7th International Conference on DEVELOPMENT AND APPLICATION SYSTEMS
Suceava, Romania, May 27 – 29, 2004

DATABASE QUALITY- SOME PROBLEMS

Augustin-Iulian IONESCU, Eugen DUMITRAŞCU
University of Craiova
Str. A.I. Cuza Nr 13
iai100@k.ro

Abstract: *In this paper some specific problems concerning the database and information quality are presented and analyzed. There are emphasized some approaches presented in the literature and some "classical" mistakes encountered in practice.*
Keywords*: data quality, information quality, database quality, data integrity, database design.*

Introduction

Issues concerning data quality and information quality were presented by diverse authors from the beginning of database utilization but the reality emphasizes today a lot of misunderstandings and wrong approaches in database design and implementation. In the last years the increasingly dependence of decision level on the information provided by the information systems of the enterprise and the development of new forms of data collections like data warehouse, data marts, mobile databases, Internet (view like a huge database) implies the reconsideration of the old concepts developed in the early 70. The last years emphasized the costs of poor data quality not only in the businesses area. Similar costs can be found in governmental or educational organizations as well. Poor data quality is sapping all organizations of money and opportunities. Some areas in which costs are created and opportunities lost through poor data quality are (Olson, 2003) transaction rework costs, costs incurred in implementing new systems, delays in delivering data to decision makers, lost customers through poor service, lost production through supply chain problems.
Obviously, the problem of data quality and information quality must be tackled in the frames of general quality theory developed by the American and Japanese quality gurus beginning with the middle of the twentieth century and must satisfy the specifications of the ISO 9000:2000 standard.
Despite of the differences which exist between the position of the quality gurus in many aspects concerning the analysis and the implementation of the quality in the real enterprises, there are some common positions, namely:

- The quality is defined as "the fitness with the costumer expectations";
- The implementation of quality is a management problem not one of the workers;
- The benefit of quality must be measured as the costs of nonquality not as a bonus for the enterprise;

Surveys of the most popular approaches concerning the quality are presented in the last two references.
The implementation of the different methods proposed for improving the quality of the activity in the enterprise is difficult either because of the misunderstanding of some fundamental concepts at the management level or of the lack of motivation for the executive levels.
One typical error is the orientation of all resources to the treatment of the effects, not the discovery and elimination of the sources of errors. This approach determines a permanent activity of inspection and correction in total contradiction with the principle <<make it good

for first time>>. Also, as Philip Crosby remarks (Crosby 2001), many peoples ignore the existence of several views of quality:

- from management;
- from quality professionals;
- from employees;
- from the customers.

Adding to these views the different interpretations of those implied in the implementation, a lot of visions results, sometimes irreconcilable.
The struggle for the quality emphasizes a basic idea, unanimously accepted (at least at the declarative level) - even if the management must impose the quality standards in the enterprise, the imposed objectives realization implies **all** the employees. One speaks of a **quality culture** in each enterprise.

The Main Problems in Database Quality

The database schema quality

A recent report published by the Standish Group shows that 37% of projects in the information systems get cancelled, with another 50% completed but with at least a 20% cost and time overrun and often with incomplete or unsatisfactory results. This means that only 13% of projects are completed within a reasonable time and cost of their plans with acceptable outcomes. Failures are not isolated to a small group of companies or to specific industries. In almost all the drawbacks of projects in the information systems the poor database quality is involved. Many peoples blame IT for this situation but data is created by people outside IT, and is used by people outside IT. IT is responsible for the quality of the systems that move the data and store it. However, they cannot be held completely responsible for the content. Much of the problem lies outside IT, through poorly articulated requirements, poor acceptance testing of systems, poor data creation processes, and much more. The fact that data quality is universally poor indicates that it is not the fault of individually poorly managed organizations but rather that *it is the natural result of the evolution of information system technology* (Olsen, 2003). There are two major contributing factors. The first is the rapid system implementations and change that have made it very difficult to control quality. The second is that the methods, standards, techniques, and tools for controlling quality have evolved at a much slower pace than the systems they serve.
The problem is to determine when a database has an acceptable quality. In other wards this means to determine the dimensions for assessing database quality. In [7] there are presented a lot of sets of dimensions used for the evaluation of the database quality.
In principle the database quality implies:

- the database design quality
- the data integrity
- the data precision
- the data relevance
- the data protection and security
- the documentation quality

The database design quality must be analyzed at all the three levels - conceptual, logical and physical. The poor quality at one of these levels implies the fail of the project but the correct solution at each level do not guaranties the success of the project. There are some explanations for this situation. One explanation consists in the fact that in many projects each application uses its one database, even the databases are practically identical. An example is the information system for CJAS Dolj, which uses three files for the patients in three different applications. The existence of three files with practical same content induces a high level of redundancy with all well-known consequences. Another problem is the caption of all interesting data from costumer point of view. In the above presented example the lack of the possibility of identification of the tutor of a child or of the contact telephone for a person create big problems and frustration. Such problems appear because the database designer is concerned only with the solution of one application. This is the classical error and not only for the beginners - a database is dedicated only to one application.

