Data Quality Evaluation Through Non-Conformity

Selma Foligne Crespio de Pinho[1], Geraldo Xexéo[2], Ricardo Barros[3]
[1]Brazilian Navy
[2]DCC/IM/UFRJ and PESC/COPPE/UFRJ
[3]Brazilian Navy and PESC/COPPE/UFRJ
selma@casnav.mar.mil.br, {xexeo, rbarros}@cos.ufrj.br

Abstract

We present a three-step process for data quality evaluation based on non-conformities that is compatible with the ISO 9000 family of standards. The first step uses expert opinions to establish weights that convey the importance of data quality characteristics, capturing the expert's perception and expectation about data quality. The second step measures how much the database satisfies the users, using a predefined relationship between quality characteristics and types of non-conformities found using the system. The third step uses graphs and reports to allow the monitoring of the data quality of a system during its life cycle. We also developed the AQUA prototype and used it in a real case to validate the model. As a result, users are able to monitor data quality and act to improve it when necessary.

1 Introduction

Data quality assurance is an approach to avoid erroneous decisions and consequent financial losses [1]. We propose to assess data quality by non-conformities. For that, we investigate and analyze the subjective data quality characteristics important to the users, and then consider and monitor how non-conformities affect the perceived quality of data. In addition, we describe a data quality evaluation experiment using an automatic tool called AQUA.

Data quality evaluation is a practice that can be accomplished by two approaches [2]:

(i) Quantitative evaluation (objective) of the database. In this evaluation form, objective indicators are used to measure database quality regarding representation and structure.

(ii) Qualitative evaluation (subjective) of the database. In this form, subjective indicators are used to assess database quality and its usability. These subjective indicators are quality characteristics defined to capture the users' perception and expectation about data quality.

In this presentation, Section **2** gives a short introduction to data quality concepts. Section **3** lists the quality characteristics that were identified as necessary to evaluate a database qualitatively and lists the types of errors associated with each one of these characteristics. Section **4** describes the mathematical model for obtaining the quality index of data from non-conformities. Section **5** describes the necessary stages to accomplish the evaluation of the proposed data quality integrally. Section **6** describes the evaluation experiment and the AQUA prototype. Finally, Section **7** presents the conclusions and future perspectives of this work.

2 Data Quality

Quality is a multidimensional concept. Normative document ISO 9000:2000 [3] defines quality as "the totality of characteristics of an entity that allow capacity of satisfying explicit and implicit needs". Explicit needs are defined as those expressed in the producer's definition of the proposed requirements. They are composed by the utilization terms of the product, their goals, functions and expected performance. Implicit needs are those necessary for the users, although not expressed by the producer.

One can identify three main approaches to data quality in the literature: theoretical, empiric and stochastic. The theoretical approach focuses on the data that could become deficient during the production process. For example, WAND and WANG [7] have defined the dimensions of data quality using ontological concepts, based on problems that happen in the mapping of data, from real world to information systems. That study comes from the observational fact that the development and use of

information involve two transformations: the representation and the interpretation ones. This approach proposes that data deficiency could happen during the representation and/or interpretation transformations, generating, in this way, a lack of conformity between the vision from the real world and the one obtained from the information system.

In the empiric approach, data quality attributes important for users are captured. The data collected by the users is analyzed to define the characteristics that will be used to assess if it is adjusted to its tasks or not. This approach is used when data quality is based on experience or understanding of which, from the users' perspective, are the important attributes [8,9,10,11].

The stochastic approach incorporates a set of internal controls in information systems to increase the ability of these systems to prevent, discover and eliminate errors [12].

We adopt the quality model proposed by ROCHA [4], which uses the following concepts: quality goals, quality factors, criteria, evaluation processes, measures and aggregate measures.

Quality objectives or goals represent the general properties that a product should possess. Each goal is subdivided in factors, which can be further broken down in sub-factors. Factors and sub-factors define different users' perspectives about the quality of a software product.

