Strategies For Gaining Competitive Advantage In A Dynamic Environment Thru Data Quality

Andrew B. Nyaboga, William Paterson University of New Jersey, USA
Muroki F. Mwaura, William Paterson University of New Jersey, USA

ABSTRACT

Raw data should be treated as materials for manufacturing products so that organizations can identify errors which otherwise could not be detected. Information system can be studied from the same perspective of product manufacturing.

Keywords: Data, raw materials, entity, attribute, value, dimensions of data quality, granularity, composition, semantic consistency, structural consistency

INTRODUCTION

Dramatic developments in technology over the last four decades or so have improved enterprise's capabilities to store and distribute data through the enterprise databases. Although these data are ubiquitous and crucial to everyday enterprise operations, they are often of very low quality. In most organizations, there is insufficient attention to this very essential problem of Adata quality. Despite the use of a system development life cycle, systems often produce poor results that alienate customers and clients resulting in huge maintenance expenses, and reduced business efficiency, effectiveness and profitability. Those organizations that need to be ahead of the game must adhere to the dimensions of data quality. Understanding the problem of data quality, organizations must address the data and data management relationship. Most data management is performed with databases. A database is considered to be a large related set of data that may be stored and presented on physical media.

What are data? Data may be classified as having the following characteristics:

- collected purposely
- stored in a medium
- repetitive in nature
- encoded in a specific format

Data are collections of attributes and relations, alternatively defined as:

- a set of facts
- raw material of information
- result of the conceptualization of the real world and relationships

The real world consists of entities which are defined by a set of attributes, and the set of attributes is defined by a domain having type and range values. Therefore, at the lowest level, "Adatum", which is a piece of data, is defined by a triple (entity, attribute, value). This is the same as saying:

DATUM = (entity, attribute, value)

Figure 1 below gives a summary of the definitions given above.

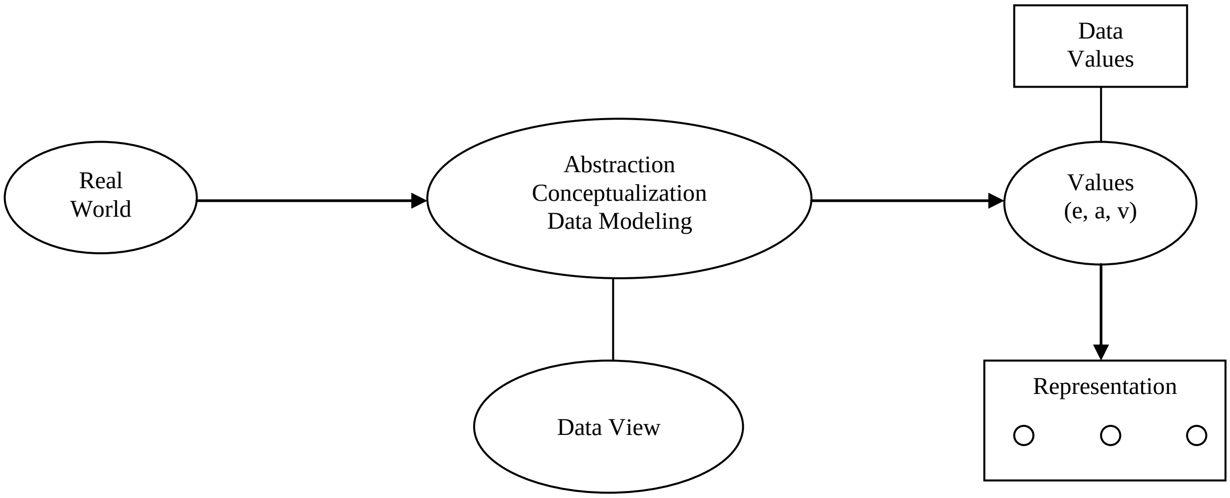

Figure 1

Looking at data an abstraction of the real world, we get three major dimensions of data quality. Each of these data dimensions has certain quality dimensions associated with it.

Data Quality Major Categories

The three categories of data quality referred to in Figure 1 are: data quality of a view, quality dimension of data values and quality dimension of data representation.

