MANAGERIAL ISSUES IN DATA QUALITY

Donald P. Ballou*

Giri Kumar Tayi

Management Science and Information Systems

School of Business

SUNY-Albany

Albany, NY 12222

ABSTRACT

Data has become the raw material for the Information Age. Accordingly, maintaining and enhancing the quality of the data resource needs to be an important organizational activity. However, doing this presents a series of managerial challenges. These arise in part from the nature of data but more from the increasingly prevalent trend of using data in ways that were not envisioned or intended. An individual or group charged with responsibility for the data's integrity must determine who is using what data for which purpose, assess the nature and extent of any deficiencies that may exist, and evaluate the impact that quality problems could have on the various uses of the data. This paper explores issues such as these, while highlighting concepts, techniques, and tools for ensuring that the organization's data is of high quality.

Key words: Data Quality, Management of Data Resource, Procedures for Data Quality Management

MANAGERIAL ISSUES IN DATA QUALITY

Mark Twain once said, "Everybody talks about the weather, but nobody does anything about it." One is tempted to make a similar statement regarding data quality. Certainly there is a steady flow of reports concerning problems with data. For example, the Wall Street Journal [18] reported that

> "Thanks to computers, huge databases brimming with information are at our fingertips, just waiting to be tapped. They can be mined to find sales prospects among existing customers; they can be analyzed to unearth costly corporate habits; they can be manipulated to divine future trends. Just one problem: Those huge databases may be full of junk. ... In a world where people are moving to total quality management, one of the critical areas is data."

Problems with data quality are not new and have been a continuing concern to those in the information systems profession. For example, John Gosden [10], writing in 1979 when he was the Vice President for Corporate Computer Services for the Equitable Life Assurance Society, stated that

> "A big mistake is the assumption that data is reliable: People do not design for bad data, but they should. I sometimes suspect that in some systems we would be better off to assume values for the data than to collect it. There are many more ways for data to be wrong than right".

Increasingly organizations are focusing attention on the problem of data quality. In part this is because the need to address data quality deficiencies has become more critical. As we move into the information age, more employees than ever work directly with computer-based data. This generates a new set of problems. For example,

> "Bad data is like a virus: a given data item gathered for one business purpose and intended to be stored in a single database may actually be used for dozens of business purposes and replicated in numerous databases" [11].

A related problem with multiple users of data is that of semantics. The data gatherer and initial user may be fully aware of the nuances regarding the meaning of the various data items, but that will not be

true for all of the other users. Thus, although the value may be correct, it can easily be misinterpreted. The separation of the data gatherer and data user has other negative consequences. Back when data were almost always tightly coupled with application programs, those responsible for the system had a good feel for what to expect, and they were able to pinpoint some of the more egregious problems. That capability of judging whether the data are "reasonable" is lost when users do not have any responsibility for the data's integrity and when they are removed from the gatherers. Such problems are becoming increasingly critical as organizations implement data warehouses.

At this time, especially in industry, there is considerable activity regarding data quality, much of it ad hoc. Problems with data quality arise, become serious enough to require attention, and are then addressed [9]. Some organizations have on-going efforts to improve the quality of data. An example is AT&T's program which concentrates on developing, instituting, and implementing procedures that are designed to ensure that data entered into computer-based system possesses the requisite level of quality [11]. Another example is the Environmental Protection Agency which has developed specific data quality objectives for its monitoring programs in making effective regulatory decisions. In particular it has identified acceptable tolerance levels for different types of data errors [7]. The US Fidelity and Guaranty, a multiline insurance company, provides a third example. It has in place a vigorous, company-wide data quality control program. The focus of the program is to make available to management accurate and timely data so that decisions can reflect the true mix of risks they insure and the associated costs and benefits. The program comprises four elements: prevention, control, assurance, and audit [15].

Some companies are instituting the position of data quality manager, an individual with responsibility for ensuring the integrity of the organization's data. (In practice data quality management may well be a group activity, but for expository purposes we shall assume that a single person fills that role). But what such an individual should do, and how he or she should do it is an issue of great importance. This person must have a sound understanding of the nature of the organization's data, must identify and enforce policies that help ensure the quality of data, and must effectively balance conflicting requirements for data quality and make appropriate tradeoffs.

Such an individual would also be responsible for maintaining the integrity of data found in a data warehouse. Addressing such managerial issues is the purpose of this paper.

