# A Methodology for

# Establishing and Maintaining

# Quality in Data Context

Presented at the
1999 Conference on Information Quality
MIT Sloan School of Management

Robert Tap
Surprise Information Corporation

October 1999

# Introduction

This paper describes "A Methodology for Establishing and Maintaining Quality in Data Context" which was developed by the Surprise Information Corporation. The methodology uses Maslow's *Hierarchy of Needs* as a metaphor to establish a continuous quality improvement process and for communicating with management about the information needs of new business processes. When adopted as a methodology, the use of the metaphor results in the improvement to the quality of information used throughout a large organization for a multitude of business purposes. The goal is the achievement of a clearer understanding of business issues, problems, and potential solutions for management action through understanding the data holdings of an organization better.

The paper describes an approach to holistic thinking about information quality. To improve the value of the data holdings, it suggests how to interact with management to implement and use a continuous quality improvement process (Deming Notion). The emphasis here is the development and use of a process that gives senior management insight based on high quality information so it can better weigh alternative business decisions about investing in the organization's information infrastructure. Such investments in infrastructure should not be limited to hardware and software. It must include time and money directed to identifying and understanding *sources of information quality problems* that inhibit the performance of an organization. Raising the quality issue to senior management requires providing them with, a focused context in which to discuss problems. The methodology expressed here can provide that context.

# Objective of This Paper

The need for such a holistic view has been expressed in earlier years of the conference.

The driving force seems to be the difficulty of comprehending the volume of data errors compounded by a lack of a relationship of these errors to business needs; thus making it hard to put a value on the errors. Without values, all errors appear to be equally bad and correcting the situation becomes an impossible task.
The objective in learning this metaphor is to provide a common language and framework to understand and communicate the results of a data quality analysis. Management and the development team need a basis to discuss data quality. Such discussions permit business decision makers to deal effectively with alternatives, to focus on results, to avoid misunderstandings, and to implement new strategies more quickly. If that exists, then they can begin to have a clear understanding of the business issues, problems, and potential solutions the current data quality levels imply. This understanding helps them to join together with IT in developing new, advanced and/or complex business and information systems.

# Approach to Answers Through Analysis

The focus of the analytical method that underpins the metaphor is the practical and continuous use of the total data set used in a business to explain the performance of that business. In order to create such an explanation, high confidence in the quality of the data that describes that business must exist. This leads to the requirement of a methodology to analyze the quality of the business data. With the creation of a framework (context) within which to review and characterize all data elements, a holistic data quality view emerges. With this holistic view, business issues and problems can be joined to the discussion to appreciate the potential values for various improvement recommendations.

The analytical output of the framework easily translates to maps that can be used to lead discussions with business area managers, to navigate data quality issues, and to understand how to read a variety of performance measures. Visually based problem solving of complex issues like these permit absorption of data that might be otherwise hard to comprehend; that is to say, a picture can help to convey an idea. Thus, using holistic ideas, frameworks, and maps as a foundation for discussion plays to those who deal well with visually based approaches to problem solving.

To put the paper into perspective, let us review the ideas we are examining:

1. The field of work we are discussing arises from the increased use of Knowledge Discovery Through Data Mining tools. Many times the answers have *puzzling characteristics* that are traceable to quality problems in the data used.
2. The subject we are discussing is a *data quality analysis capability* designed to develop issues that require management action.
3. The target audience for this approach is anyone having a data warehouse operation or a very large database operation and is responsible for *reporting to management* on data quality issues.
4. The initial focus is on *quality of performance measures* as a means to understanding the information that is extracted from the data resource.
5. Either a warehouse or straight database can be used as *source* material.
6. The approach, while explained in a top down manner, is performed as a bottoms up, *rapid prototyping information analysis effort.*
7. All work is *outcome focused.*
8. The aim is *to improve* one's business capability in through enhanced decision-making.

# Establishing the Language for Discussion

**How do you usually think and talk about data at work?**

Discussions about data are often filled with technical jargon that can make a business manager's eyes glaze over. Terms like bits, bytes, bandwidth, key, and index; packets, addresses, connectivity, and message abound. Frequently, context is completely absent. The person(s) who is supposed to be informed by the data is not being informed. He has been provided with a sealed box of tools and no key.

Technical people are often like your telephone company. They help you place a call, but they don't get involved with what you say in your phone conversation. But for people in business, CONTEXT is EVERYTHING. Thus, the first thing needed is to learn a context language for discussing a quality process for your data. The operational metaphor for the language is based on the work of A.H. Maslow's, "A Theory of Motivation" with which many professionally trained managers are likely to be familiar. It is sometimes called the Hierarchy of Needs.