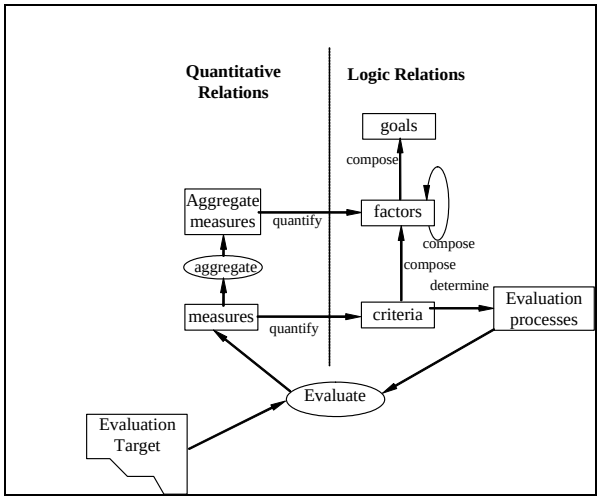

Figure 1 - Rocha's Model of Quality

Moreover, factors should be divided in primitive quality attributes that can be measured, called criteria. One or more alternative evaluation processes must be established for each criterion.

Measures can be objective or subjective. Most of the time we aim for objective measures, but the special characteristic of specific domains, like software development, sometimes forces us to appeal to subjective analysis.

The numerical or qualitative measures that quantify the criteria are what remain after the evaluation is complete. These measures must be aggregated to quantify the factors. The relations among measures and aggregate measures are known as quantitative relations.

Based on Rocha's Model, this work defines the goals, factors and its respective sub-factors, evaluation criteria and process in the database context. The application of Rocha's Model is described in Section **3.**

2.1 Quality and Non-Conformity

In Statistical Process Control (SPC), one traditional method of analyzing the quality of a process is by using control charts for non-conformities. A non-conformity is a defect according to some specification of a quality attribute. The Control Chart counts non-conformities detected in the units produced by the process, be it a manufacturing one or not.

When counting non-conformities it is also possible to distinguish between the degrees of effect that a defect can have in a produced unit. This is very interesting for complex units that can be affected in

different ways by different non-conformities, and with different results, according to the users` perspectives. This approach is known as a "Demerit System", which can be defined in the following classes [12]:

- Class A - serious or critical defects that affect the product's essential function, preventing its use;
- Class B - serious defects that reduce the the product's efficiency ; and
- Class C - small or irregular defects that, without modifying the performance of the product, constitute imperfections.

An unit's demerit value can then be calculated by multiplying the number of non-conformities of each class by the weight of that class. Each class has a weight that is defined according to the specific problem for which this technique is adopted. The result of this procedure is the weighted average of non-conformities in a product. In this way, a quality evaluation by non-conformities uses all non-conformities found, taking into account the weight of each one of them, to verify if the quality of the evaluated product has degraded.

3 Quality Factors for Data and Associated Non-Conformities

In this section, we define the list of data quality factors and sub-factors. This is organized according to three quality objectives: usability, conceptual reliability and representational reliability [13,4]. Further on, we associate one or more types of errors to each quality attribute, defining, in such way, the non-conformities that influence them.

To monitor an evaluation process based on non-conformities, we use a single criterion for each quality attribute, based on the number of failures or non-conformities that occurred during the use of the database product. This also defines a single evaluation process, i.e., the counting of non-conformities during a definite period of time.

In this way, all non-conformities or errors occurred should be registered, so that the quality degree of the stored data is measured, based on how the form of each quality characteristic was affected by these errors.

A detailed account of this research can be found in PINHO [14], but this model can be substituted by any other quality model based on the concept of quality factors, i.e., ISO- 9000 compatible.

3.1 Quality Factors

To reach the necessary adequacy of this list, we collected the opinion of 27 (twenty-seven) specialists from academic, military and enterprise institutions. It was necessary that they acted as data consumers that, from several perspectives, use regularly data to make decisions. Also, to identify the most adequate participants, we also studied their professional profile, that is, education degree and experiences with information systems.

Each one of then attributed a weight which varied from 0 to 4, to each characteristic, from a initial set of quality characteristics to express their importance. The proposal was to eliminate the quality characteristics that were not related to data quality according to the majority. The inclusion of a new quality characteristic occurred whenever some specialist identified its absence from the initial list.

The result of that research was a set of characteristics able to represent users' quality expectations regarding a database product to be evaluated, which is found in Table 1.

3.2 Types of Errors Associated with the Characteristics

During the field research with the specialists, they also analyzed an initial list of possible type of errors associated with the characteristics. The inclusion of a new type of error occurred whenever any specialist identified its absence from the initial list. At that moment, they also attributed a weight to it, in order to express the degree of importance of the new type of error.

Table 2 shows the type of errors which were collected and validated. These errors were associated with the affected characteristics.