THE NATURE OF DATA

The U.S. Department of Labor has estimated that well over 40% of those in the workforce will be engaged in data services by the year 2000. Clearly data is the essential ingredient, and hence those who work with data, especially the data quality manager (DQM), cannot be effective without a sound understanding of the nature of data. To acquire this, one is tempted to start with Webster's definition of data: "factual information (as measurement or statistics) used as a basis for reasoning, discussion, or calculation." Although appropriate in general, this is not especially helpful to the DQM. Neither the Encyclopedia Americana nor the Encyclopedia Britannica has a category "Data", or for that matter does the Encyclopedia of Computer Science and Engineering. It appears that something almost everyone uses on a daily basis is not discussed in a systematic fashion in the general literature.

Not surprisingly perhaps, publications aimed at the information systems community are more helpful. For example, Alter [1] defines data to be "facts, images, or sounds that may or may not be pertinent or useful for a particular task." As is usually done, he distinguishes data from information, the latter being "data whose form and content are appropriate for a particular use." He identifies five types of data- formatted data, text, images, audio, and video-, a taxonomy considerably broader than that associated with traditional data processing systems.

In a very real sense data constitute the raw material for the information age, and yet any analogy is tenuous at best. Unlike physical raw material, data does not get consumed and in fact can be reused repeatedly for various purposes. Whereas physical raw material has a tangible existence, stored data consist of the equivalent of magnetized spots and can vanish instantly leaving no trace. By themselves the things we call data may well be meaningless unless placed in some context. An obvious example would be physical measurements for which one needs to know the units before the stored values have any relevance.

With physical raw material one has a pretty good idea what quantity is required. With data that is not so, as some will complain of the need for more data whereas others will object to an information overload. The value of the raw material to an organization, from an accounting perspective at least, is clear. In contrast the value of data depends almost entirely on its uses, which may not even be known . Furthermore, data cannot have value unless they are accessible, sometimes a problem for those in the next room but not for others thousands of miles away.

Dimensions of Data Quality

It is not sufficient to state that the data are wrong or not useful, for that provides no guidance as to how one should go about improving the data. To report that the data are inconsistent indicates a problem rather different from saying that the data are out of date. Part of the DQM's job is to elicit from data users as precisely as possible exactly what it is with the data that they find to be unsatisfactory. This is facilitated by an awareness of the dimensions of data.

It has long been recognized that data are best described or characterized via multiple attributes or dimensions. For example, in a paper that focused on how information systems can amplify or diminish deficiencies in data, Ballou and Pazer [2] identified and discussed four dimensions of data quality: accuracy, completeness, consistency, and timeliness. An example of the role of these dimensions can be found in Laudon's study of data problems that exist in the criminal justice system [12]. Accuracy could refer to recording correctly facts regarding the disposition of a criminal case, completeness to having all relevant information recorded, consistency to a uniform format for recording the relevant information, and timeliness to recording the information shortly after the disposition. Imagine the impact on individuals of poor quality on any of these dimensions!

Recently Wang, Strong and Guarascio [20]analyzed the various attributes of data quality from the perspective of those who use the data. Their analysis began by soliciting information from users regarding various descriptors attributable to data. This resulted in over 100 items. Factor analysis was then applied to these attributes to group them into about 20 categories each consisting of attributes that users reacted to in a similar fashion. For example, the attributes "unbiased" and "objective" were grouped to form one of the categories. These were further analyzed to identify four broad groupings;

Intrinsic data quality, Contextual data quality, Representational data quality, and Accessibility data quality. In terms of the dimensions identified by Ballou and Pazer [2], accuracy belongs to intrinsic DQ, completeness and timeliness to contextual DQ, and consistency to representational DQ.

WHAT IS DATA QUALITY MANAGEMENT?

The data quality manager is responsible for the quality of the raw material of the information age. Such an individual must traverse largely unexplored territory. The ultimate goals are uncertain and in any case probably cannot be fully achieved. Conflicting interests tend to push the data quality manager in differing, sometimes opposite, directions. There will not be adequate resources for the task. Yet the consequences of not ensuring the quality of the data resources can be catastrophic. Our discussion of what the data quality manager does begins with a definition.

> Data Quality Management involves specification of policies, identification of techniques, and use of procedures designed to ensure that the organizational data resource possesses a level of quality commensurate with the various current and potential uses of the data.

