



A Comprehensive Approach to Data Quality Governance

Thomas C. Redman, Ph.D.

A Navesink Consulting Group White Paper

A shorter version of this paper was published in the CAiSE'05 Workshop Proceedings

Correspondence to:

Thomas C. Redman, Ph.D.
President
Navesink Consulting Group
732-933-4669

tomredman@dataqualitysolutions.com

www.dataqualitysolutions.com

A Comprehensive Approach to Data Quality Governance

Thomas C. Redman, Ph.D.

ABSTRACT

While most organizations will enthusiastically agree that data are key assets, they also acknowledge that these assets are in poor shape. Too often the data needed for critical decisions and operations are unavailable, poorly defined, out-of-date, incorrect, or otherwise unfit for use.

Further, traditional hierarchical organizational structures are ill-suited to managing data. Therefore in this paper we propose a comprehensive approach to data quality governance. Our model features six mutually-reinforcing components:

- A *Data Council* for senior leadership,
- A *Data Quality Staff* for managing meta-data processes and conducting certain staff functions,
- *Information Chain Management* for data created inside the organization,
- *Data Supplier Management* for data received from outside the organization,
- Specific roles for *the Information Technology department*, and
- Continuous improvement through chartered *improvement teams*.

Each component has proven its mettle in leading-edge organizations. Further, we believe that (in some industries at least) the time is ripe for creation of a “Chief Data Officer” position. This person will head the Data Quality Staff and drive implementation of other elements of governance.

KEY WORDS

Data, Data Quality, Governance, Chief Data Officer

INTRODUCTION AND SUMMARY

It is becoming increasingly clear that current hierarchical organizational structures are not well-suited to managing the organization’s “data” as valued assets. Support for this claim comes from many quarters—here we cite two:

- The abysmal condition of the data assets. Most organizations cannot answer even the most basic questions, such as how much they have, how much new data they create, which are most important, or who manages them. People cannot find the data they need, and the data are poorly defined. Too much data are inaccurate and sources disagree, leading people to mistrust the data.
- The observation that management structures and data flow are misaligned. Most “management” is up-and-down the organization chart (vertically). But data flow horizontally across the organization, from person to person and department to department. In most organizations these horizontal flows are unmanaged.

Virtually every organization is bedeviled by poor quality data and the cumulative cost is enormous.

Further, the impact is enormous. Some of the most important issues of our time, from international security, to financial reform, to improved health care, stem, at least in part, from poor quality data. Virtually every organization is bedeviled by poor quality data and the cumulative cost is enormous.

In this paper we propose a comprehensive approach to data quality governance. The approach features six interrelated components:

- A Data Council, charged with aligning data to the organization's most critical strategies and pressing needs and addressing the organizational, political, and social issues that hinder the effort.
- Data Quality Staff, charged with administering the overall data program and "owning" so-called meta-data processes.
- Information Chain Management, charged with ensuring that internally-created data are of high quality.
- Data Supplier Management, charged with ensuring that the data the organization acquires from the outside are of high quality.
- The Information Technology Office, charged with ensuring that provisions for data quality are "built into" new systems, databases, and technologies.
- Improvement teams, chartered by any of the above to fix specific problems and/or pursue specific opportunities to improve data.

Each of these components has already proven its mettle in leading-edge organizations. The proposed governance approach synthesizes these components into a powerful whole, each component supporting and reinforcing the others. Even more importantly, roles and responsibilities for managing data are clearly defined and assigned to departments that are capable of carrying them out. This last point, involving departmental capability, bears mention because so many organizations have improperly assigned "data" to their Information Technology Departments.¹ But IT neither creates nor uses much data, so it is singularly ill-suited to the task.

Before introducing our approach to data quality governance, we first review pertinent features of the *nature of data in organizations*. These comments segue into a discussion of several salient points that directly impact data quality governance. As an example, we pointed out above that data flow horizontally across the organization. So it stands to reason that the most important of these data flows must be explicitly managed.

We then describe what we hope to achieve with the proposed governance approach. Ultimately, the goal is to have exactly the right (and correct) data in the right place at the right time any time data are needed. This vision is compelling, but it is indeed a tall, perhaps unachievable, order. So we also set some nearer-term objectives.

We then introduce our approach to data quality governance. We describe the most important roles each component plays and how these components fit together. As a practical matter, it would be extremely difficult to implement all components at once, so we suggest some ways that organizations can get started.

Importantly, we believe that a sea change is taking place in some industries. For at least a decade, every organization would readily admit that "data are among our most important resources." And many recognized the problems cited above. But few devoted any real attention to data. For the past year or so we have seen increasing signs, in some industries, that organizations see a new

For the past year or so we have seen increasing signs ... that organizations see a new urgency in data generally and data quality specifically.

¹ Herein we use the term "organization" to mean the enterprise of interest. It could be a corporation or government agency or a portion, such as a Business Unit, thereof. We use the term "department" to mean any lower-level portion of the organization.

urgency in data generally and data quality specifically. We believe this will lead to creation of “Chief Data Officers” with roles similar, in many respects, to Chief Financial Officers. Thus the Chief Data Officer will lead the Data Quality Staff and drive implementation of the governance model discussed herein.

OBSERVATIONS ABOUT “DATA” IN ORGANIZATIONS

Motivation, requirements and guiding principles for our governance model stem from many quarters.

Poor quality data lie at the root of issues that capture our collective attention and will not let go.

First, while virtually everyone agrees that “data are the key asset of the Information Age,” data are perhaps the organization’s most underutilized asset. We already noted that most organizations cannot answer the most basic questions about their data, never mind routinely use them to create value. The poor quality of an organization’s data further underscores this point. Too much data are poorly defined, inaccurate, out-of-date, or otherwise unfit for use.

We’ve already noted that the impact is enormous. One need not work through the intricacies of IT architectures to understand this point. One need only read the newspaper. Poor quality data lie at the root of issues that capture our collective attention and will not let go (Redman, 2004). A partial list includes:

- Intelligence failures,
- Financial reporting,
- Census undercounts and overcounts,
- The year 2000 Presidential election, and
- The bombing of the Chinese Embassy in Kosovo.