In other cases the structure of the database is too restrictive and any modification in the requests on the database contents raise the necessity of schema changing, the changes in the data integrity rules or in attribute semantic. A well-known example is the sub estimation of the length of the attribute Name. The appearance of a patient with a longer name implies the modification of the definition of attribute and the redesign of the reports. Apparently very simple, this situation generates irritation and lost of time.

All this presented cases are the consequence of a poor analyzes in the first stages of design process, namely the miscorrelated analyzes and implementation of applications respectively the inadequate generalization of some conjectural values.

The modification of the database structure is imposed either by the design errors or by the necessity to assure the concordance between the structure of database objects and the changes appeared in the business environment. It is the designer task to anticipate the possible changes in the business environment or in the customer needs and to create a database scheme that supports these modifications. For example, if the relationship between two relations is at the current moment of type 1:M but it is possible that some changes in the business rules will transform this relationship in one of type M:M, the relation will be treated from the beginning like being M:M.

The normalization of the database structure is one of the sources of the poor behavior especially in the large databases. Many author emphasize that a perfect designed database may be useless because the time of the information reconstruction is unacceptable. In the last years many authors consider that the *Boyce-Codd Normal Form* or in some situation the *Third Normal Form* are sufficient for practical proposals. A well controlled redundancy of data is sometime preferable if assures a reasonable time for information retrieval. The new generation of DBMS for little systems offers a relative poor set of join operators, especially they do not offer the possibility to perform the outer join or reunion. For the simulation of outer join it is necessary to use a kind of data redundancy, namely a line for each possible value of foreign key, even the line contains only null values but the foreign key.

The relation fragmentation is generally related with the distributed databases but there are some cases when the fragmentation may be used in a centralized database for reducing of the time necessary for the realization of a join or for the minimization of the interactions between different users of the database. A well performed fragmentation can also increase the flexibility of the application implementation.

The importance of the conceptual model of a database is accepted today by almost all database designer (exist yet some designers which consider that the conceptual modeling is important only for those who are not capable to design an effective database) but the existence of a lot of conceptual models, each with countless dialects, the lack of standardization, decrease the impact of conceptual design. In despite of what their promoters clams, each conceptual model has weak points. The ER models are very poor in the capture of business rules and only in last years some variants borrowed some concepts from other models for the representation of more sophisticated data integrity constraints. The ORM models are better concerning the caption of complex integrity constraints, the readability of the model and the transformation in the relational model are the weak point of this model. The object-oriented models are generally too complex and they did not replace the old models in the database design practice.

Batini, *et al*. [1] propose and details a set of criteria for assessing database schema quality: completeness, correctness, readability, minimality, self-explanation, extensibility, expressiveness, and normality.

In the present paper is discussed only the first criteria, the completeness. *A schema is considered complete when it represents all relevant features of the application domain*. In principle completeness can be checked by looking in detail at all requirements of the application domain and checking to see that each of them is represented somewhere in the

schema and on other hand checking the schema to see that each concept is mentioned in the requirements. The problem is *who guaranties the correctness and the completeness of the requirements.* A set of requirements may be complete from the point of view of one user but incomplete from the point of view of other users. For example the telephone number of a patient is very important for the current doctor but it is not relevant for the statistical applications of CJAS staff. Some users view the same data in different forms each useful for one proposal. The independent application for one user can become only a component of the application for others users or some application can be integrated in a one more useful application. In many situations the requirements referred at the old experience of the user and really not feet the user's expectation. The difference between current data and historical data or between strategic and tactical application is generally difficult to understand for the users. All this situations and a lot of others must be emphasized during the first step of the design process, the analyze of information system. Analyze is a very difficult activity and the poor result of analyze can affect all the design process. A well conducted analyze can put in evidence a lot of anomalies in the analyzed domain and eliminates many sources of the dirty data. A classical mistake consist in the design of a database scheme taking in account only one application or only the needs of only one department and neglecting the interaction with others application or the possibility of the implementation of other applications sometime in the future. The integration of the databases or applications only after the implementation is a poor practice which consumes a lot of human and financial resources.