Although there have always been individuals charged with management of data quality for a particular system, only recently has the need arisen for a broader, organization-wide perspective. The responsibilities and activities of the DQM are varied but ultimately involve determining who uses the data for what purpose, identifying the data quality needs of these users, and then doing something about any data quality deficiencies that may exist. In addition the DQM should be sensitive to the need to upgrade the data on a continuous basis to enhance its value.

Various facets of data quality management are discussed by Wang [19], who places the material in the context of Deming's quality management principles. These include articulating a vision of data quality, establishing responsibility for data quality, educating personnel regarding the importance of data quality, teaching appropriate data quality skills, and institutionalizing systematic data quality enhancement procedures.

Clearly many aspects of the above approach to data quality management are dealt with at the macro level and impact organizational policies. Although the DQM would be involved with this, he or she, however, has the primary responsibility for the design and implementation of tasks and activities which support organizational policies. These tasks range from the tactical to the strategic and have been at the center of our work over the past several years. Insights stemming from these efforts are incorporated into this paper.

Data Quality Management Tasks

In the past who used what data for which purpose was obvious. Files were closely linked to systems, and it was well known what the systems did. In on-line database environments all data are potentially available to all employees. In practice only some small subset of the data is available to everyone, and use of other data sets is limited to certain individuals or groups. By data set we mean any identifiable collection of data in a traditional file or some subset of it, a collection of related files, one or more relational base tables, a relational view, related images, and so forth. What the users do with data made available to them may well be an open question. Some may have specific, well defined needs that occur on a regular basis. (Report generation is an example). Others may use data for monitoring purposes. (The VP for sales may wish to browse through sales data.) Another use could be for decision support. (HR personnel might wish to analyze the impact of a proposed change to the retirement system). The data quality needs of these examples are rather different. In the first case consistency and completeness are important, in the second timeliness may be critical, while in the third effective format may be the key. However, the differing needs cannot be addressed unless the data quality manager knows what the uses are. All responsibilities of the DQM follow from this knowledge.

Information Gathering

There are various approaches for obtaining information as to who uses what data for which purpose, but whatever is utilized must involve obtaining information from the organization's employees and customers. One of the authors and Harold Pazer, a colleague, have found that a multistage process for acquiring the desired information is effective. The process helps in understanding how data get created, used, and managed in an organization. The first step is to

distribute a Data Inventory form to all employees within a unit. This questionnaire elicits information on existing data sets, uses of the data sets, those responsible for gathering and entering data, perceived importance of the uses of the data, problems with the data, and so forth. This information is solicited for each data set with which an employee might have contact. The second part of the form solicits information on future data and system requirements. All employees are required to return the Data Inventory forms even if they should have no involvement with any data set.

Based on the returned Data Inventory forms, the DQM in conjunction with various unit managers identifies individuals from whom it would be beneficial to gather additional information. These people are prioritized (High, Medium, Low) as being appropriate for follow-up interviews. We have found it beneficial for the interviews to begin with an unstructured phase during which the employee is encouraged to discuss one or more systems/data sets, perceived problems, uses of the data, and in general any concerns related to the data. The second phase of the interview is structured and seeks additional and more detailed information regarding the data sets. Special emphasis is placed on the quality of data sets (especially vis-a-vis its intended uses), and questions are included which relate to current or anticipated sharing of the data. This information provides the basis upon which the DQM can proceed.

During the information gathering phase, one needs to ascertain precisely the data quality problems. Although users may complain about the quality of the data in general terms ("The data they expect me to use sure is lousy."), such feelings are usually based on rather specific, observed deficiencies. The data may be out of date, data from one unit may be incompatible with data from another unit, a few values may be orders-of magnitude wrong (misplaced decimal points), a large number of values may be a little wrong (inventory estimates are used in place of counts), and so forth. A key aspect of the information gathering phase is to elicit input from data consumers as to exactly what they perceive to be wrong with the data . Unfortunately, obtaining information regarding data deficiencies is not necessarily straight forward, as the problem may be quite different from what the user suspects. An obvious example relates to confusion regarding the meaning of a data item. One department may complain that the sales data are inaccurate when in fact the problem is that different

definitions are being used as to what constitutes a sale. The more data are shared, the greater is this kind of problem.