Fortunately, most data quality issues don’t make the front page. But virtually every organization is bedeviled by bad data—increasing cost, angering customers, compromising decisions, and making it more difficult to align departments. One study puts the costs to American business at over \$600B/year for *customer data alone* (Eckerson, 2002). And, “data quality” in some form or fashion is almost always cited as the number one or number two issue hindering successful implementation of new technologies, from data warehouses, to enterprise systems, to data mining.

Leading-edge organizations have made and sustained two orders of magnitude improvement and reduced the costs of poor data quality by two-thirds. They have done so by following the prescriptions offered above (adopting the philosophy of preventing errors at their sources, minding the social issues, proper management accountability, focus, measurement, etc.). They have connected data customers and data sources. For data created internally, they have managed and improved their most important information chains, in effect bridging the “white space” between departments. They manage their so-called “meta-data processes”—data modeling, data standards, and data cataloging. For data obtained from the outside, they have implemented supplier management with their most important suppliers. Finally they have a relentless focus on improvement—both on eliminating root causes of erred data, and addressing new data needs with new data (Redman, 2001).

Our first observation about the nature of data in organizations is that virtually every activity, from responding to a simple customer request to setting strategic direction, involves data. All sorts of data for all sorts of purposes. Herein, we consider “data to be of high quality if they are fit for their intended uses, by customers,² in operations, decision making, and planning.” (Redman, 2001). According to this definition, those using data are the ultimate arbiters of quality. And data customers can have a stunning range of needs for data. Some require “up-to-the-minute” data, others want data to be quite detailed, and still others want data presented in a special format. All data customers want the data to be accurate, relevant to the task at hand, clearly defined, and readily available in an easy-to-understand format. (See English, 1999; Fox, 1994; Huang, 1999; Redman, 2001; and/or Wang, 1996 for discussions of the various “dimensions of data quality”).

...virtually every activity, from responding to a simple customer request to setting strategic direction, involves data.

Second, “Data” is itself a multi-faceted notion. Philosophers, computer scientists, and information systems engineers have debated just what “data” (and other Information Age assets such as “information,” “knowledge,” and “wisdom”) are for generations. We’ve selected the definition used herein (see Fox, 1994 and Tsichritzis, 1982) because it best exposes how data are created and used in organizations, which is essential if they are to be managed effectively.

Thus, data (or a data collection, data set, etc.) consist of two related components: a *data model* and *data values*. Data models are abstractions of the real world that define what the data are all about. Models define entity classes, the collections of things (physical or ideal) of interest, and attributes and relationships, which specify pertinent features of the entities. As an example, the reader is an entity. His employer is interested in the entity class “EMPLOYEES” and attributes such as NAME, DEPARTMENT, SALARY, and MANAGER.

The Internal Revenue Service is also interested in the reader, as a “TAXPAYER.” It is interested in some of the attributes that interest the Employer, but many others, such as INTEREST INCOME, that do not. The reader is, quite obviously, the exact same person, but each organization has distinct needs and interests, so their models are different.

This example underscores an important point. Data are subtle and nuanced. An EMPLOYEE and a TAXPAYER are similar, but EMPLOYEE data will not meet the IRS’s needs. To modify Mark Twain slightly, “the difference between the right data and the almost right data is the difference between ‘lightning’ and a ‘lightning bug.’”

“Customer” is another good example. Each organization defines “customer” in subtly different ways. Brokers call their customers “accounts,” consultants call them “clients,” department stores call them “shoppers,” doctors call them “patients” and so forth. Each organization can model “customers” differently, focusing on attributes that best help it define and execute its strategy and/or niche. And the data model becomes the organization’s *lingua franca*. But the potential for confusion is enormous because each department also tailors its definition of customer to suit its purpose. Thus the Shipping and Marketing departments define ‘customer’

² Note, as used in the definition of quality, the “customer” is a customer of the data. He, she or it may be an internal customer or an external customer. The organization also has customers and they need not be data customers. If it is not clear whether we are referring to an “organizational customer” or a “data customer,” we will explicitly use the term “data customer.”

differently. And when people within the organization speak different languages, communication is impaired.

These points are true for all data. With respect to governance, management must recognize and exploit the unique opportunities that data bring to create value. Management must also recognize and mitigate the unique challenges.

Returning to our definition of data, each attribute must have a well-defined domain of possible values. For example, all possible nine digit numbers defines the domain of possible values for SOCIAL SECURITY NUMBER (SSN). Data values are to be drawn from the domain and assigned to attributes for specific entities. A single datum is thus a triple:

Datum = <entity, attribute, value>
An example datum is:
<John Q. Public, SSN = 123-45-6789>

One further complication is that “datum” is an abstraction. Note that one never actually sees a datum. Instead, what one actually sees are *data records* on paper, via computer applications, in tables or charts, etc. And note that the same datum may be recorded and presented in a virtually unlimited number of formats

Now consider the variety of contributions needed if even a single datum are to meet the needs of even a single data customer.

- The datum must be modeled properly and clearly defined,
- The datum value must be current and correct, and
- The datum must be accessed from a database and presented to the customer (often through an application) in a way he, she, or it can understand.

The operations that perform each of these functions are separate and distinct. In particular, data modeling, creation of data values, and development of databases and application programs are distinct activities, performed by different people, usually in different departments. And any may take place outside the organization. The important consequence for data quality governance is that all must be aligned to a common goal, yet managed separately.

As a practical matter, we usually think of a “data customer” as an individual person or department. Meeting their needs individually, or one at a time, leads to duplication of data and disparate systems. The high degree of redundancy and the numbers of systems that “don’t talk to each other” attests to this point. While a certain amount of duplication may be desired, unfettered duplication and disparity is clearly not in the best interest of the overall organization. So governance must recognize the notion of “organization as data customer” and address its needs as well.

Importantly, practitioners have verified that eliminating the root causes of error, at their sources, yields superior results at lower costs.

