apply a default value for the erroneous data

When accuracy is more important than completeness, it may be appropriate to reject the error. When the data is known to be erroneous, but it is within the tolerance level, then it is appropriate to accept the error. When the correct value can be determined, then the error can be corrected. Lastly, when the correct value cannot be determined and completeness is very important, then a default value can be substituted for the erroneous data. Regardless of the course of action taken, it is very important that the data warehouse users understand the data quality implications of the chosen remedy.

When data is corrected coming into the data warehouse, we are faced with an interesting situation. The data warehouse ceases to match the source system, and becomes more accurate than the source system. Minimally, the difference needs to be explained. Ideally, the process that was used to correct the data in the data warehouse can trigger a transaction that also corrects data in the source system.

The next section describes a four phased approach to understanding the data, cleansing each source, integrating data from multiple sources, and enhancing the value of the data.

FOUR PILLARS OF DATA QUALITY MANAGEMENT

Once the data quality management initiative is sanctioned, the specific data subjects to be addressed and their priorities are determined. (A method for doing this is discussed within the Getting Started section.) A four-phase process for achieving successful data quality management for any particular set of data follows. Product data is used as an example within this paper. As you will see, proper tools can significantly reduce the effort required to perform each of the four steps.

DATA PROFILING

Data profiling is the process of gaining an understanding of the existing data relative to the quality specifications, as shown in Figure 2. This must be your starting point. Just as you can't establish driving directions to a destination unless you know your starting point, you can't take measures to improve data quality unless you know the quality of your existing data. If you're building a data warehouse, this step is typically called "source system analysis".

This process consists initially of looking at the actual data. Two key questions to ask are whether the data is complete, and whether the data is accurate.

Issue	Example
Out of Acceptable Range	Patient Age = 185 ()
Non-Standard Data	Main Str, Main Street, Main ST, Main St.

Invalid Values	Data can be "A" or "B" but Value = "C"
Differing Cultural Rules	Date = Jan 1, 2002 or 1-1-2002 or 1 Jan 02
Varying Formats	(919)674-2153 or [919]6742153 or 9196742153
Cosmetic	jon j jones transformed into Jon J Jones
Verification	ZIP code does not correspond to correct City & State

Figure 2 – Data profiling determines if the data is complete and accurate

Data completeness has two facets. The first entails analyzing whether every product is included in the database. To perform this analysis, we need to understand what constitutes a product. For example, are we only interested in products that we manufacture or are we also interested in competitor products? If we're interested in competitor products, we often need to use external sources to obtain the information. The second facet of data completeness entails having data in each required field. For example, if we have a field for the product's launch date, is it populated?

Data accuracy is another dimension. With data accuracy, we are concerned with the values in the field. Just because a product has a value in the launch date field does not mean it is correct. With data profiling, we can look at the data file as a whole and discover a disproportionate number of products with a common launch date such as November 11, 1911, which would result from a data entry field being filled with 11-11-11. This information provides us with a starting point. We are now in a position to do something about it.

Another discovery made during profiling is that we have too much data. An examination of our data may indicate that we have more products than we thought, and a review of some of the data could reveal duplications.

DATA QUALITY

In this step, we build on the information learned in data profiling to understand the causes of the problems. For example, the data profiling activities could reveal that we have duplicate data. The analysis portion (validation) of data quality uncovers the symptoms – different representations of the same product due to inconsistencies in the data. With the use of the technology, additional duplicate data may be discovered as well.

To the extent that we can define valid domains, we also validate the values of the data to identify specific quality problems. Once we identify the specific data problems, we are in a position to do something about it. Here, we can choose one of four basic options previously stated:

- Exclude the data: if the problem with the data is deemed to be severe, the best approach may be to remove the data.
- Accept the data: even if we know that there are errors in the data, if the error is within our tolerance limits, the best approach sometimes is to accept the data with the error.
- Correct the data: when we encounter different variations of a customer name, we could select one to be the master so that the data can be consolidated.
- Insert a default value: sometimes it is important to have a value for a field even when we're unsure of the correct value. We could create a default value (e.g., unknown) and insert that value in the field.