There are situation when an apparent good design, acceptable from the user point of view can generate complicated problems in the future. By example, if the attribute Address is considerate like an atomic string of characters, if in the future it is necessary to develop a data warehouse or a global database, the identification of the component of each address will be a laborious and delicate problem.

The data quality

An important problem is the confusion between some concepts:

- Some authors consider the terms *data quality* and *information quality* as being synonyms. In fact the two terms refer two distinct concepts, namely *data* and *information*.

Taking into account that the databases are utilized like basic elements in the decision process, the information can be defined as <<the incertitude removed concerning the realization of an event among a set of possible events>>.

The data can be considered as <<information removed from context>> or <<the representation of information on a material support>>. On one hand this means that the same data can represent different information and on the other hand the same information can be represented in a lot of forms (data).

The transformation of the data in information implies the placement of the data in a well-defined context and the interpretation in according to the user needs, his/her culture and the technical resources. The information is the result of the data processing by human or automatic means.

Evidently, even if the data quality is guaranteed, the resulted information can be of a poor quality or even can be wrong. There are some causes that determine these situations to occur:

- An unacceptable confusion, which persists especially in some product documentations, is that between the terms *data validation* and *data accuracy* (data precision).

In fact, data validation supposes the checking of the data values for the allegiance at a set of values defined in the data integrity constraints. Such tests are very easy realized by means offered by DDL or by triggers. It is well known from the beginnings of the databases that the validation of the data is not the guaranty for its accuracy.

Olson emphasizes [12] two characteristics of accuracy: form and content. Form is important because it eliminates ambiguities about the content. The representation of a date can create many misunderstanding because the user would not know whether the date was invalid or just erroneously represented or what is the "real"

date. The date 11/02/2004 means 11 February 2004 for an European but 2 November, 2004 for an American. A value may not be considered accurate if the user of the value cannot tell what it is.
The uniqueness of data can also create big problems. Data can be different if the option "case sensitive" is activated but they are not if this option is deactivated.
The guaranty of data accuracy is a very sensitive problem, which imposes a permanent activity for correcting, fixing and analyzing corrupted or erroneous data. For this activity the organizations spend a lot of time and resources. Larry English consider [4] that <<the assessment of data quality has value only if it is used to awareness of process failure and results in process improvements that eliminate the causes of defective data. The *ultimate* goal of information assessment must be to assure that processes and maintaining information quality that consistently meets all information customers' requirements >>.

It must emphasize the existence of two distinct situations in the approach of data accuracy:

- Because of the moral or/and legal considerations, no error can be admitted. For example, there is a little probability that a candidate accept to lost the first place because the result was registered with one *acceptable* error. In such cases the Philip Crosby principle "zero tolerance" is imposed.
- In some situations the accuracy has a statistical meaning and the errors are tasted on a sample of data according to statistical principles. Such situation is acceptable for a census.

There are also more subtle aspects of data accuracy. For example it is possible that a patient presents corrupted data concerning the incidence of chronicle diseases in his family because he doesn't know that his parents are not his biological parents and that the information concerning his relatives are not relevant for the medical assessment.
Another example is represented by the declarations about the number of birth or abortions because many women consider these problems like crucial secrets of there live.
The data accuracy is time dependent. The assertion concerning data accuracy is not consistent if it doesn't precise the moment of data assessment. This means that the actuality of data is a very important component of data quality. The use of non-updated data can result in very poor conclusions, sometimes with catastrophic effects upon the evolution of the enterprise.
The NULL values can create a source of misunderstandings because sometime the null means "nonsense value" and other time NULL means "current unknown values" or "not important data" or "I'm not sure". The use of default values is another source of poor data quality. Many people consider that the effort to modify these values is not justified if they do not affect theirs immediate interests. The ambiguity in interpretation can generate big anomaly in statistical applications or data mining.
A problem not sufficiently discussed is the meaning of concept *accurate data*, who must guaranty the data accuracy and which are the legal responsibility for the provider of poor data quality. In fact the problem is with *what* must be compared the registered data, in the data accuracy assessment process. A possible answer is the comparing with the primary documents when these exist. But sometime, especially when the primary documents are holograph documents this process can create more confusion. Another possible solution is the comparing with data already registered in another database that is considered sure. Many blame IT. However, data is created by people outside IT, and is used by people outside IT. IT is responsible for the quality of the systems that move the data and store it. However, they cannot be held completely responsible for the content. Much of the problem lies outside IT, through poorly articulated requirements, poor acceptance testing of systems, poor data creation processes, and much more.
The chosen solution whatever, a legitimate question arises: "*how correct are the correct*