Perhaps the most obvious observation about data is that organizations have so much of them. And they create more everyday. Virtually every activity creates data, just as every activity uses data. Estimates of the “doubling life” (the time it takes for the quantity of an organization’s data to double) range from six months to three years (Whiting, 2002). The high rates of data creation mean that attempts to find and correct errors in the data are ill-fated. There is simply too much. Instead quality must be managed at the sources. Importantly, practitioners have verified that eliminating the root

causes of error, at their sources, yields superior results at lower costs (Redman, 2001). This result seems to hold independent of the type of data, the rate of data creation, etc.

The next observation is the speed with which data get around, moving from one department to the next in the blink of an eye, and changing as they go.³ What began as a simple customer order for a teal sweater, winds its way through fulfillment, inventory management, and billing. The data are aggregated with other data as part of financial reporting. And, sooner or later, someone uses them to try to understand why teal has become more popular than aqua. This means that there can be many “data customers,” with different needs, far downstream of data creation. Further, some data leave the organization. Some are sold, some are in response to regulatory or other concerns, some are passed on to the organization’s customers as part of day-to-day operations, and so forth.

Interestingly, while some data have many uses, other data are never used at all. Yet organizations continue to create more, adding cost and diluting focus. Thus an important governance task is to align the effort, focusing on the data needed to meet the organization’s most important objectives.

Fourth, recall that data are intangible. They have no physical properties, which complicates measurement. Indeed, all important data quality dimensions are abstract and so are difficult to measure. For example, you can’t tell by direct examination whether most data are correct. And “that which doesn’t get measured, doesn’t get managed.” This problem can be addressed, but it takes skill and experience.

Intangibility contributes to an “out of sight, out of mind” feeling about data. After all, they are stored neatly away in modern databases (and hence the responsibility of IT). So there is a tendency to pay less attention to data, particularly given the other compelling and highly visible problems facing the enterprise. This problem is exacerbated by the lack of accepted methods for valuing data. Ultimately, value is decided in a marketplace. But most data are simply not for sale! Some rather basic tools help estimate a portion of the costs associated with poor quality data. But ultimately leadership must take action without an unassailable business case to do so.

Paradoxically perhaps, everything about data seems politically charged (see also Davenport, 1993 and Strassman, 1994). A question as basic as “how many characters should we allot to the ‘address field’?” generates passions that seem beyond all reason. Never mind issues of privacy, ownership, or the proper role for IT (Redman, 2004b). Data ownership provides an excellent example. People and organizations innately sense that “data leads to power in the Information Age.” So they try to acquire all they can get and don’t readily share with others. There are other organizational, social, and political barriers as well. These issues demand the personal attention of an organization’s most senior leaders.

**...everything
about data
seems
politically
charged.**

Finally, our governance model is based on what works. Leading-edge organizations have made and sustained order of magnitude improvements and greatly reduced the costs associated with poor data quality. They do so by preventing and greatly reducing the costs associated with poor data quality. They do so by preventing errors at their sources, addressing social issues, and

³ At a somewhat deeper level, the speed of data movements stem from the fact that data can be digitized, electronically transmitted, and copied at extremely low cost.

aligning management accountability. They connect data customers and data sources by managing and improving their most important information chains, in effect bridging the “white space” between departments. They manage their most important suppliers and they have created data standards. Of course they measure data accuracy and other important aspects of data quality. Finally they have a relentless focus on improvement—both on eliminating root causes of erred data, and addressing new data needs and new data (Redman, 2001 and English, 1999).

WHAT WE HOPE TO ACHIEVE

Ultimately of course, one desires “exactly the right (and correct) data, clearly defined and easy-to-read, in the right place at the right time to conduct an operation, make a decision, or set a strategy to advance the organization, at reasonable cost.” It is an easy vision to conceive—and who would ever object to any part of it?

- Since, as noted above, “data are subtle and nuanced,” why settle for data that aren’t directly to the point (i.e., the right data)?
- Unless there was no choice, who would ever use data values that aren’t correct and up-to-date? There is too much risk!
- Who would ever use data they can’t read or understand? How could they?
- Aren’t modern information technologies designed to ensure that data are indeed in the right place at the right time?
- Day-in and day-out, aren’t data bound up in everything people do? And isn’t the whole purpose to help the organization pursue its goals, operate effectively and efficiently, discover new niches, etc.?

Compelling as this vision may be, achieving it will take a long time. And it is unreasonable to expect that any organization could implement all components at once. So our approach to governance aims to help organizations evaluate their individual circumstances and implement components sequentially.

Most importantly, our approach to governance aims to help organizations consistently improve the quality of their most important data.

Most importantly, our approach to governance aims to help organizations consistently improve the quality of their most important data:

- Make it easier for people to find, read, and interpret the data they need to do their jobs.
 - Improve the accuracy of data values one to two orders of magnitude.
 - Better understand its data assets and increase the value of those assets.
- Be aware of and manage the organizational, political, and social dimensions. In particular to bridge the gaps between IT and “the business.”

OUR APPROACH OF DATA QUALITY GOVERNANCE

A summary graphic of our approach to data quality governance approach is presented in Figure 1. The figure features all components, the major roles each plays, and several of the most important interactions between components. Note that the improvement teams complete projects assigned them by “management” which in turn acts at the direction of the Data Council. Note that data quality governance is overlaid on the existing organization and each layer supports those below it.