The specific approach taken may differ for each data element, and the decision on the approach needs to be made by the business area responsible for the data. It is also important to note that the data quality activity improves the quality of the data that already exists. It does not address the root cause of the data problems. If the enterprise is truly interested in improving data quality, it must also investigate the reasons that the data contained the errors and initiate appropriate actions, including incentives and changes to business procedures to improve future data.

DATA INTEGRATION

Data about the same item often exists in multiple databases. This data can take virtually any form (customer name and address data, product data, etc). Data quality management / integration can be applied to virtually any data problem, as shown in Figure 3.

One company, for example, had two product files – a master product extract from its US-based ERP package and a product extract from Europe. The company sold the same products in both areas, but the products may be sold under

different names, and the product, brand, and description patterns in each file were based on the data entry personnel.

Submitted Data	Standardized Data
DataFlux Corp	DataFlux Corporation
DataFlux Inc	DataFlux Incorporated
DataFlux Co.	DataFlux Company
MR JOHN SMITH	Mr. John Smith
Mister Jonathan Smith	Mr. John Smith
Mster John SmITh	Mr. John Smith
#(877)8463589 ext250	877-846-3589 ext 250
8778463589 x250	877-846-3589 ext 250
International Paper	International Paper Company
IP	International Paper Company
Intl Paper	International Paper Company
Interntl Paper	International Paper Company

Figure 3 – Data from different sources must be integrated

The first challenge in data integration is to recognize that the same customer exists in each of the two sources (“linking”), and the second challenge is to combine the data into a single view of the product (“consolidation”).

With customer data, we often find a common field (e.g., tax identification number) that can be used to identify commonality. When this occurs, then multiple records for the same customer can quickly be identified. With product data, this is often not the case. To help the company cited above, data quality management technology was used to match the product information across the two different systems. This automation eliminated a task that required almost half of a subject matter expert’s time on a continuous basis.

The common data across both files was simply a product description. The US file contained the description, brand name, and product identifier all in one field and in various patterns. The European file contained just product descriptions, also with varied patterns and abbreviations. The data quality management technology was used to perform the following processes:

- Parse the description from the US file into product-specific attributes and into brand name
- Reconcile the differences in brand names
- Reconcile the differences in product attributes (short forms, abbreviations, etc.)
- Phonetically match the reconciled data
- Display reports of matching products

Teaching the engine how to handle this company’s business data enabled the company to save 24 man-weeks per year – in each of its seven business units.

DATA AUGMENTATION

Data augmentation is the last step for increasing the value of data. Data augmentation entails incorporating additional external data not directly related to the base data, to gain insight as shown in Figure 4. With customer data, it’s very

common to combine internal data with data from third parties to increase an understanding of the customer and his or her buying potential and loyalty. We might also obtain data about the behavior of customers with certain attributes. By combining that data with data about our specific customers, we can segment customers more effectively to identify specific opportunities.

Address	4001 Weston Pkwy
House Number	4001
Street Name	Weston
Street Type	Pkwy
City	Cary
State	NC
ZIP	27513
ZIP+4	2303
Carrier Route ID	34
Delivery Point ID	1
County Number	183
County Name	Wake
Congressional District	4
Record Type	Business
Latitude	35.8306
Longitude	-78.7841
Census Block Group	371830535122
Census Block Group Digit	2
Census Tract	053512
County FIPS Code	183
State FIPS Code	37

Figure 4 – Example of Data Augmentation

Another example of data augmentation is the incorporation of data based on the address or full postal code of a customer. Knowing, for example, that all the houses in a particular Zip+4 area were built in 1980 and have 3,000 – 4,000 square feet provides us with data to target certain product offerings.

Data augmentation related to addresses is common. It also applies to other business data, such as products. Sales analysis often entails understanding sales patterns of product sales. In the manufacturing industry, the company's knowledge of the actual consumer sales is limited – it only knows what it sold to the wholesaler or retailer. Information about the actual sales needs to be acquired from third parties – either the retailers or data providers such as Nielsen. Internal sales data can then be augmented with this information to gain an understanding of the actual sales patterns and to help make decisions concerning delivery times and quantities to ensure that the retailer is never out of stock.

Additionally, the information on the company's product sales can be combined with information about competitor sales. The competitor sales information is obtained from the third party, but typically includes only a product identifier. Other information about the competitor product then needs to be obtained, and all of this has to be encoded in a way

that enables the company to make the appropriate comparisons.