Analysis of the Impact of Data Deficiencies

One reason it is important to identify various uses of the data is so the DQM can gauge the seriousness of the deficiencies in the context of user needs. For example, most people would agree that using different alpha-numeric codes in different divisions of the organization to represent the same item is not desirable. Yet this need not create any difficulties. If these data items remain confined to the division of origin and are never combined across divisions, then there is no need to change anything. If they are combined but on a regular basis for the purpose of, say, a standard, monthly report, it is probably simpler to map the various codes to a standard set within the report generating program than to change all of them at the source. However, if the data items are made available across divisions on an ad hoc basis as would be the case with a data warehouse, then the DQM needs to do something to resolve the inconsistencies.

The need to assess the potential impact of data deficiencies is by no means limited to inconsistent coding schemes. As another example suppose certain numbers (say, preliminary monthly sales) can be in error by 30%. If many such numbers are averaged, if there is no systematic bias to the errors, and if this is the only use of the numbers (preliminary average monthly sales), then there probably is no need to take drastic, corrective actions. If on the other hand, one such number is used in the denominator of a ratio (deseasonalizing the sales), there could be very serious consequences, depending on what the ratio is and especially how it is used. When a data value is used for one and only one purpose, then systems designers can ensure that particular value possesses the requisite data quality. However, in the case of multiple uses, a level of data quality that is fine for one user may be inadequate for another.

Unfortunately, there is no simple way to determine with certainty the seriousness of the impact of data deficiencies. A good starting point, however, is to consider the perception of those who use the data in some way on a regular basis. If they do not have reservations regarding the data, and if those who receive reports, etc. based on the data are satisfied, then there is a basis for believing that any

deficiencies with the data, for this use at least, are not serious. At the other extreme, if people refuse to use data because they do not trust them, then something is seriously amiss. It is the in-between situations that most tax the judgment and insight of the data quality manager.

Ascertain the Nature of Data Deficiencies

To paraphrase the earlier quote by Gosden [10], there are many ways for data to be wrong. This is especially true in multi-user environments, for users may well have differing data quality requirements. It is necessary for the DQM to have an awareness of what could lead to inadequate data quality. A first step in understanding how data can go bad is to recognize the fact that data have multiple attributes or dimensions, as discussed earlier. A set of data may be completely satisfactory on some of these dimensions but inadequate on others. For example, one dimension is accuracy, another timeliness. It can easily happen that for a certain use the stored data values are accurate but not sufficiently timely. Thus there is a data quality problem on one dimension but not on another. For this use one would conclude that the data's quality was not satisfactory, a result of the timeliness problem.

Each person who works extensively with a set of data can identify or list the ways that data can be inadequate. Lists corresponding to different sets of data may contain common elements, but they will not be identical. One cannot say a priori "Here are the four ways this data can become deficient." Only through a lengthy and rather messy process can a DQM become sufficiently acquainted with the data and its uses to be in a position to identify potential problems with the data. These problems will be organization-dependent. For example, one would anticipate that the kinds of data problems encountered by a commodities exchange would be rather different from those found in a university.

Determination of Appropriate Level of Data Quality

One might wish for a state in which all of the organization's data is perfect in every way. Achieving that, however, could well bankrupt the organization. Nor is having such pure quality necessary. Does anyone care if the figure for the amount for paper consumed by an organization is accurate to the page? The trick is to know whether the quality is sufficient for the use. If the quality level is higher than required, then resources are being spent unnecessarily to maintaining quality that

is not needed. If the quality is lower than required, then at a minimum sooner or later the application will suffer credibility problems.

Determining an appropriate level of data quality is clearly context dependent. Deciding on an appropriate level is especially difficult when differing users have differing needs. One might be tempted to state that the use requiring the highest quality should determine the overall level of quality. But what if that use is rather minor and unimportant to the organization, whereas the major use of the data set does not require anywhere near such a level of quality? Thus the DQM must not only determine the level of quality required by each use of the data set, but in the process needs to balance conflicting requirements for data quality.

Choosing Procedures to Ensure Data Quality

There are two basic approaches to ensuring data quality. The first is to keep problems with the data from developing. Various principles can be employed to achieve this, many of which have an analog in the manufacturing world. These include capture of data in machine readable form; use of as few data entities and values as possible; limiting storage locations for data. The second approach is to fix problems as they arise or are identified. The former is preferable, as it tends to eliminate credibility problems and in addition usually is less expensive. However, that choice is not always available. For example, suppose the merger of two divisions results in the consolidation of their data centers. Further suppose that certain information in one case was reported and stored on a bi-weekly basis and in the other on a monthly basis. Such data cannot be combined directly for reporting, decision support or any other purpose. New data files based upon the existing incompatible files would have to be generated to take the differing reporting periods into account.