You may recall that Maslow's theory can be applied to individuals, a community, or other organizational environments.

There are five levels in the Hierarchy of Needs.

### I. Physiological

Maslow tells us that at the base level of existence, the **physiological level**, one must deal with and resolve survival needs of the most fundamental kind, namely survival of self and the species, before one can focus on issues at the next level. So man dealt with getting enough food to eat by hunting and gleaning and then, after many centuries, turned to agriculture, and brought his initial technology along in the form of pottery and some shelter to create a sustainable environment.

### II. Safety

With the attainment of a relatively survivable environment, the next level of problem involves insuring **safety** by figuring out ways protect oneself from physical harm. Walls to keep animals out and groups of mud cottages and wood houses to deal with rain, snow, and cold grew from campsites and villages to stone forts and castles with complex walls.

### III. Belonging

When an individual was a member of a village, he was expected to take on responsibility for various things as part of **belonging** to the community. When they performed well, they received the acceptance and friendship that can be generated in the community.

### IV. Esteem

Individuals would receive **esteem** from others in the form of status and praise. This in turn improved the self-esteem of the individual and motivated them further. If individuals did not perform well, they faced the perils of physiological existence alone.

### V. Self-actualization

Finally, those individuals who felt strongly about belonging to the group and had strong self-esteem would become extremely capable at whatever talents they had developed. These talents were offered to the community in many forms from scribe or soldier to ruler or king. This **self-actualization** of capability is the highest level that Maslow described. (It was encapsulated in one of the most successful advertising campaigns in America, namely the US Army's *Be All You Can Be, Joint the Army* recruiting ads).

So, how does all of that relate to an analytical method to understand data quality? Current thinking says, "Data is an Asset." So are the people who handle the data. Technology and data supports the development of people and organizations. Process ties this all together.

The methodology described in this paper creates tests and evaluations to qualify an organization's data at five levels of performance capability that parallel to those described by Maslow. If the data cannot qualify well at a base level, it is unlikely that results prepared from that data for higher-level purposes will be truly useful.

Let us continue our language development. At the equivalent **physiological** stage, one performs an **Adequacy Analysis** of the Physical Data. It measures the most basic capability of your data to produce knowledge; it is the converse of the old garbage in/garbage out notion.

For the **safety** stage, one performs a fairly standard **Quality Analysis** of the Physical Data along single tread lines.

**Belonging** relates to an analysis of the capability of the data to support an Analytical Framework (adding infrastructure) by performing an **Options/Connectivity Analysis**.

If one is comfortable with the preceding results, **esteem** can be developed from the data in terms of presenting its **Meaning and Interpretation Relationships.**

Finally, **self-actualization** or Knowledge emerges through the **"Eureka"** discovery process. This is finding an "outside the box" answer and requires some outside the box thinking. One might read Tom Peter's "The Pursuit of Wow" to understand the thinking and analysis required at this level. It is in this analysis where new relationships in understanding business problems and solutions emerge.

# Obtain Data about the Data (Testing)

To execute the analysis and determine if your data can qualify its performance at each level, specific analytical tests have been created.

At the lowest level, an **Adequacy Analysis** of the Physical Data is performed through **tests of completeness and emptiness.** These tests are applied separately and sequentially to each and every table/file that comprises the corporate database. There may be several hundred tables within the database. Each table is under investigation at this point because the quality of each is related to the source in the most fundamental way (i.e. you purchased it, you recorded it, or you built it from other source material).

If the results of the first test are satisfying, a basic **Quality Analysis** of each and every one of the physical data elements is performed. Again, we are talking about hundreds of elements. The analysis produces **information on ranges, means, and distribution of occurrence patterns.** At this point a fundamental understanding has been learned about every row and column that is potentially possible in your corporate database.

At the third level, the weaving together of tables and data elements begins in its most simple form. An understanding of the results from an **Options/Connectivity** Analysis **derives basic relationships** between records, tables, and files of time, space, and/or hierarchy and begins to establish the Analytical Framework. These relationships are asserted and tested through construction of basic and logical performance measures. Your business knowledge experts review the results of these measures to determine their truth and validity.

If one is comfortable with the preceding results, more complex data analysis is undertaken in terms of presenting **meaning and interpretation** that is supportive of the business and its desired outcomes. Again, business experts review the resulting statistics or performance measures for understanding and representations. This fourth stage is the final stage seeking confirmation of beliefs and understanding of the experts. Here too the various contradictions observed while doing the confirmation work are reviewed and discussed to gain insight about formulation of the analysis in the fifth stage.