Data augmentation is another opportunity for technology to help in data quality management. With data augmentation and merging capabilities integrated into these tools, the manual effort to develop meaningful information from business data is significantly enhanced.

The value of our data can be substantially increased if we understand it, ensure its quality, integrate it, and augment it. The next section provides you with valuable tips on getting started with your data quality management initiative.

GETTING STARTED

The challenges in deploying a data quality management initiative are significant, but they are not insurmountable. The initial efforts should encompass:

- Education
- Stewardship
- Partnerships
- Four-phase program
- Technology support

EDUCATION

Support from the key stakeholders will be better if they understand data quality management and are committed to its deployment. The education needs to consist of a combination of theory concerning this subject, case studies from other organizations, and specific issues within your organization's data. It will be much easier to gain support for the initiative if people recognize that the organization is either wasting money or is missing business opportunities due to data quality management deficiencies.

STEWARDSHIP

A steward is a person who is called upon to exercise responsible care over possessions entrusted to him or to her. In the case of the data steward, it is the person responsible for exercising care over the data assets of the organization. There are numerous responsibilities that need to be fulfilled. Some of these are fulfilled by the business data steward, while others are handled by IT team members. Responsibilities can be classified as acquisition, management, dissemination, and disposal

Data acquisition responsibilities include:

- Business processes that create or change data
- Systems that handle the data
- Establishing authorities with respect to capturing and updating the data
- Establishing the validation rules for the data
- Establishing the business rules for handling the data
- Establishing the quality constraints for the data

Data management responsibilities include:

- Developing and maintaining the data models
- Understanding the demographics of the data
- Establishing and enforcing data naming standards
- Establishing meta data requirements and complying with them
- Managing data redundancy
- Backup and recovery of the data
- Data archival and restoration

Dissemination responsibilities include:

- Defining access security rules and complying with them
- Establishing standard query and reports
- Providing capabilities to the users
- Managing system use
- Monitoring output data quality
- Providing appropriate meta data

Disposal responsibilities include:

- Establishing and complying with retention rules
- Deleting data in compliance with business practices

DATA STEWARDSHIP SKILLS

Data stewards need both technical and interpersonal skills. Technical skills include a basic understanding of data modeling, a basic understanding of database management systems, a strong understanding of data warehousing concepts, facilitation skills and technical writing skills. Interpersonal skills include a solid understanding of the business, excellent communications skills, objectivity, creativity, diplomacy, being a team player, being well-respected in the subject area and in the knowledge of the overall organization.

ESTABLISHING DATA STEWARDS

One of the challenges facing organizations is identifying the people who need to perform the stewardship role. Following is an approach for identifying the people responsible for each set of data. The approach employs a matrix for determining the appropriate representation on data stewardship teams for each major category of data that will be managed. The matrix has two axes: data subject areas and business functions. The intersection of each subject area and function is annotated to reflect whether or not that type of data is used in that business function. Where the data is used, the annotation describes the way in which it is used: created, read, updated, or deleted. Matrices of this type are sometimes referred to as 'CRUD' matrices. As a prerequisite, the organization must already have defined its subject areas and business functions.

Lay out a matrix with subject areas on the X axis, and business functions at the desired level of detail on the Y axis. It is sometimes helpful to add the next-higher level of business functions as groupings on the Y axis, to improve legibility. See Table 1 below for a sample of how this matrix might look. Note that in this sample, fictitious functions and subject areas are used.

Table 1: Sample Blank Matrix

		Customer	Product	Order	Contract	Part	Supplier	Employee	Financials
Finance	Fin Func 1								
	Fin Func 2								
	Fin Func 3								
Sales	Sales Func 1								
	Sales Func 2								
Supply Chain	SC Func 1								
	SC Func 2								
	SC Func 3								
Etc.									