Ensuring that the data have and maintain the requisite level of data quality is best accomplished via the judicious use of procedures and incentives. This is the approach adopted by AT&T Bell Labs for their data quality efforts [17]. Those who generate, enter and maintain data must be properly trained and supervised and need to have a personal stake in doing the job well.

An anecdote emphasizes the need for personal involvement and supervision of those who generate, enter, and maintain data. Several years ago a populous state required each hospital within

that state to file a report on its activities during the previous year. Since from the hospital's perspective there was little benefit or reward for doing this mandated task, it was usually assigned to a lower-level employee who was told in effect to "Fill out the forms." Not surprisingly these individuals tended to be sloppy, especially with respect to the hard-to-fill fields. For example, it was not uncommon to use last year's report as a starting point and from there make a fictitious adjustment. A few years of this would completely destroy the data's integrity!

Procedures to clean up dirty data are varied and depend on the nature of the problem, user needs for better quality, and the cost involved. Inspection, rework, and imputation are the basis for the procedures. At one extreme each field of each record could be inspected, verified, and reworked using the data's source. For example, with some personnel systems copies of the data fields are forwarded periodically to those most likely to detect errors, the employee for some of the fields, the supervisor for others. Doing this is time-consuming and costly but is very effective. In case of missing data values statistically-based imputation procedures can be used to obtain reasonable values [14]. At the other extreme procedures can be used to automatically identify outliers which are then individually checked out. Clearly this approach does not catch all errors but would detect many of the most serious ones.

Perhaps the best way in general to detect errors is to have those most familiar with the data scan the fields on a regular basis looking for suspicious values. (This is inappropriate for large data sets.) This in conjunction with data audits that examine periodically both data quality maintenance procedures and through sampling their effectiveness helps pinpoint areas needing attention.

Selection of Data Quality Enhancement Projects

Although the DQM has at his or her disposal various procedures and tools to maintain and enhance data quality, in practice selection of what data sets are to be maintained at which level is constrained by resources and political considerations. Often there is the need for the use of some sort of cost-benefit analysis. As is usually the case, costs are easier to identify and quantify than benefits. That said, the measurement of benefits can also be straight-forward in some cases. An example would be changing procedures to ensure accurate inventory values. At the other extreme would be projects that enhance the organizational data infrastructure. An example of this would be reformatting or

reworking data from different sources so that combined data can be shared across the organization for ad hoc purposes. The DQM needs to prioritize the various data quality projects while being sensitive to these considerations. Later in this paper we explore the problem of determining which project to undertake in an environment where benefits may be ambiguous or even unknown.

WHAT MAKES ENSURING DATA QUALITY DIFFICULT?

One can argue that ensuring the quality of data is much more difficult than is the case with manufactured goods. Inspection has traditionally been used in the latter case at various points of the production process to ensure adherence to predefined standards. Although inspection is costly and not perfect, its use coupled with attention to quality production processes can yield high quality products. Production processes are self-contained. Although the output of one manufacturing process may be input for another, the requirements of the second process are well known and so the intermediate product can be designed with those in mind.

In a world of multiple, ad hoc data users, much of the certainty of the production environment vanishes. The raw material, the data, may well be of uncertain quality and its uses may be known only in part. The effectiveness of possible inspection procedures is uncertain if data undergo a series of ad hoc processing steps. It is possible technically to combine data that were never meant to be combined. In addition to these difficulties there are several factors that complicate the job of a data quality manager.

Uncertainty regarding what constitutes the data resource. For what collection of data should the DQM be responsible? The answer is neither obvious nor simple. One might be tempted to reply "all data stored in computers". What about engineering drawings? And word processing documents? To narrow the scope, one might suggest certain key systems. But what if these systems use data from unreliable, external sources? Should the DQM be charged with working with those sources to improve the quality? Where do the responsibilities of other IS personnel end and those of the DQM begin? To what degree should the DQM be involved with specifying or reviewing traditional edit

checks? It is easy to ask such questions, much harder to answer them. But if the DQM is to be successful, some clear delineation of his or her responsibilities is essential.