Table 1 – Objectives, Factors and Sub-factors of Data Quality

Objectives	Factors	Sub-factors
Usability	Adequacy	Availability of information / Age of data
	Efficiency	Opportunity / Efficiency of execution
	Applicability	Relevance / Utility
	Profitability	Lucrativeness / Aid at user work / Competitiveness
Conceptual Reliability	Believability	Appropriate amount of data / Accuracy / Completeness Coverage
	Integrity	Robustness / Precision of data / Consistency / Easy of signaling / Accountability
	Functionality	Retrievability / Flexibility / Interoperability / Security access
	Legibility	Understandability / Adequacy of information
Representational Reliability	Uniformity	None
	Manipulability	Availability of documentation / Traceability

Table 2 – Type of Errors Associated to Data Quality Characteristics

Characteristics	Type of Errors
Availability of information	Unavailable data
Age of data	Data stored without update for long time
Opportunity	Answer delay turning it not useful
Efficiency of execution	Data recovery delay / Data record delay / Execution task delay
Relevance	Offered service is not accomplished in correct or complete way
Utility	Functionality is not important to organization tasks
Lucrativeness	Low productivity or financial loss by difficulty to use of data
Aid at user work	Use of offered services complicating users' work
Competitiveness	Use of data does not help the acquisition of market advantages
Appropriate amount of data	Insufficient quantity of data
Accuracy	Data manipulation results in incorrect information
Completeness	Lack of necessary field in forms / Lack of information
Coverage (depth)	Information offered by data is not enough
Robustness	Error-insertion in database after abnormal situation of system operation
Precision of data	Stored data does not represent correctly its meaning in the real world
Consistency	Two or more different values stored in database
Easy of signaling	Lack of alert to indicate incorrect or non-conform data entry / Lack of alert to indicate incorrect data manipulation
Accountability	Absence of record of data modification or manipulation authorship
Retrievability	Absence of mechanism to recover the affected data, in case of failure / Delay or difficulty to recover affected data, in case of failure
Flexibility	Difficulty in data manipulation (expansion adaptation or aggregation)
Interoperability	Impossibility of interacting with other database
Security Access	Absence of safety mechanism to access control of system / Lack of definition of data access scope allowed for each type of user
Understandability	Little objective information generating doubts
Adequacy of information	Understanding demanded by information is incompatible with the users
Uniformity (no subfactors)	Unsatisfactory data presentation
Availability of documentation	Lack of documentation in order to assist data verification and data localization process / Lack of documentation that enables data association with its source
Traceability	Absence of mechanism that enables to trace a information in order to locate it

4 The Model

4.1 Mathematical Model

The whole model is based on the Statistical Process Control of quality that suggests the use of the Demerits System, as we have seen in Section **2.1**, to observe how the non-conformities that occurred can help in a continuous effort to improve a process [5,16]. The Demerits System is used to classify the non-conformity according to its severity, because it can be inadequate to consider that a product is perfect (flawless) or imperfect (with defects). It is also interesting to assess the frequency with which defects occur in each unit.

The model is divided, basically, in five steps that express the general idea of the evaluation criterion. In this way, the demerits system is used and applied in a hierarchical structure, as follows.

4.2 General Idea of Criterion Conception

Whenever a quality characteristic is affected by some non-conformity, the general quality of database product (DBP) is degraded.

Figure 2 shows that error-occurrence Q*i* affects the associated characteristic C*i*, which then affects the quality of stored data in DBP. The variables have the respective meanings:

DBP - database product

C*i* - *i*-th quality *characteristic o*f DBP

W_i - weight of the characteristic C*i*;

e_{ij} - type of error *j* associated with the characteristic *i*, where each type of error has:

q_{ij} - weighted average quantity of error type e*ij* that occurred , considering the severity of the error-occurrence; and

p_{ij} - weight of the error type e_{ij}.

Q*1* , Q*2*, ... , Q*i* , ... , Q*m* - average rate (error/min) of error-occurrence that affect C*1* , C*2*, ... , C*i*, ... , C*m*, respectively.