Unless there is a very small number of business functions represented, populating this matrix is best done incrementally. For each increment, gather knowledgeable experts for a particular business function and perform the following steps:

- Review the subject area model to ensure the participants understand how the subject areas have been defined. Read aloud each subject area definition.
- Review the functions on the matrix to validate that the group understands and has knowledge of each of the functions.
- Population can be done by functional area or by subject area. If this is being developed top-down, then a functional approach can work well. If this is being developed bottom-up, then a subject area approach will generally work better. With the subject area approach, the first subject area is selected based on what is needed to support a project.
- Starting with the first function (of the functions relevant to this group), ask the following for each subject area:
OR
- Starting with the first subject area, ask the following for each function:
 - Does the function use this kind of data?
 - If yes, what is the nature of that usage? Does this function CREATE new instances of that type of data? Does it UPDATE the contents of that type of data? Does it merely READ or access that data? Does it DELETE instances of that type of data?
- Place the notations C, R, U, or D on the matrix as appropriate.
- Repeat this process for the remaining functions to be addressed by the group.

The complete table should look similar to the sample in Table 2.

Table 2: Sample Matrix Populated

		Customer	Product	Order	Contract	Part	Supplier	Employee	Financials
Finance	Fin Func 1	C			C,R				C
	Fin Func 2	R		U	C				C
	Fin Func 3	U			U			U	C
Sales	Sales Func 1	R		C		R			
	Sales Func 2				U			U	
Supply Chain	SC Func 1		C,U			R,U	R		R
	SC Func 2		R	R			C		
	SC Func 3						U		R
Etc.									

Once the matrix has been fully populated, it can be used to determine functional representation on the data stewardship teams. There should be a data stewardship team for each subject area. That team should be populated as follows:

- For each function that Creates, Updates, or Deletes data in that subject, the stewardship team for the function should have process owners from those functions.

Process owners are the people that drive the data stewardship process for a given subject area. They are responsible for scheduling working sessions, developing definitions, monitoring data compliance, and publishing stewardship results and working products. Process owners must be represented at all data stewardship meetings and working sessions.

- For each function that uses data in that subject, the stewardship team for that subject should have participants from those functions.

Participants are other interested parties who do not have the responsibility for driving the stewardship process for a subject area. They are invited to participate in all working sessions, and may be consulted for additional information if required.

PARTNERSHIPS

Data quality management requires a concerted, cooperative effort by people throughout the organization, and partnerships are critical. These partnerships include the commonly recognized ones between the information technology groups and the business units. In addition, information technology groups must partner with each other and business groups must partner with one another. Vertical partnership is also important. The message needs to be clearly communicated up and down, and across, the organizational structure.

FOUR PHASE PROGRAM

For each data subject area, you can then proceed through the four phase approach previously described. Armed with information from the data profiling activities, you can identify the most significant opportunities for data quality improvements within a particular data subject area. Focus on these opportunities, and perform the data quality improvement activities in each data source as well as the data integration activities to combine the data. You are then in a position to look for opportunities to augment the value of the data so that your company can better serve its customers or make better, more informed business decisions.

TECHNOLOGY SUPPORT

The data quality management initiative will only be successful if it is pursued as a partnership and is supported by technology. The business and IT groups need to recognize their respective roles and representatives from different

business units must cooperate to resolve issues that cross organizational boundaries. Data quality management technology capable of supporting the functionality noted in the description of the four phase program should be acquired to improve the effectiveness of the program and to significantly lower the effort required to manage data as an asset. The reduction in the on-going resources is in line with using a capital investment to lower the manual effort associated with any task, and may go a long way towards improving the return on investment of the data quality management program by reducing the overall program costs over time.

Sophisticated software and methodologies exist today that can help you solve the challenges of data quality management. We recommend that you investigate these available tools before you attempt to and then apply the four-phase approach to data quality management.

CONCLUSIONS

Deploying a data quality management program is not an easy task, but the rewards are enormous. Deploying a disciplined approach to managing data as an important corporate asset will better position your company to improve the productivity of its information workers and to better serve its customers. This is no longer an option, particularly in today's competitive and regulatory climate. To move forward, the key stakeholders must be educated, a stewardship function implemented, and appropriate technology must be acquired. With these in place, the four-phase program can be effectively pursued.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jonathan G. Geiger
Executive Vice President
Intelligent Solutions, Inc.
P. O. Box 4857
Boulder, CO 80306
Cell Phone: (858) 603-1699
Fax: (858) 452-2799
Email: jgeiger@intelsols.com
Web: www.intelsols.com

View Slides