Perceived lack of importance. In some ways data quality and computer security are analogous. Almost everyone agrees that ensuring computer security is an important activity, but at budget time it tends to get shortchanged. It has been said that nothing increases the budget for computer security like a well-publicized breach or disaster. Similarly, ensuring data quality is widely recognized as a valid and important activity, but in practice only a few people list it as a top priority. One of the major responsibilities of the CIO and the DQM should be to sensitize executives and managers to the importance of ensuring data quality.

Multiplicity of potential problems with data. There are so many ways for data to be wrong. One assumes that the foreign exchange rates in today's paper are correct, and yet in actuality the values were out of date before the paper was even printed. For a particular file every value could be both correct and timely, but certain records (rows) may be missing entirely. The sales data may be accurate, timely and so forth, and yet be of little use if an inappropriate reporting period is used. The figures given to the Securities and Exchange Commission may be correct but irrelevant as far as investors are concerned. Estimates may be provided for the inherently unknowable (What percentage of employees are doing their best?) with no way of judging the validity of these estimates. No one can anticipate everything that could compromise the integrity of the organization's data. Awareness of this does not provide a solution but is a necessary ingredient for effective data quality management.

Inadequate documentation of data. It is easy to reach into a database and pull out data, a consequence of the marriage of data base management systems and data communications. The worth of that data is another matter and can be known only if the user has access to documentation (preferably on-line) that clarifies the meaning, source, timeliness and other attributes of the data. The need for some sort of documentation is highlighted by the following:

> "Madnick [Professor of Information Technology at MIT] notes that at one U.S. insurance company ..., net premium is the standard business measure -- and yet the company has 17 different definitions for net premium, each equally valid" [16].

Uncertainty regarding the seriousness of deficiencies. As indicated, this problem is primarily a consequence of the ability to use data for multiple purposes. If all uses are well-known, it is certainly easier to determine risks. However, even if this is the case, data errors can have surprising consequences. It has been shown that random errors can have a profound impact on forecasts to the point that the forecasting methodology itself is inappropriate [3]. In that study a software package was used that not only found the best fit using linear regression but also tested several other types of regression as well (quadratic, log linear, and so forth). Not only did the forecasts change as errors were corrected (that was to be expected), but more seriously the forecasting methodology changed. It is difficult to underestimate the damage that can be caused by data of poor quality. A naively simple error in one case could cause catastrophic damage in another context.

TECHNIQUES FOR FACILITATING DATA QUALITY MANAGEMENT

Effective managers combine knowledge of their sphere of responsibility with judgment. Activities designed to ensure and enhance the quality of the data resource will inevitably require that judgment play a prominent role. Nevertheless, the application of judgment is most effective when applied within an established framework. Furthermore, it should be supported by the use of various procedures, techniques, and models. An early example of such a framework is the decision calculus enunciated and applied by Little [13]. The models supporting the framework can fall anywhere along the qualitative - quantitative continuum. Those at the qualitative end rely more heavily on soft data, are not likely to yield an optimal solution, but are usually flexible. Those at the quantitative end have specific data requirements, produce optimal or near-optimal solutions, but do not handle soft factors well. In this section we describe several of the procedures that can be used by the data quality manager to facilitate his or her work.

The first procedure provides a framework for project selection and is qualitative in nature. The second identifies which of the existing data sets should receive priority for data quality enhancement and is a quantitative technique that optimizes allocation of resources. The third focuses on the processes that produce predefined information products. It is a model that can be used, among other

things, to analyze the impact of changes in the quality of the data on the quality of the information product.

Data Quality Project Selection. Ultimately the data quality manager must identify what activities or projects to undertake to maintain or enhance the quality of the data resource. As usual, some sort of cost-benefit analysis is appropriate. Quantifying benefits is in general more difficult than with costs but is especially so for data quality projects, a consequence of the fact that the same data set can be used by several applications for rather different purposes.

Selection of projects follows the usual pattern. The information gathering phase is designed, among other things, to identify opportunities and problems in the sphere of data quality. The DQM must identify at least one solution for each problem or opportunity and then estimate the costs and benefits associated with that solution. Cost-benefit ratios can then be used to select projects.