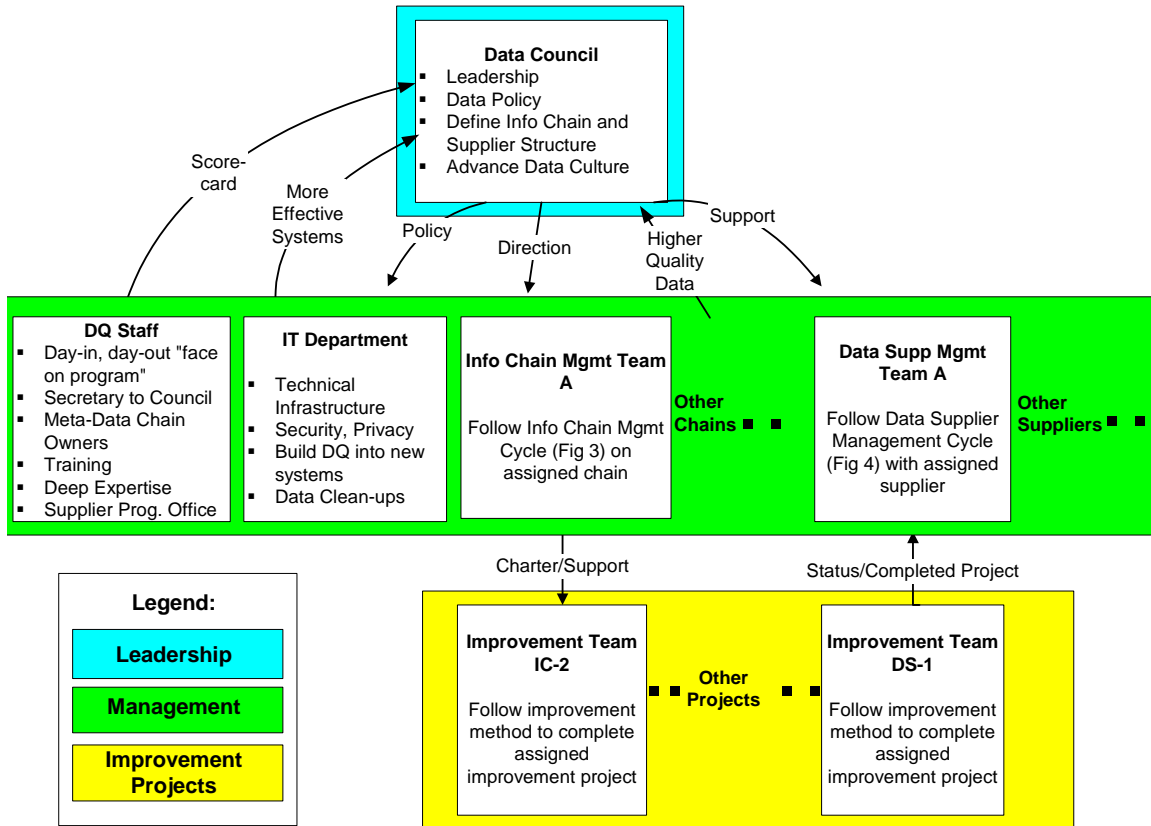


Figure 1
Our Comprehensive Approach to Data Quality Governance
© Navesink Consulting Group

The next sections describe each component and some of its major interactions with other components in turn. Ideally, we would like to present the components in a prescriptive order that instructs “build this component first, that one second, and so forth.” But each organization has unique needs and abilities. Some are “top-driven,” others “bottom –up,” still others “middle-out.” These factors should guide the order in which each organization establishes governance.

With this caveat, we’ll describe the components top-down.

Data Council

We now turn our attention to the most senior levels of the organization. Experience confirms that practically anyone can impact quality, within his or her span of control. And a few talented individuals can impact data quality within their spans of influence. But without senior leadership, these efforts almost always fall short. The seniority and position of the individuals perceived to be leading the data quality effort dictate its success more than any factor. So our proposed governance approach calls for a Data Council—the more senior and broader, the better. For organizations whose major products are data, the Chief Executive should head the Council.

The seniority and position of the individuals perceived to be leading the data quality effort dictate its success more than any factor.

Underscoring the need for the proposed Data Council, Dr. Joseph Juran (Juran, 1993), one of the foremost pioneers of the quality revolution in manufacturing summed it up this way,

“They thought they could make the right speeches, establish broad goals, and leave everything else to subordinates... They didn’t realize that fixing quality meant fixing whole companies, a task that can’t be delegated.”

And owing to intangibility, the vast quantities of new data, and the speed of data movement, leading the quality revolution for data may be even more difficult.

To be effective the Data Council must carry out certain functions. Per Dr. Juran, these may not be delegated. First, the Data Council must provide visible leadership. It must clearly articulate the business purposes and strategy for the data quality effort. Some organizations simply must reduce cost, others need higher-quality data so they can better execute a strategy of customer intimacy, still others sell data in an increasingly competitive marketplace that is demanding faster access to correct data. As part of the effort, the Council must determine the organization’s positioning of quality with respect to the competition. Obviously a leadership position within its industry conveys certain advantages to an organization. But “paving the way” also requires greater skill and effort. So a strategy that places the organization on the “leading but not bleeding” edge may better suit many organizations. Still others may elect a middle-of-the-road strategy. Importantly, if the history of the quality revolution in manufacturing is any guide, a strategy that assumes that data quality can be ignored is almost certain to fail.

One reason that the strategy and business purpose are so important is that they help focus the effort, in effect enabling managers to figure out, for themselves, which data and which dimensions of data quality are most important. Eventually people need guidance such as:

“Our most important data are the data that we send along to our clients. Right now they are complaining about errors. We simply must halve the number of errors clients see in the next twelve months.”

that they can translate into more specific actions. Thus the owner of an order fulfillment chain may conclude:

“The last step in my information chain is invoicing and customers complain that they are overcharged all the time. I’d better focus my efforts there.”

Second, the Council must develop and implement the Data Policy. This policy links the direction set and specific management accountabilities. We have noted several times that data quality is best managed at sources. So the policy must charge those that create data, including suppliers and owners of information chains, with ensuring that they do so correctly the first time. This part of the policy can be articulated as:

“Don’t pass bad data onto the next person. And don’t accept bad data from the previous person.”

The Data Policy must also address the roles played by IT and thornier issues such as data ownership, data sharing, and privacy. There are no silver bullet solutions to these issues. Unless the Council takes them up, chaos is sure to reign. We know of one Business Unit that has at least twenty-five overlapping but disparate copies of its most important data. It may be unrealistic (even undesirable) to expect “one gold copy.” But reducing the number to (say) five will reduce the problem by 80%.

Policies are much easier to define than they are to implement. So the Council should begin by benchmarking where it stands and each year or so, sponsor a review to evaluate progress. This can be done on a departmental basis, with those departments meeting their obligations being “certified” and those making the fastest progress singled out for special quality awards.