I. Data Quality of a View

A database model is a collection of a universe of typed data, typed operations and typed relations that constitute an “Aapplication world” that, in principle, may be formally or logically described by a database view.

A view is a structural description of a data model, i.e. it gives the structure or the structural format of the database (data) structures as specified by a schema or scheme. Additionally, it gives a set of logical sentences known as integrity constraints that are validated by the data model. The term validated means that all of these sentences may be interpreted as true statements using the contents of any database state. The database state may be thought of as contents of the database structure at any fixed time before or after all the data manipulation or transaction is performed. Thus, a relational model is described by an entity relationship view, an object oriented model by an object oriented view, etc.

A quality view is generally considered to satisfy certain dimensions of quality among which we identify:

- Content
- Scope
- Level of detail
- Composition
- Consistency
- Reaction to change

A. Content: refers to actual facts or “ideas” represented by data. Content may be further characterized by the following sub-dimensions:

- relevance
- clarity of definition
- Obtainability of values

Relevance: Are data described by data values relevant or needed for a view or subview?. Some pertinent questions are: What are the authoritative sources of data? Will the data be useful and useable to current or intended applications? How do you collect the data and how do you measure errors?

Clarity of definition: The components of a view, entities, attributes and value, must be defined to have meaning. What is a graduate student? An auditor? A full time student? A full time equivalent? or An engineer? Definitions are needed by either the data collectors or the potential data users.

Obtainability of values: Data are not tangible. It is complicated to get correct data as there are many factors involved, i.e. the legal factors, expenses and the easiness of collecting or obtaining correct values. One can obtain a person's date of birth, but the information cannot be used for hiring purposes because there will be legal ramifications.

B. Scope: Includes comprehensiveness (broad view to satisfy all applications) and essentialness. View does not include unnecessary information.

Comprehensiveness: this refers to the need to identify needs of multiple users and multiple applications and technical requirements. View may also have to take into account the future as well as present needs.

C. Essentialness: What is essential? What is not? When data are collected, they are assumed to be important, but not all the data generated have value. Consider the following three problems caused by unnecessary data:

- Unused data - it is hard to check quality without usage. Take an example of the John Deere tractor. How do you know how long it lasts? Because it has been used before.
- Redundant information may cause anomalies.
- Unjustifiable acquisition: What about the cost? Analysis of cost benefit? What is important is that non-essential data may interfere with essential data, hence, causing damage.

D. Level of Detail: Level of detail refers to both quality and quantity of data that must be included. It consists of the subdimensions of granularity of semantic entities and domain precision. Granularity:More granular is an indication of more details and richer information for future or unanticipated needs, but it is also more expensive in space and time and thereis a greater possibility of errors in data acquisition. Precision of domain: This is not the same as accuracy because precision is defined by domain structure, e.g., the numberof domain values(such as color or weight) for a given attribute. The greater the number of values in the domain, the greater the precision. For example, James' height is 72 inches. This measurement in terms of inches provides a greater precision than if the measurement had been given in feet and inches.

E. Composition: Composition is concerned with the way semantic entity types such as attributes and domains are structured or grouped .Major groups include the following: Naturalness - semantic types should correspond to a natural semantic unit. Occurrence Identifiability - individual entities of each type should be individually identifiable, e.g. by primary key in the relational model. The Social Security Number of American citizens will not be effective if the database contains Canadian citizens as well. The term identifiability requires that each entity be different from the other which can be easily achieved by the primary key. Homogeneity - semantic data types are homogeneous. For example, attributes should apply to all entities of a type. Additionally, attributes should have a natural range but there is a possible danger as in the following example:

Department	Memo Number
A	100-199
B	200-299
C	300-399

What will happen in this case if Department A or B issues more than 100 memos? However, this scheme is sometimes useful as in numbering of apartments in high rise buildings. Thus, apartment 4C is on the fourth floor or apartment 5A is on the fifth floor. The data contains secondary information.