The difficult part is the determination of the benefit of a possible project. The benefits resulting from improved data can be highly subjective, and their evaluation may well differ from one individual or group to the next. Since a particular data quality project will affect multiple uses of a data set, it is necessary to evaluate the benefit for each use. It can happen that a project that is beneficial for one application will be deleterious for another. This occurs when a tradeoff (such as the accuracy-timeliness tradeoff) exists, and the project improves one dimension at the expense of another. A net benefit for a particular project is obtained by aggregating the benefits across the various uses that are impacted. It should be kept in mind that for certain uses the "benefit" could be negative. A procedure for evaluating benefits systematically in such an environment is given in Ballou and Tayi [5]. That approach utilizes a multi-stage process designed to reduce ambiguity regarding project benefits. At the beginning and end the involvement of senior IS personnel is critical.

Enhancing the Quality of Stored Data Sets. As mentioned, it is preferable to keep problems with data from developing, but one does not always have that option. Unfortunately the DQM does not have full knowledge as to the seriousness or extent of any identified data quality deficiency. He or she could examine a data set carefully, determine precisely all deficiencies, and study thoroughly the impact that these deficiencies would have on uses of the data set. However, doing that is expensive, and analyzing

a few data sets in this way could easily exhaust the budget for data quality assurance. At the other extreme the DQM could commit funds to enhance the data quality for those sets that appear to be most deficient. After the fact it could well happen that in actuality these data sets were not in as bad shape as thought initially.

Clearly some balance needs to be struck between determining which data sets are in the greatest need of attention and actually doing the work of enhancing their quality. Striking this balance is complicated by considerable uncertainty regarding the nature of deficiencies in the data sets, their frequency or extent, and their impact. Ballou and Tayi [4] developed a quantitative procedure that uses an iterative approach in conjunction with an integer programming formulation designed to address the uncertainties and conflicting needs. Among other things their approach provides guidance to the DQM regarding how limited resources should be committed to determining quality levels of different data sets vis-a-vis improvement of the data sets.

Reengineering to Enhance Quality of Information Products. In many cases the output of an information system can be thought of as an information product, something whose value resides primarily in its information content. This value is affected by various data quality attributes. Who, for example, wishes to learn about last week's baseball results in today's newspaper? Clearly the quality of such an information product is dependent upon the quality of the input data coupled with the effects that the various data processing and quality assurance activities have on that data. The functional relationship, however, between the quality of the input data, the effectiveness of the processing and quality assurance activities, and the quality of the information product is less clear. Furthermore, for complex systems it may be difficult to determine what set of activities is most responsible for any delay that may arise in producing the information product. These problems, of course, are also found in manufacturing systems that produce physical products. Recent efforts by Ballou, Wang, Pazer and Tayi [6] have produced a model of manufacturing information systems that specifies and tracks various parameters which ultimately provide measures of the quality, timeliness, value, and cost of the information product. The model facilitates analysing possible improvements to the quality of data

inputs and the resulting change in quality to the information product. It also permits analysis of various alternative information manufacturing system configurations.

The techniques discussed above span the entire qualitative-quantitative continuum and address different facets of data quality management. They range from selection of individual projects to possible reengineering of entire, information systems, either existing or potential. There are, of course, many other tools the DQM can utilize to achieve the organization's quality goals.

CONCLUSIONS

It is becoming increasingly evident that inadequate data quality is a major problem for organizations. Mark Hansen [16], a principal at QBD Solutions, a firm based in Cambridge, Massachusetts,

> "surveyed about 50 Fortune 500 companies in 1990 (while he was at MIT) [and found that] 'two thirds of them reported significant problems with the quality of their data'".

World class organizations are continuously seeking ways to harness the full potential of their data so as to achieve their competitive and strategic goals. However, the degree to which they can accomplish those goals depends critically on the quality of the data, and the procedures and policies used in furthering that quality. The use of data in organizations has changed over time and so has the importance of data quality. Today, more employees of an organization deal with data than ever before. The data is created by one group (order entry) in one part of the world and used for decision making by another group (marketing) thousands of miles away. Just a decade or so ago all data entry, processing, and use were carried out in centralized environments and met specific needs. In such an environment detecting sources of error and ensuring data quality was a relatively straightforward and manageable task. As more and more products compete in the global market place and with the advent of global telecommunication networks, organizations are being forced to manage a large amount of data which may be in different forms - voice, data, image, audio, and video - and could be generated by many individuals and institutions dispersed all around the world. In the past as long as the data production and consumption were separated, inspecting quality into the data using traditional quality controls and error checking processes was a reasonable approach for the organization. However, as

data increasingly becomes one of the most important strategic resources for an organization, designing quality into the data should become the primary focus of all efforts in ensuring data quality.