Third, the Data Council is responsible for the information chain and data supplier structure. It must identify the most important end-to-end information chains and data suppliers and put owners and management teams in place. The Council must hold them accountable to the Data Policy, and it must invest them with sufficient authority to make needed improvements. It is especially important that the levels of accountability and authority are commensurate with one another. Note that our governance approach does not call for a re-organization along information chain lines. So the Council must mate authority and responsibility in the face of a hierarchical organization chart.

The Data Council also bears the responsibility for advancing the “data culture,” a multi-faceted, amorphous, poorly understood topic (Redman, 2004b). The proposed governance structure called for here stands in stark contrast to the current lack of concern for data. And “advancing the data culture” involves no less than the Data Council getting people, entire departments, and the entire organization to think and act differently. We have no simple prescriptions for Councils beyond:

- Continually reinforce that “data matter.” Link data, strategy, decision-making, and business objectives as tightly as possible. Define and publish top-line targets for improvement and track progress toward those targets. Create and hold up role models. Create urgency about the topic.
- Adjust the compensation system to reward improvement.
- Be visible. One way to do so is to participate in a data supplier team. There are many others.
- Adopt a change management paradigm (e.g., Kotter, 1996) and follow it.

Finally, if people are to take on their new roles they need training. Our proposed governance approach calls for the funding of training to come from the Data Council.

Chief Data Office and Data Quality Staff

The second ingredient of our proposed data quality governance approach is a Data Quality Staff. The head of this staff should be given the title Chief Data Officer (CDO),⁴ but he or she is often affectionately known as the “Data Quality Czar” or “Czarina.” He or she reports to the most senior member of the Council. He or she leads a small staff of full-time employees and the Staff is empowered by the entire Council.⁵

One of the CDO’s most important jobs is Secretary for the Data Council. In that capacity, he ensures that the background work needed for the Council to be effective is complete. Thus the CDO:

- Develops the business case and strategy for data quality and works with the Council to integrate it into overall business plans.

⁴ It may seem logical to title this person the Chief Information Officer, Unfortunately most organizations have misused that title for the Chief Information Technology Officer.

⁵ Some refer to members of the Data Quality Staff as “stewards,” a term we do not favor.

- Translates the Data Policy into top-line measures of success, defines and assembles the Data Quality Dashboard, and helps the Council understand the results and their implications.
- Ensures that the Council stays on track by focusing it on its most important decisions.
- Manages the yearly “certification program” for Business Units, external suppliers, etc.

The CDO is the day-in and day-out “face” of the data quality program. She and her staff work throughout the organization to help others understand their obligations under the Data Policy, integrate data into their plans, and create “urgency” so the data program isn’t given short shrift.⁶

The CDO is the day-in and day-out “face” of the data quality program.

The CDO and his staff assemble the overall Data Quality Program. They are responsible for articulating the philosophy of preventing errors at their sources, and importing the needed techniques to manage information chains and suppliers, measure quality levels, conduct improvement activities, and so forth. One subtle, yet important role of the DQ Staff is to make these techniques as easy to use as possible. The apparently simple task of renaming the six-sigma DMAIC technique as, for example, “Our Bank’s Quality Improvement Cycle” is really much more subtle. It involves selecting techniques to best suit the organization, translating them into the organization’s language, and using the organization’s case studies to illustrate these techniques. The DQ Staff may document this work in the form of a handbook.

An important component of the Chief Data Office is the Supplier Management Program Office. It serves the following roles:

- It translates the overall Data Policy into policy relevant to data suppliers.
- It makes sure that a department is assigned to manage each of the most important data suppliers and resolves any issues that arise when two or more departments compete for a supplier’s attentions.
- It sets organization-wide goals for data supplier management and tracks progress against them.
- It assembles the overall Supplier Scorecard.
- It operates the Supplier Certification Program.
- It helps resolve especially difficult or important problems.

The DQ Staff either provides or arranges data quality training and education.

Data Quality Staff also builds deep expertise in all aspects of quality management so that these skills can be applied on especially challenging or important issues.

Finally, we noted earlier that the owners of meta-data chains may reside in DQ Staff.

Information Chain Management

A *process* is defined as “a set of interrelated work activities, usually characterized by specific inputs and repeated value-added steps, which produce a specific set of outputs.” Note that the term “process” embraces a wide variety of actions. One follows a process when one brews coffee. The “order fulfillment” process that starts with a customer order, continues with a credit

⁶ A somewhat more cynical view of this work is that many people resist new thinking, tools, and techniques. One frequent comment is “This doesn’t apply to us because...” So the work of the Data Quality Staff aims to mitigate the excuse.

check, concludes when the customer's account is properly credited, and includes all steps in between (such as delivery) is also a process. Finally the steps a large company follows to select new companies to buy, complete the transaction, and integrate the purchased company into the organization is also a process, even though it may take years to complete one cycle.

Note too that the term "process" is a relative one. "Taking a customer order" itself fulfills the definition of a process, though in the context of end-to-end order fulfillment it is simply a step or sub-process.

Technically, the focus here is only on the "data portions" of processes, as "data" are one category of inputs, value-added steps, and outputs. Others include raw materials, finished goods, services, cash, and so forth. But the data portion of a process, which is called an *information chain*, is often dominant. The order fulfillment process is a good example. Even the "brew coffee" process creates some data—at the very least it leads to a note "Time to order more coffee" when the supply is low. So as a practical matter, for many important situations, "data," "process," and "information chain" are (nearly) interchangeable. Figure 2 summarizes these points via the so-called "customer supplier model."