F. Consistency: The consistency of a view means that the definitions among related components have a consistent meaning and that the components are structured consistently. It means that the consistency of a view has two characteristics:

- semantic consistency
- structural consistency

Semantic consistency: Semantic consistency means that related components of a view must have consistent meaning. For example, must a class have at least one student? A team have at least two members? An employee work for one and only one department? Both domain and integrity constraints, including constraints on null values help insure consistency. The various constraints defined for the object relational model (ORM) are really integrity constraints and data models help insure semantic consistency.

Structural consistency: Structural consistency refers to the consistency of attributes across semantic types. For example, attributes should have some definition across different components.

G. Reaction to Change: Reaction to change may be broken into two subdimensions: namely, robustness and flexibility. Robustness is a views' ability to reflect real world changes and changes in user requirements without its design needing to be changed. Flexibility is the capacity of a view to change in order to accommodate new user of application demands.

II. Quality Dimension of Data Values

The dimensions of data quality pertaining to data values are: accuracy, currency, completeness and consistency.

Accuracy

Accuracy is the degree of agreement between a particular value or a set of values and an identified source which supplies a value or values. In a relational model, values come from a data domain. Accuracy is also defined as the nearness of a value of an attribute to some value in the attribute domain which is considered to be a true value.

Currency

Currency refers to the degree to which a specific piece of data is up-to-date in the data creations processes, timeliness between when data is due and when it is received. Whereas cycle time is the time a sub process takes to create or utilize values, i.e. the time needed to complete an acquisition or usage cycle.

Completeness

Completeness is the degree to which values are present in the semantic items that require them. For example, does an entity-relational data model have the proper number of entities? Of attributes? For the following example, assume we are working with a relational model (other model choices are possible). A mandatory attribute requires a value. If an attribute may or may not have a value, it is referred to as an optional value. An inapplicable attribute does not allow a value. For example, consider a student telephone number. Under the definitions supplied above, null can mean at least four things:

A. The attribute is applicable and the student has a telephone number, but it is not known.
B. The attribute is applicable but with a special value in the domain of student telephone number assigned because the student has requested an unlisted student telephone number.
C. The attribute is inapplicable. For example, the student does not have a student telephone.
D. It is not known if the student telephone number is applicable because it is unknown whether the student has a telephone number or not.

Consistency

Data are said to be consistent with respect to a set of constraints if they satisfy all constraints on the data model. Constraints may be on the same semantic entity or across semantic entities. A data structure is consistent with a view and represents a legal database state if and only if the data of that structure satisfie all constraints of a view. Thus consistency is necessary but not sufficient for correctness. For example, if the state of New York zip code is 07474, it may be semantically consistent with the constraints of a data model but is probably not correct because 07474 is the zip code for William Paterson University in the state of New Jersey. We must know the exact intended semantic association to denote the combination of New York and 07474. Of course, William Paterson University is not in New York, but does geographically face New York, looks out upon New York, etc.

III. Quality Dimensions of Data Format

Earlier we defined data as a component of entity, attribute and values (eav). In this section, we add format (f) and physical instance (I). Thus, datum = (e, a, v, f, I). Format is the mapping from the domain to a set of symbols. The quality dimension for the format is a component of appropriateness, clarity, universality, precision, flexibility, ability to represent null values, and efficient usage of the recording medium.