It is becoming clear that organizations of the future will predominantly be involved in offering knowledge-based products and services [8]. As an example, currently most credit card companies provide the card holder with a monthly report of all the individual transactions. Instead, the company could offer a value-added service by summarizing the transactions into card holder specified categories thereby transforming the raw data into an information-based tool that could be used by the customer for, say, tax preparation, travel expense accounting etc. Going a step further, the company could extract the unique characteristics or information underlying each customer's pattern of transactions and then utilize it to provide knowledge-based services which could meet specialized needs of an individual and/or specific groups of card holders. Automatically forewarning a card holder in the event of any transaction deviating from his or her unique pattern could be one such service.

The single most essential ingredient for providing value-added information services is the quality of the data. Accordingly, the organizational processes and policies which are deployed to ensure data quality should be focused, as required or needed, on raising the current data quality levels from their best level of 1 - 5 percent error rate to 1 - 5 errors per million (epm), that is attain a three orders-of-magnitude improvement. In fact, those organizations that are effective in extracting relevant information from data of high quality and converting that information into useful and productive knowledge will be the ones which can succeed in the fiercely competitive global market place.

References

1. Alter, S. Information Systems: A Management Perspective. Addison Wesley, Reading, MA, 1991, p. 81.

2. Ballou, D.P., and Pazer, H.L. Modeling data and process quality in multi-input, multi-output information systems. Management Science, Vol. 31 No. 2, (1985): 150-162.

3. Ballou, D.P., Pazer, H.L., Belardo, S., and Klein, B. Implications of data quality for spreadsheet analysis. Database, Vol. 15 No. 6, (1987): 509-521.

4. Ballou, D.P. and Tayi, G.K. Methodology for allocating resources for data quality enhancement. Communications of ACM, Vol. 32, No. 3 (1989): 320-329.

5. Ballou, D.P., and Tayi, G.K. Determining priorities for projects with uncertain benefits: an application to data management. European Journal of Operational Research, Vol. 76 (1994): 206-217.

6. Ballou, D.P., Wang, R.Y., Pazer, H.L., Tayi, G.K. Modeling information manufacturing systems to determine information product quality. Cambridge, MA, MIT Sloan School of Management Working Paper TDQM-94-05, 1994.

7. Blacker, S.M. Data quality and the environment. Quality, (1990): 38-42.

8. Davis, S., and Botkin, J. The coming of knowledge-based business. Harvard Business Review, Vol. 72, No. 5 (1994): 165-170.

9. Goodhue, D.L., Quillard, J.A., Rockart, J.F. Managing the data resource: A contingency perspective. MIS Quarterly, Vol. 12 No. 3, (1988): 373-392.

10. Gosden, J.A. Some cautions in large-scale system design and implementation. Information and Management, Vol. 2 No. 1, (1979): 7-14.

11. Huh, Y.U., Keller, F.R., Redman, T.C., and Watkins, A.R. Data quality. Information and Software Technology, Vol. 32, No. 8 (1990): 559-565.

12. Laudon, K.C. Data quality and due process in large interorganizational record systems. Communications of ACM, Vol. 29 No. 1, (1986): 4-11.

13. Little, J.D.C. Models and managers: the concept of a decision calculus. Management Science, Vol. 16 No. 8, (1970): B466-B485.

14. Little, R.L. Editing and imputation of multivariate data: issues and new approaches. In G.E. Liepins, and V.R.R. Uppuluri (Eds) 1990.

15. LePage, N.J. Data quality control of United States Fidelity and Guarnaty Company, in G.E. Liepins and V.R.R. Uppuluri, (Eds) Data Quality Control: Theory and Pragmatics. New York, Marcel Dekker, 1990.

16. Kiely, T. Keeping your data honest. CIO October 15, 1992, pp 42-46.

17. Redman, T.C. Data Quality: Management and Technology, New York, Bantam Books, 1992.

18. Wall Street Journal Data Bases are Plagued by a Reign of Error. May 26, 1992.

19, Wang, R.Y. Toward total data quality management. In R.Y. Wang (Ed.) Information Technology in Action: Trends and Perspectives Englewood Cliffs, NJ, Prentice Hall, 1993, pp. 190-196.

20. Wang, R.Y., Strong, D.M., and Guarascio, L.M. Data consumers' perspective of data quality. Cambridge, MA: MIT Sloan School of Management Working paper, 1994.