considered data?" Obviously, from moral and legal point of view, the concordance between the registered data and those presented in the primary documents or in the witness database represents a guaranty of data accuracy. But for the data mining workers this guaranty is not sufficient. Really, such guaranty is not possible. For example, who can guaranty that the evaluation of a candidate represents the true measure of his knowledge? In some special situations (for example in contests) more examiners/arbitrators are utilized and a mediation of the results is made, but this is only an acceptable compromise.

Taking into account the problems discussed above the conclusion is that from the informational point of view the data accuracy *is a trust problem*.

The introduction of data marts and data warehouses in the architecture of information system in many enterprises had as first objective the elimination of wrong data. The analysts shall be provided with clean data, no matter what was the source of these data. The identification of sources, the extraction of data and data cleansing are sometimes laborious and expensive processes and the result may be a collection of data without resemblance with any data source. The cleansing process does not eliminates the errors at the source and the data warehouse user must be informed about the source of the data and cleansing procedure for understanding why a source was preferred. This situation brings into question again the problem of the trust in the data accuracy and data timeliness.

The increased utilization of multimedia documents modifies the idea about what precision means. Defining information quality as the fitness of the information obtained with the customers' expectation, it is very difficult to objectively define the quality of a picture or of a song registered in digital format. The quality of multimedia information is sometime essentially dependent on the quality of display, printer, scanner or boxes and less on the quality of data. For example an excellent representation of a landscape on the display is unrecognizable in the copy realized with a chip printer.

The data availability is less discussed in the literature as an aspect of data quality. This topic is generally related with the very large databases for which a 7x24 availability is frequently imposed. But in our days the problem of data availability is a real problem for all the users of an databases however tiny it is. Sure, the data availability problem is strongly related to the hardware of the systems on which databases are implemented but sometimes the impossibility of access the data is caused by the necessity of the reconstruction of a database or the redistribution of the data on the secondary memory or the lack of space on the disk or in the main memory. Many users don't understand that the space occupied by the data is only a part of the secondary and main memory necessary in the search process.

An important component of data quality as the customer perceives it, is represented by the time wasted in obtaining the answer at the query. Apparently the value of this parameter is very strong related with the hardware quality but there are many situations when it is especially dependent on the information system architecture. A typical example is when the customer is very content with the database quality in the first month of exploitation but in time the value of the access time to the data becomes unacceptable. Because the decrease of performance is obviously related to the increase of data quantity, the selected solution is frequently the acquisition of a more powerful server or of a more sophisticated (and expensive) DBMS. In some cases, as a result of a more detailed analysis it comes out that the new acquisitions resolve the problems only on short term because the data volume continually increases. This happens because of a wrong approach of the database design, namely the introduction of current data and historical data in the same table, which lead in time to the accumulation of a huge data amount and the increase of access time. Many of the data preserved in the database are seldom used in particular applications and some data haven't been used for months or years. In such cases the separation of current data from the historical

data can produce a substantial decrease of the access time without notable additional expenses.
The large use of data replication in the distributed databases created new problems concerning the data consistence because despite of the guarantees claimed by the software firms, there are no possibilities to assure a perfect real time synchronization between the different replicas. A particular case, frequently encountered in the regions with a low level of the development of computer networks is a primitive form of mobile computing, namely the periodical load of the data in a personal computer from a centralized database, their processing and use at local level and then the transfer of the results in the centralized database. The lapse of time between two synchronizations is sometime weeks or month. The result of such policy may be disastrous, with negative effects in the economical and social evolution of an enterprise. On other hand there are situations when the lake of synchronization is not very important, like in the case of updating for a list of available products or the communication of the discovery of a new butterfly in the Amazonian jungle. In conclusion the problem of the data synchronization must be evaluated for each case in concordance with the customer's expectations and the real cost of the data nonquality.
Persons or organizations collect, store and manage a huge amount of data, but don't usually document where it came from or what is should be used for. Such data become in short time useless. For avoid this situation, a guidebook of metadata must be developed for the users to know how to interpret the data and how to use it.
In many of the above-presented cases, the information obtained from the data is predictable and the customer can evaluate how the results cover the expectations. This means that the customer know very well what fields are expected and sometime how many registrations will be obtained. He can verify if the result contain all the needed data and only the needed data. Because users assume a centralized DBMS to always have all the information to provide a *complete* result, the quality of such a system is measured by its response time.
There are however situation when the customer doesn't know what the result will be. This happens when the user obtains the information from Internet. Many free applications can be found in the Web such as meta search engines, integrated stock information systems, or bibliographic services. Each of these programs must integrate the information obtained from many independent and autonomous sources. The response time of an Internet source is less important compared to its ability to provide the information queried for. Inherently, nobody claims *all* the information concerning his field of interest but each user is interested for the *relevance* of the obtained information. To gain the full advantage of multiple sources a user must query all available sources and integrate the results. Even if automated, this is a tedious and often expensive process, and some criteria to determine which sources are to be preferred over others. The measure of source relevance must take into account both the number of objects provided by the source, and the amount of information per object it provides. Another problem is the impossibility to repeat in time the same query with the same result and the difference between the results provided by different search engines. Such elements arouse a frustration sentiment for the customers.