A. Appropriateness : Appropriateness refers to both the user and task and is frequently the most important format dimension. For example, data structured as a list (for example, the result of a structured query language - SQL - is usually presented as such) may be more appropriately presented to a decision maker in the form of a graph. As another example, data structured as a graph may not be appropriate input for a certain hardware device. Therefore, appropriate may vary across user groups.
B. Clarity : Clarity represents a correct interpretation of the values represented. For example, the meaning of data presented at a nominal or ordinal level may be less clear to a human being than data presented at an interval level or at a ratio level. (Recall from statistics that ordinal data are defined as data that can be ordered but whose difference cannot be determined or is meaningless, whereas interval data are data where differences can be computed but that may not have an inherent starting point and ratios are meaningless. Ratio data are like interval data but with an inherent starting point and for which ratios are meaningful.) After all, what is a cypher message without the cypher?
C. Universality or Portability: Universality promotes formats that can be used by as wide a variety of users as possible. For example, 3127 does not require knowledge of English where as three thousand one hundred twenty-seven does.
D. Precision: Precision refers to having distinct representations for any two elements. Precision is pertinent for quantitative data. Note that a data domain may contain an infinity of values, which no medium can represent. Precision also applies to other data formats such as the abbreviation format for labels.
E. Flexibility: This data dimension is the degree to which a format accommodates changes in a conceptual domain. It allows for changes in user views or in the storage media. As an example, consider a grading system that allows a format of only one digit, say “1” through “9” Is 1 the highest or lowest? This system does not allow for an expansion of grades while a one digit letter system does.
F. Ability to Represent Null Values: This refers to the degree to which a format can represent null values. For example, consider a motorist whose license reads “None” who receives a ticket intended for cars that do not have a license plate. Good formats provide ways to represent null values.
G. Efficient Usage of the Recording Medium: Efficient usage of the recording medium can conflict with other requirements such as flexibility. It is appropriate to store data values in a string format using codes instead of storing them as an image. If codes are used, they can be reproduced when needed.

SOLVING QUALITY PROBLEMS WHILE GAINING COMPETITIVE ADVANTAGE

In a manufacturing industry, the product manufacturing system uses the input raw materials to produce a physical product or output materials. Similarly, an information system uses input raw data to produce data products or input data. The data product from the information system may be treated as raw data in another data manufacturing system. To gain competitive advantage, organizations must treat input raw data and product data the same way they treat any product during the manufacturing process. Information systems can be studied from the same perspective of product manufacturing. Figure 2 below depicts an analogy between manufacturing data products and physical products.

Figure 2

		Product Manufacturing	Data Manufacturing
Raw Data + Materials from the Environment	Input	Raw Materials	Raw Data
	Process	Material Processing	Data Processing
	Output	Physical Product	Data Products

By treating raw data as materials for manufacturing a product, organizations will be able to identify some errors which otherwise could not be detected. For example, United Parcel Service (UPS) approached their business by understanding that to obtain their goals, customer satisfaction was paramount. UPS realized that to achieve that objective, the quality of data on which they based their strategic business decisions must be of good quality. The same approach has been applied by Federal Express which has made sure that their field information systems capture raw data, e.g., service request time, package pick-up time, delivery time and feed back from customers. Additionally, the captured data had to be processed correctly at various stages of the package delivery process. By doing so, UPS and Federal Express produce high quality data products that provide the basis for delivering customer service that has set the standard for their industry.

To ensure that data products are produced or manufactured according to certain quality specifications, the following characteristics must be considered:

- Quality requirements in raw data when captured from environment
- Quality verification of raw data during the work in progress and the final product
- Actions taken when the data products do not conform to the required specification

SOLVING THE QUALITY PROBLEM

Before any raw data are put in the system, the source of that raw data must be verified. This procedure will allow the user to implement data edits and correction procedures to meet accuracy standards. The portion of incoming transactions that fail to meet the requirement is eliminated. Second, after the input of the data, verification procedures for data accuracy, e.g., edits and the examination of randomly chosen data and records, are required to test for authenticity. This will also allow or enable the user to test the accuracy of that data stored in management information system records that might be in error. By following this procedure, the methods associated with the accuracy of data are examined. This will include: the portion of the income data that failed, the portion of income transactions that are erroneous and whether the stored data in the MIS record are in error for some reason.

To reduce the quality errors in any permanent fashion, it is of utmost significance to make sure clean data are entered into the database. This is a difficult problem that requires database edits, clean ups and provision of a data tracking system that will check all software, hardware, records, and databases, and analyze the records or the object data as they pass through the information process. In particular, a capture system which utilizes a tracking database may be needed to download records to the tracking database for tracking and analysis.

A capture system is a combination of hardware, software, and manual methods that collects samples of data as they pass through information processes. In continuous sampling, data arrive randomly over time with no rigid grouping by any characteristic. In particular, a date and time stamp of arrival and departure of the sampled data entity is written into the tracking record. The other methods to be used may include a statistical quality control to determine what assignable and special causes are present in the information process and the appropriate action to be taken at that particular time. An assignable cause, A.K.A. common cause, is a source of variation internal to a process, e.g., unintended software truncation in data. A special cause is a source of variation which is external to a process, e.g., bad data caused by a delay in receiving untimely transmitted data from another process.