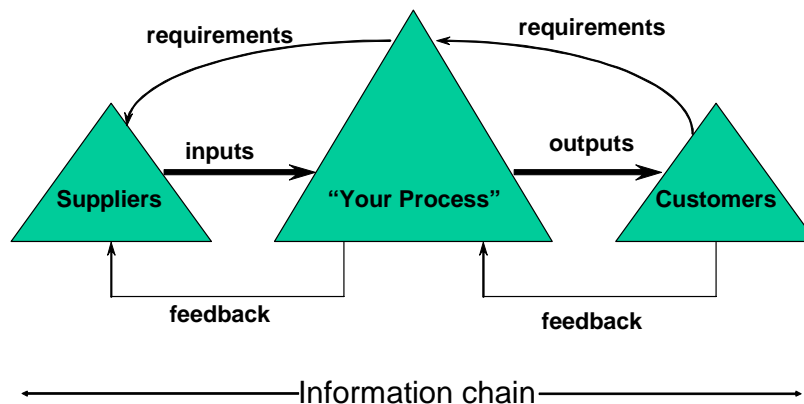


Figure 2
The Customer-Supplier Model

© Navesink Consulting Group, 2000

Our governance approach calls for the organization to explicitly identify and manage its most important information chains using the Information Chain Management Cycle

Our governance approach calls for the organization to explicitly identify and manage its most important information chains using the Information Chain Management Cycle (Figure 3) or its equivalent. The most important chains include:

- Those that parallel the flow of goods and services. The order fulfillment chain of a direct merchandiser is one such chain. For a shipper, the chain also starts with a customer request and tracks the movement of the container from truck, to boat, to rail, and site to site. For a hospital, the chain starts when a patient checks in to the hospital, follows her to her room, clinics, operating theatres, labs, and so forth.
- Those that yield a great amount of intellectual capital, including research and development, engineering, and marketing.
- Financial chains, i.e., those that involve money, including financial reporting.

- Any other chains critical to the organization.

Note that though these chains are critical to the business, no one department “owns” the data as they cross departmental lines. Our governance approach specifically does not call for re-organization.⁷ Rather Information Chain Management is overlaid on the current organization chart. A team of managers, drawn from the various functional areas has responsibility for the end-to-end performance of their assigned process. One especially senior manager, often called the “owner,” is assigned to lead the team. Assembling this team, naming the owner, and making sure its responsibilities and decision rights are clearly defined is step one of the Information Chain Management Cycle.⁸

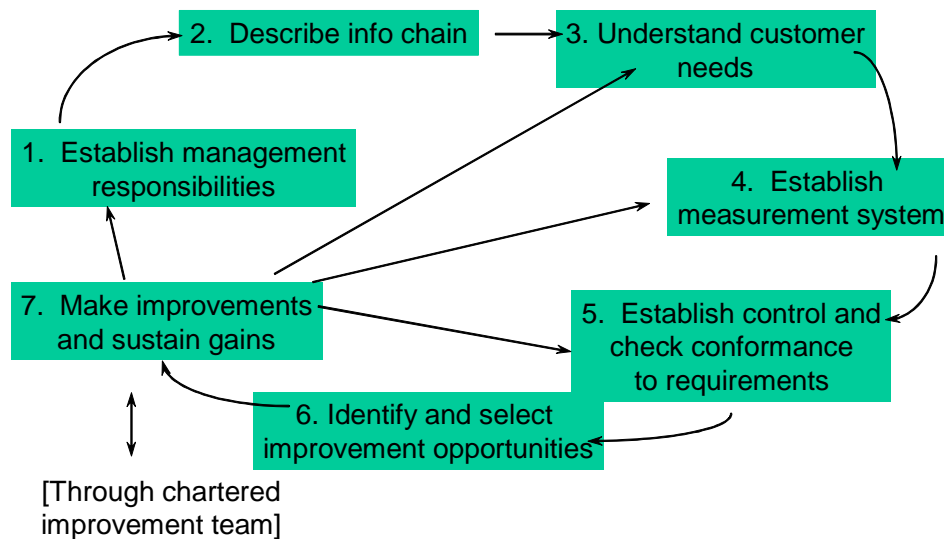


Figure 3
Information Chain Management Cycle

© Navesink Consulting Group

We do not wish to camouflage the difficulties in implementing Information Chain Management, as it represents a form of “matrix management.” This point is one reason the “Data Council,” discussed later, is so important.

In addition to the information chains cited above, there are three so-called “meta-data chains” that are especially pertinent to data:

- Data modeling which, as previously discussed, determines which aspects of the real world are captured in data. Today, most data modeling is conducted by the IT department.

⁷ Others, notably Hammer, believe that organizations should organize around process (Hammer, 1996). Hammer may well be right. Indeed, our argument that “the most important flows of data must be managed” supports Hammer’s position. But we believe that the hierarchical organization is unmatched at promoting functional excellence, which is a necessary ingredient into process excellence. Further, we have seen numerous examples where cross-functional teams effectively managed processes. So we are not calling for wholesale re-organization, yet anyway.

⁸ Organizations that already have a process management structure in place need not redefine their processes. They must however, make sure that process managers understand their responsibilities for data.

- Data cataloging which aims to develop and keep current an inventory of an organization's data. Today most data catalogs are also maintained by IT using data repositories. While repositories meet the needs of the systems development community, they are poorly suited to data customers.
- Data standards, which aims to define and agree on common definitions, code sets, and algorithms to be used across the organization. Relatively few organizations have the needed standards or standards chain.

Therefore, our governance approach requires that each of these three meta-data chains must be specifically identified, an Information Chain Management Team must be assigned, and the chain must be managed end-to-end. Doing so can be especially problematic, because they don't "fit" well anywhere. One could make the case that the owners of these chains should reside:

- In IT, as many aspects of the data modeling process are quite technical and the people who do this work currently report to IT.
- Within a Business Unit, as heavy involvement of "the business" is needed to ensure that the data reflect the real work of the organization.
- In DQ Staff, since "data" are its full-time job.

For now, we recommend that the owners of these three chains report within the Data Staff and that the management chains from each be drawn from all appropriate departments (Pautke, 2004).

Data Supplier Management

The fourth component of our governance approach is Data Supplier Management. It seeks to satisfy the same ends as Information Chain Management, namely to ensure that the needs of data customers, far removed from data creators, are met. The most important difference is that data creators lie outside the organization.⁹

Our governance approach calls for organizations to specifically identify their most important data suppliers. Suppliers of purchased data (credit reports, market data, engineering specs, etc.) fit this criterion. Other important data, such as provided on invoices, may come bundled with or as a byproduct to supplied product and service.