Conclusions

No matter how the data are obtained, how they are stored, what data models are used, information quality plays a crucial role. The temptation to focus only on the data quality is wrong because the goal of a quality program is to improve the quality of the product. The quality method must focus on the information "costumers" to understand their quality requirements in order to perform their processes efficiently and effectively. Information customers include all persons within the enterprise who need information to perform their work and all stakeholders, including end customers and shareholders, who depend on information [4]
In a time of decreased budgets, the cost of poor data quality is very important. According to

several leading data quality managers, this cost may be expressed as:

Cost of Poor Data Quality=
Cost to Prevent Errors + Cost to Correct Errors +
Cost to "Make Good" for the Customer

Some organizations spend all their time fixing problems leaving no time to prevent them.
The quality projects are not chip. But the prize of nonquality may be most expensive.
The lake of quality reduces return on investment, diminish staff productivity, harms the credibility of the organization.
The misunderstanding of the concepts of data quality or their naïve application can result not only in lost of money. Sometime the poor information quality may kill. Two catastrophes do to poor information quality are analyzed in [8].

References

[1] Batini Carlo, Ceri Stefano, Navathe Shamkant (1992) *Conceptual Database Design: an Entity-Relationship Approach* - The Benjamin/Cummings Publishing Company, Inc.
[2] Crosby Philip (2001) *The Views of Quality* - PCAII, Inc, Articles, April-June
[3] English Larry (2000) *Plain English on Data Quality: Seven Deadly Misconceptions* - DM Review
[4] English Larry (2002) *The Esssential of Information Quality Management* - DM Review
[5] English Larry (2003) - *Total Information Quality Management* - DM Review, September
[6] Fisher Craig, Kingma Bruce (2001) "*Criticality of Data Quality as Exemplified in Two desasters*" - *Information&Management* 39, 109-116
[7] Hoxmeier John - *A Framework for Assessing Database Quality* - http://osm7.csbyu.edu/ER97/workshop4/jh.html
[8] Fisher Craig, Kingma Bruce (2001) *Criticality of Data Quality as Exemplified in Two disasters* - Information&Management 39 109-116
[9] Naumann Felix - "*From Databases to information Systems-Information Quality Makes The Difference*" www.citeseer.nj.nec.com/naumann01from.html
[10] Naumann Felix, Rolker Claudia - "*Assessment Methods for Information Quality Criteria*"- www.citeseer.nj.nec.com/naumann01from.html
[11] Naumann Felix, Johann Cristoph Frytag, Ulf Leser - "*Completeness of Information Sources*" - www.citeseer.nj.nec.com/naumann01from.html
[12] Olson Jack (2003)- "*Data Quality—The Accuracy Dimension*" - Morgan Kaufmann Publishers
[13] Rolker Claudia, Kramer Rolf (1999) *Quality of Service Transferred to Information Retrival: The Adaptive Information Retrieval System* - Proceedings of the 1999 ACM CIKM International Conference on Information And Knowledge Management, Kansas City, Missouri, USA, November 2-6, ACM press, pages 359-404
[14] Sakar Pushpak (2002) *A Paragon of Quality - Intelligent Enterprise*, October 8
[15] Shultz David (2002) "*Facing Facts*" - PCAII, Inc. August 2002
[16] *** - *The quality gurus* - www.dti.gov.uk/mpb/bpgt/m9ja0001
[17] *** - *Views of Quality Gurus* - www.geocities.com/sandeep_Kumar_chhabra/compare.htm