RE-ENGINEERING OF BUSINESS PROCESS TO ENHANCE DATA QUALITY

Improving communication between sub-processes (especially processes that involve data exchange between people and machines) of an information process can lead to substantial improvement in data quality and hence organizations gaining competitive advantage in the process. Some strategies for changing the process that may help include the following:

- Establish a common language for the same information across databases. This will eliminate the need for normalization and translation changes at different points. It will also improve communication between people and computer processes.
- Reduce manipulation and handling of data (since they can incorrectly change the data). It will also eliminate intermediate disks, tapes and other forms of data transmission.
- Excluding unneeded processes is paramount to meet the data quality objectives. Some data that you get from the previous sub-process can be gotten from an earlier process, thereby bypassing intermediate processes where data can be corrupted.
- Improve input screens by using a single screen rather than forcing operators to switch between several different data entry screens. For example, what is the meaning of input data on the screen? Is input a verb or an adjective? Does it mean to take action by inputting the data or is it simply labeling the data above as input data? Use context sensitive online help that gives a list of standard entries for each data item and/or define normalizations, translations, and functional changes performed on the data. It is also necessary to check individual data items and records with a program that allows unacceptable input during manual entry.

To be competitive in this global environment, the use of this data must know the origin of that data. Much of the data today may come from outside the user's or corporation's group. If a supplier's data does not meet your requirements (company requirements), you can insist that the supplier institute data tracking or else seek another supplier. It is also important to inform the supplier of data of the consequences if their data does not meet the requirements desired by your company. External suppliers may be given rewards or penalties according to the accuracy of the data they supply.

DATA QUALITY REQUIREMENTS AND ENFORCEMENT

There are many types of data errors that can cause problems in any data generating environment. Errors are like weeds that occur spontaneously if they are not guarded and put under control. There are several steps that can be taken to enforce data quality in databases. The major ones include the following:

- Maintenance of schema quality
- Verifying data entry in fields
- Checking data dependencies
- Enforcing data constraints if known in advance
- Checking the compatibility of data schema

Necessary measures are needed to ensure that data schema are not changed. The consistency of data schema is paramount when different tables are merged. Additionally, some measures of quality may be enforced while data entry is taking place. Type and range checking must be performed. For example, ages cannot be a negative number and telephone numbers must have a certain number of digits depending on geographical location or country.

The best type of data quality enforcement is to prevent any error before it happens. For example, it is necessary to prevent the data error before it is entered into the system. It is necessary to check and find obvious mistakes; for example, inputting an invalid value for an attribute. It is also important to look for unlikely values and issue a warning where necessary to the data entry operator. Other methods that can improve data quality include the following:

- Using special values for unknowns to reduce or eliminate confusion
- Identifying and checking to make sure values are unique
- Ensuring validity across systems by the use of version numbers.

The other method of detecting data errors is by doing data quality audits. The first step is to list all types of errors and their frequency of occurrence. Sometimes it is not that easy because some errors are not easy to detect. One of the possible methods is to do data sampling from the population. The random sampling of data is the cheapest way of tracking errors and it will also eliminate extra costs. Possible methods to check errors in a data quality audit include the following:

- Checking for logical constraints within the fields, records and tables
- Checking for functional dependencies
- Range checking for integers and reals
- Checking for patterns in strings
- Checking for statistical constraints

Data must be treated like gold because it is the foundation of knowledge. The model for data processing follows three steps. The first step is ensuring the quality of that data being processed. The second step is analyzing that data appropriately; and the third step is to interpret the data and give the meaning. Figure 3 depicts the three step model of data processing.