Finally, Knowledge emerges through the unveiling of new relationships, those not thought of earlier. Called the **"Eureka"** discovery process, this analysis seeks new hypotheses that might be suggested by the contradictions. **New hypotheses** are raised and relationships **tested**. Maps showing paths that can be logically followed (or not) are drawn to find those that would likely produce an analytical conclusion that is not a foregone conclusion. Thinking outside the box is necessary to discover and suggest relationships. Emerging software is beginning to support such thinking. One is **seeking discovery** of something new rather than just the confirmation of a hypothesis.

# Test of Completeness and Emptiness

A completeness/emptiness test is done for each and every table/file of data under review.

Begin with a review of the technical specifications and documentation of the database that is the subject of the data quality testing. This is the theoretical statement of what should be contained in the data under review.

It is considered a theoretical statement, as it seems it is seldom tested in the real world and only reviewed on an exception basis. Often this exception arises from the observation of a few odd errors and someone's request it be looked into. It seems to the IT department that these occasions frequently arise in demonstrations of new technical capabilities to senior managers.

Count the number of records in a table for which all data elements are present. Then compare it to the total number of records in the table and present the result as a percentage of total records that are complete.

Next, count all records in which at least one data element is missing and record the position in the record of the missing elements. This will create the pattern of emptiness. If one has n data elements in a record, one has $2^n$-1 possible error patterns. For 10 elements, this yields 1023 possibilities.

Surprisingly to many, normal results of an analysis show only a dozen or so emptiness patterns. This is because the emptiness is not the result of random error. Instead, specific and identifiable actions are the source. The resulting emptiness patterns are reviewed to discover the logical source(s) for their existence. In a recent study, a single pattern of emptiness emerged. It was learned that the source of the emptiness was the inability of the staff member assigned to purchase the data to be able to afford a complete table. Thus, the staff member suggested the most frequently used records be complete and the others contain only incidental information. Unfortunately he forgot to tell anyone of his decision. Several analyses later it was learned their results were extremely misleading. By this time, the corporation had made a significant investment based on those results. If a database has a high level of completeness and a small number of patterns of emptiness (that are likely to produce no significant analytical degradation), one moves to the next stage of analysis.

## Ranges, Means, and Distributions

Next, for each element in each table, review system documents to learn the prescribed definition and compare them to results produced by analyzing the actual data in that field. Look for unique occurrences of values in an alphabetically ordered list. Review related source documentation expectations for mean values and presumed distribution patterns of occurrences. Values are produced and a rank ordered list created, most frequent to least frequent.

Compare the results to the standards used in data definition documents to determine if the data set is producing meaningful answers and that the answers are within the values proscribed by the standard.

Generally, at least 80% of the results come from 20% of the data elements on the list. Should the results differ significantly from this general expectation, one would have to investigate the logic of that difference. At this time, people familiar with the business from areas like sales, marketing, and operations should be reviewing the lists with the analysts to determine the reasonableness of the observed results.

## Derived Relationships

In this analysis, investigate the logic of the results and the meaning of basic relationships by comparing outputs of the same or very similar data element when it occurs in different tables. Time is often the most commonly occurring data element around which to build relationships for testing. Time can be planned, scheduled, actual, and/or final. Frequently there are implied relationships between all four and often each is recorded through a variety of means.

Characteristics like format, population make up, linkage, and its unique identification capability are checked. This sounds so simple in concept and it is; but it is frequently not possible because of data system issues arising from their construction in a structure-dependent database. Current database languages are a common source of this problem. At this point, relationships asserted in the documentation and associated definitions are constructed and tested (e.g., scheduled time versus actual time and the associated sequence of events relationships). Computed relationships are constructed and tested for logic and likelihood.

## Meaning and Interpretation

Simple performance measures are now constructed to begin to get at the real meaning of the information in the data collection.

A good first test would use the simple measure produced by a Ziffian ranking distribution of all the values to a data element. The Ziffian distribution test illustrates behavioral independence or dependence as expressed in the particular measure. Arrange each occurrence of a data element value in rank order, high to low. Plot the log of the value versus the log of the rank. The results come in three forms. A straight line, two lines where the largest values appear constrained, and two lines where the lowest values appear constrained. The first line is the "God punishes everyone equally" picture. It means there is no constraint to the values you are reviewing; the data element's value is an independent factor. In the second, it frequently means a constraint, like market entry or monopoly factors, are present and constraining the value. The third suggests your data system is not capturing all the cases (under sampling) that are relevant to this element.

Logic can be applied to results to agree or disagree with the constructed information. That is to say, based upon your prior use of the data in its business applications, one should have an idea that the business system under observation is dependent or independent with regard to the measurement. These calculations confirm that notion or begin to tell you that your notional idea is false or that your data measurements are extremely poor.