**...part of the
supplier
management must
reside in the
business**

Note the similarities between the Data Supplier Management Cycle (Figure 4) and the Information Chain Management Cycle. Step 1 calls for defining specific management accountabilities. Many organizations have a "contracts" or procurement group" that resides in staff and handles business arrangements with suppliers. It cannot be expected to have the deep expertise needed to understand how the nuances of how the data are used. So part of the supplier management must reside in the business.¹⁰ Thus the most important customer of an individual supplier leads day-in, day-out management for the entire organization. In some cases this is easy—accounts payable has always been responsible for ensuring that invoices are correct before it pays them. Using the supplier management cycle the focus shifts from finding and fixing errors to receiving correct invoices in the first place.

⁹ Many organizations find it easier to implement Data Supplier Management than Information Chain Management.

¹⁰ As we shall discuss shortly, there is an important role for Data Quality Staff as well.

Even when many departments need the data provided by a single supplier (as an example, many departments may use data supplied by credit agencies), it is usually straightforward to determine which is in the best position to manage the supplier. When contentions arise they should be handled by the Supplier Management Program Office, which resides in Data Quality Staff.

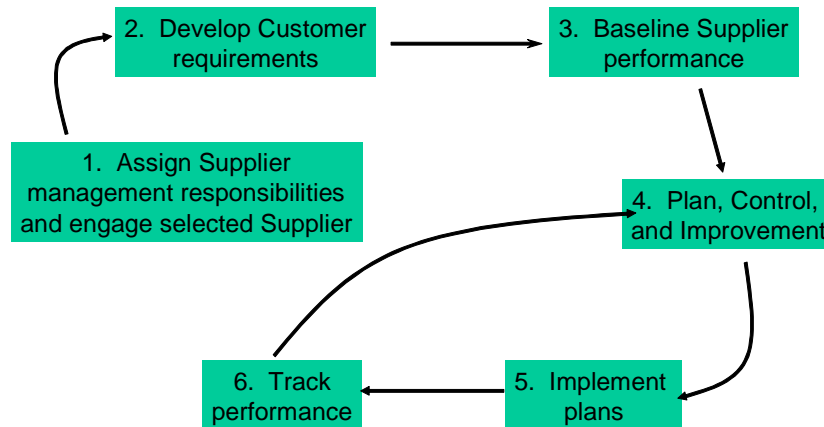


Figure 4
Data Supplier Management Cycle

©Navesink Consulting Group

Importantly, Data Supplier Management is effective within the organization as well. The financial reporting chain is a good example. All financial transactions are inputs. So it is perfectly appropriate for the financial reporting chain to treat the order fulfillment chain as a supplier.

Information Technology Office

Next consider the roles of the IT department. Previous sections of this paper make it clear that the most critical issue that governance must address is proper management accountability. Some organizations, perhaps noting that “all the important data are in the computer,” blithely assume that the IT department is responsible for data. Yet, as we have also noted, IT neither creates nor consumes much data, so it is a singularly poor choice.

That said, IT clearly has critical roles to play. First, IT is responsible for the “technical infrastructure,” including the organization’s databases, its communications’ infrastructure, and computer applications. This is no small task, as the quantities of data being created and stored are growing exponentially. There are dizzying arrays of available technologies, and more are invented all the time.

Second, the job of implementing security and privacy policies often falls to IT.

These are both daunting and thankless tasks. One reason they are so daunting is because a body of law and common practice are yet to develop. A second reason is that the risks are so poorly understood and hackers are so ingenious. At the same time, legitimate customers want faster, easier access through more devices. Wi-Fi is but the most current example.

Third, IT clearly has a role to provide (or purchase) applications to help business processes perform faster, more economically, and with greater capacity. From the data perspective, these applications:

- Must ensure that people can access data and understand them and
- Help those who use them create high-quality data for later data customers.

These in turn, require:

- Alignment between process and application.
- That the tools needed to create high-quality data be “built into” the application. For examples,
 - Edit checks that help people correct their own mistakes,
 - Measurement capabilities to help the Information Chain Management team know how well customer needs are being met and identify ways to improve the process,
 - Utilities that help users quickly and easily learn data definitions.

It is important not to underestimate how challenging alignment can be. The current trend in applications is toward so-called “enterprise systems,” such as Customer Resource Management (CRM) and the failure rate of these systems is alarmingly high. Not surprisingly, “data” and “alignment with process” are often cited as the two most important contributing factors for these failures (Friedman, 2004).

A fourth task for IT is implementing data clean-ups, meaning finding and fixing erred data, changing formats, adding new fields, etc. Of course data clean-ups are the “task of last resort,” but sometimes new customer needs require them. The task is best led by IT. Organizations are well-advised not to conduct data clean-ups unless the information chains creating new data are of high-quality (there are rare exceptions to this rule). But some argue for immediate clean-ups and fixing the chain later. The trap is that the clean-up addresses the immediate problem and the underlying issues are not resolved. When this occurs, subsequent clean-ups are inevitable. Thus our governance approach calls for IT to lead one-time clean-ups only.

... Thus our governance approach calls for IT to lead one-time clean-ups only.

Data Quality Improvement Teams

The heart and soul of any Data Quality Program are efforts to improve. Our proposed governance approach calls for continuous, project-by-project, improvement. Over the years literally dozens of techniques for data quality improvement have been used. In general,

- The Data Council, Information Chain Management Teams, Data Supplier Management Teams, and Information Technology may “charter” improvement projects to solve specific problems and or pursue specific new customer needs.
- Project teams, with broad representation, are formed and can negotiate their charter with the sponsoring group.
- Teams follow a predefined problem-solving process. There are many such processes and today the six-sigma Define, Measure, Analyze, Improve, Control (DMAIC) process is popular.
- Teams are disbanded upon completion of their projects.

It sometimes happens that continuous improvement is too slow. Therefore our proposed governance approach also recognizes re-engineering as a means to completely redesign

information chains. Since the scale of a re-engineering project and the attendant risk is so high, re-engineering teams must be sponsored by the Data Council.