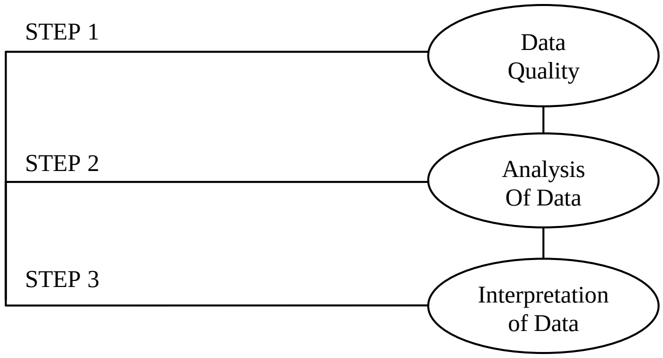

Figure 3

Data quality needs to be managed effectively to ensure effective results. Without quality data, business level strategies will be useless and the result will be an inability to compete effectively.

It has been argued that knowledge is power. Knowledge grows out of the interpretation of information which has been captured from different places. The interpretation of that information is founded on data. We can intuitively say with precision that bad data leads to erroneous knowledge which leads to poor decision making.

AUTHOR INFORMATION

Dr. Andrew B. Nyaboga earned his Ph.D at Stevens Institute of Technology, Hoboken New Jersey in 2000.Currently he is an associate professor of Accounting and Law at William Paterson University in Wayne New Jersey and teaches

courses in accounting and accounting Information System. His research interests are in technology management, knowledge management and strategic Management. His work has been published in numerous refereed journals.

Muroki F. Mwaura, William Paterson University of New Jersey, USA

REFERENCES

1. Arnold, S.E. (1992). Information Manufacturing: The Road to Database. *Quality Database*, 15(5), 32.
2. Baroudi, J.J. and Orlikowski, W.J. (1988). A Short Measure of User Information Satisfaction: A Psychometric Evaluation and Notes on Use. *Journal of Management Information System*, 4(4).
3. Bailey, J.E. and Pearson, S.W. (1983). Development of a Tool for Measuring and Analyzing Computer User Satisfaction. *Management Science*, 29(5).
4. Bailey, K. and Dunn, K. (1991). Six Keys to Quality Service. *Executive Excellence*, 8(9), 14.
5. Bailey, R. (1983). *Human Error in Computer Systems*. Englewood Cliffs: Prentice-Hall, Inc.
6. Bowen, P. (1993). Managing Data Quality in Accounting Information Systems. *A Stochastic Clearing System Approach*, Auburn University.
7. Delen, G.P.A. and Rijsenbrig, B.B. (1992). The Specification, Engineering and Measurement of Information System Quality. *Journal of Systems Software*, 17(3).
8. Haavind, R. (1992). Federal Express Wins in the Tough Service Category. In *The Road to the Baldridge Award: The Quest for Total Quality* (pp. 71-80). Stoneham: Butterworth Heinemann.
9. Huber, G. (1990). A Theory of the Effects of Advanced Information Technologies on Organizational Design, Intelligence and Decision Making. *Academy of Management Review*, 15.
10. Parsage, K. and Chignell, M. *Intelligent Database Tools and Applications*. John Wiley & Sons, Inc. New York.
11. Percy, T. (1993). Business Re-Engineering: Does Data Quality Matter? *Software Management Strategies*, Gartner Group, 1.
12. Porter, M. and Miller, V.E. (1985). How Information Gives You Competitive Advantage. *Harvard Business Review*, 63(4).
13. Prahalad, C.K. and Hamel, G. (1990). The Core Competence of the Corporation. *Harvard Business Review*, May-June.
14. Redman, T.C. (1992). *Data Quality: Management and Technology*. New York: Bantam Books.
15. Ronen, B. and Spiegler, I. (1991). Information as Inventory: A New Conceptual View. *Information and Management*, 21, 239-247.
16. Sack, I. (1995). Lecture Notes on ΑData Quality@. Stevens Institute of Technology.
17. Taguchi, G. (1981). *On-Line Quality Control Production*. Tokyo. Japanese Standards Association.
18. Te=eni, D. (1993). Behavioral Aspects of Data Production and Their Impact on Data Quality. *Journal of Database Management*, 4(2).
19. Wang, R; Storey, V. and Firth, C. (1993). Data Quality Research: A Framework, survey, and Analysis. Sloan School of Management.

NOTES