On the assumption that some of your data represents different periods (days) of observation, one makes Paratto Distributions to study the change in the mix of outcomes. Such analysis can also assist in understanding business actions like market penetration of new products versus cannibalization.

Similarly, these different periods can be compared to discover changes in volume and velocity. When the source of these difference can be understood, they can often be translated into cycle time improvements for the various operations of the business.

# Hypothesis and Discovery

As your business team reviews the performance measures, options, and outcomes from the analytical data generated in the prior activity, some issues that have been under review become more clear though some remain foggy. At this point, turn attention to the latter to try and understand what assumptions of the organization are being challenged by the results of the analysis. By studying the unexpected things, one moves closer to discovering truth or knowledge about ones business that has remained hidden.

Intuition or strong beliefs weakly held is the starting point for this analysis. Hypothesis testing and a general failure of the analysis to confirm the initial hypothesis is an important clue to the discovery of new knowledge. Frequently one learns that these strong beliefs are based on opinion or limited experience extrapolated into a general understanding. Often they have been impeding progress in essential areas of the business. By working through the beliefs, assumptions, and basis with knowledge of the *quality* of the underlying data, the truth, or lack of truth emerges from the measurements that either supports the idea or not. Management can then respond accordingly.

# Business Process Actions

At each stage of the preceding analysis, there are serious questions for the management of the business.
From the Test of Completeness/Emptiness, recommendations on how to invest in better source data follow two lines of reasoning.

(1) If the source table can be considered a reference table, what alternative sources, timing, and costs are there to improve to the product currently in use?

(2) If the source table is transaction processing based, what would it cost to improve that data collection activity and would it be worth it?

Consideration can also be given to stop collecting and/or processing data that is and has been trash. New collection methods can be considered and/or created as new technology is deployed.

Distribution Analysis can help focus discussion on the need to know more about the details of specific transactions and the willingness and/or need to invest in installing devices to improve data collection accuracy in particular ways.

Discovery of redundant capability and weak data permits getting rid of source problems or minimizing the resources required to obtain it. The more one is able to deal with this problem, the easier it will be to create and operate a "record data once; read many" policy. Capturing the data once and reusing it for many purposes is at the heart of a quality data system. The resulting efficiency and reduction in rework make the policy economically sound too.

Finally, all of this work involves discussions with management about their performance measures for the organization. It frequently leads to adopting additional or complementary measures that they hope will add clarity to their views of the business.

Management now should have a clear understanding of how the various performance measures differentiate the business, the significance of changes in each of these measures, and the effectiveness of the organization's response to these changes. Management can have a high level of confidence in their internal information sources, capabilities, and performance measures.

Because of this new found capability, the idea that a company's information can be easily used in more supportive ways should emerge. It can result in a revised approach to obtaining focused information for new business issues. Following the hierarchy (methodology) described in this paper can result in a more value-based business relationship between the IT unit and management. It can prevent undercutting by anecdotal story telling, criticism, and lack of confidence that arises from every little data error. "Control" of the business as defined by Peter Drucker is achieved.

# Conclusion

In conclusion, these are the benefits of engaging in a continuous improvement business process for creating quality based decision support data. By using the data holdings of an organization and focusing on the data context and its relationships, the business meaning of the data is at the core of the analysis and discussions. It tends to minimize discussions about the efficiency or effectiveness of the IT and its various data collection processes. These values return to IT for internal management to address in support of the business discussions and the support of the business elements for improvement in the IT area.

The business relationships that are revealed through the analysis provide improved knowledge about the organization's behavior and form a basis for forecasting the business responses to future actions of the management. If management wants to track specific new performance measures, implementation will be easier since the requirements are likely to be more focused and understood.

This approach provides senior management with a powerful review of the knowledge base for their business capability and an understanding of the quality of their data. They can discover hidden relationships that were found by interrogating the data. They may desire to restructure some data for application access in order to continue to monitor a particular business situation. The methodology is designed to communicate business needs between the IT unit and management. It can lead to a better understanding of the cost and action alternatives needed to keep *knowledgeably* ahead of the competition.

The table that follows is provided to assist the reader in relating the terms introduced in this paper.

# Comparison of Terms

| Maslow | Analysis | Tests |
|---|---|---|
| physiological | adequacy | completeness, emptiness |
| safety | quality | ranges, means, distribution |
| belonging | options/connectivity | derived relationships |
| esteem | meaning/relationships | meaning, interpretation |
| self actualization | eureka | hypothesis, discovery |