GETTING STARTED AND ADVANCING

We do not expect organizations to implement data quality governance all at once. Instead, they can implement a component or two at a time, learning as they go. For example, an organization starting from square zero may elect to start in several different ways:

- An organization that is dependent on external data may start a data supplier program.
- An organization that is implementing a new enterprise system may start by rationalizing the responsibilities of “IT vs. the business.”
- An organization that is experiencing problems with revenue collection that it suspects begins far upstream may start by implementing Information Chain Management on its “order fulfillment chain.”
- An organization whose senior management is frustrated by recurring data issues may start by implementing a Data Council.
- And so forth.

Organizations that have already implemented one or more element can use our approach to governance to plan their next steps. As examples, organizations can use this approach to:

- Help understand and address internal politics that have slowed an existing quality program.
- Help address inadequacies, gaps, and shortcomings in current efforts such as six-sigma improvement programs, stewardship, and data management by building the missing components.

Eventually of course, organizations must implement all, or nearly all, of the governance model if they are to fully address their needs for high-quality data.

Eventually of course, organizations must implement all, or nearly all, of the governance model if they are to fully address their needs for high-quality data.

FINAL REMARKS

We began this paper with the claim that the traditional hierarchical organizational structures are ill-suited to managing data and we explained why. We then discussed the “nature of data in organizations” and their implications for data quality governance.

Our proposed governance approach calls for six mutually reinforcing components. Thus “leadership” is provided via a *Data Council*, “management” through *Data Supplier Management*, *Information Chain Management*, specific roles for *the Information Technology department*, and a small *Data Quality Staff*, and “improvement” through *improvement teams*.

Each component has proven its mettle in leading-edge organizations. And collectively, they provide governance top-to-bottom and across the organization. This approach to governance “works” because:

- It helps identify the most important data and focuses the effort. “Data” and the organization’s most important issues become linked to one another.
- It clarifies management accountabilities, specifically assigning those who create data responsibility for the quality of what they create. Conversely it helps ensure that those who are ill-suited for certain tasks are not assigned.

- It recognizes that data flow horizontally, across departments and organizations, and ensures that the most important of these flows is specifically managed.
- It provides needed structure so that information chain and supplier managers and improvement teams get the support they need.

Our governance approach is not (yet) intended to replace the current organizational structure, rather it is designed to be overlaid on it.

REFERENCES

Davenport, Thomas H., *Process Innovation: Reengineering Work Through Information Technology*. Boston: Harvard Business School Press, 1993.

Eckerson, Wayne W., "The Data Warehousing Institute Report Series, Data Quality and the Bottom Line, Achieving Business Success through a Commitment to High Quality Data," 101communications, LLC, 2002.

English, Larry P., *Improving Data Warehouse and Business Information Quality*, New York: John Wiley & Sons, 1999.

Fox, Christopher J., Levitin, Anany V., and Redman, Thomas C., "The Notion of Data and its Quality Dimensions," *Information Processing and Management*, 30, No 1, p. 9-19, January 1994.

Friedman, T., "Data Quality 'Firewall' Enhances Value of the Data Warehouse," *CIO Update*, Gartner Research, April 13, 2004.

Hammer, Michael, *Beyond Reengineering*, New York:Harper Collins 1996.

Huang, Kuan-Tsae, Lee, Yang W., and Wang, Richard Y., *Quality Information and Knowledge*, Upper Saddle River: Prentice Hall PTR, 1999.

Juran, Joseph M., "Made in USA: A Renaissance in Quality," *Harvard Business Review*, (July-August 1993):47.

Kotter, John P., *Leading Change*, Cambridge: Harvard Business School Press, 1996.

Pautke, Robert W. and Redman, Thomas C., "Understanding Meta Data: The Modeling, Cataloging and Standards Process," Unpublished manuscript, August 2004.

Redman, Thomas C., *Data Quality: The Field Guide*, Boston: Digital Press (Butterworth Heineman), 2001.

Redman, Thomas C., "Data: An Unfolding Quality Disaster," *DM Review*, August 2004, p. 21.

Redman, Thomas C. (b), "Barriers to Successful Data Quality Management," Contributed chapter in *Studies in Communication Sciences*, Volume 4, Number 2, December 2004, pp. 53-68.

Strassmann, Paul A., *The Politics of Information Management*, New Canaan, CT: The Information Economics Press, 1995.

Tsichritzis, D.C., Lochovsky, F.H, *Data Models*, Englewood Cliffs, NJ: Prentice Hall, 1982.

Wang, R.Y., Strong, D.M., "Beyond Accuracy: What Data Quality Means to Consumers," *Journal of Management Information Systems*, Vol. 4, 1996, pp.5-34.

Whiting, Rick, "Tower of Power," *Informationweek.com*, February 2002.

<http://www.informationweek.com/story/IWK20020208S0009> Link last checked 6/17/05



Dr. Thomas C. Redman is President of Navesink Consulting Group, based in Little Silver, NJ. Known by many as the "Data Doc," Dr. Redman was the first to extend quality principles to data and information. By advancing the body of knowledge, Tom's innovations have raised the standard of data quality in today's information-based economy.

Tom has consulted with many organizations in a wide range of industries. His proven, repeatable tools, techniques and roadmaps have helped clients in telecommunications, financial services, computer products, dot-coms, logistics, consumer goods, and government agencies understand the importance of high-quality data.

An internationally recognized author and speaker, Dr. Redman has authored numerous papers and published three books, *Data Quality: The Field Guide*, (Butterworth-Heinemann, 2001), *Data Quality for the Information Age* (Artech, 1996) and *Data Quality: Management and Technology* (Bantam, 1992). He was invited to contribute two chapters to *Juran's Quality Handbook, Fifth Edition* (McGraw Hill, 1999). Dr. Redman holds two patents.

Navesink Consulting Group
34 Sycamore Avenue, Suite 2H
Little Silver, NJ 07739
732-933-4669 (T) / 732-576-8910 (F)
www.dataqualitysolutions.com
info@dataqualitysolutions.com