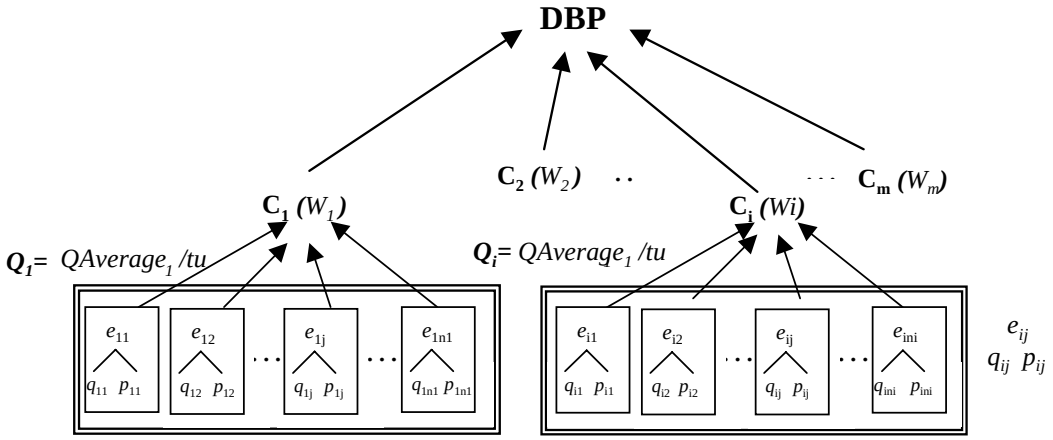

Figure 2 - General Schema of the Evaluation Model

Considering $Q_i = \frac{QAverage_i}{tu}$, where $QAverage_i$ is the weighted average quantity of errors that affect the C_i, and *tu* is the accumulated time of use of the product until the evaluation moment (expressed in minutes).

Considering that *use* is given by the time of use of the product that is being evaluated, that value can be obtained, for example, through the access control to a section *logout - login*. If *tu* is the accumulation of these values until the moment of the evaluation, then *tu* is constant for all $QAverage_i$.

Step 1: Calculation of q_{ij}

The variable q_{ij} corresponds to the weighted average of error type e_{ij} that occurred . For this, it is necessary to verify the error-occurrence together with the degree that expresses its severity for each type of error. It should be noted that an error-occurrence can have a different severity from another one, even when it belongs to the same type of error. That severity range can vary from 0 to 4, as shown in Table 3. This way, we have:

$$q_{ij} = \sum_{k=0}^{4} o_{ijk} . g_k \Big/ \sum_{k=0}^{4} g_k$$

where:

O_{ijk} corresponds to the occurrence of error e_{ij} with severity degree g_k; and

g_k corresponds to the degree that expresses the error-occurrence severity, k can vary from 0 to 4.

Step 2: Calculation of Qi - Average Rate of Error-Occurrence that Affects the Characteristic Ci

Based on the registered set of errors, it is possible to identify which characteristics are affected by them. This way, the average rate of error-occurrence that affects the characteristics is calculated as:

$$QAverage_i = \sum_{j=1}^{n} q_{ij} . p_{ij} \Big/ \sum_{j=1}^{n} p_{ij} \qquad Q_i = \frac{QAverage_i}{tu}$$

Where n is equal to the quantity of error type associated with Ci.

Step 3: Calculation of X - Average Quantity of times that DBP was affected

Considering the average rates of error-occurrence that affected each quality characteristic (performed in the previous step), the calculation of the average quantity of times that DBP was affected by non-conformities related to characteristics is performed as the following:

$$X = \sum_{k=1}^{m} Q_k . W_k \Big/ \sum_{k=1}^{m} W_k$$

where:

W_i is equal of the weight of Characteristic C_i , and

m corresponds to the quantity of quality characteristics that affect the database.

As described in Section **2.1**, the statistical system of demerits uses a weight scale that is determined according to the problem to be solved [5].

The scale of the weight of quality characteristics adopted in this work is described in Table 3.

Table 3 –Weight Scale

Weight Attributed by the evaluator	Corresponding Weight used in the model	Meaning
0	0	It indicates that the presented characteristic does not have any importance
1	1	It indicates that the presented characteristic has little importance
2	3	It indicates that the presented characteristic has importance in some circumstances, but not in others
3	9	It indicates that the presented characteristic is very important
4	27	It indicates, in an absolute way, that there are no doubts that the presented characteristic is essential

This non-linear distribution of weights guarantees an index that expresses, more efficiently, the data quality degradation. Then, the weight attributed by the user to each characteristic (which can vary from 0 to 4), is modified as shown in the corresponding column in table 3.

Step 4: Association of calculated X to an index (percentile) of quality

It is impossible to know the quality of the stored data with only the calculated X value, that is, the average of times that the DBP was affected , ,. It is also necessary to calculate the percentile of corresponding quality.

In order to have a quality index of the stored data that can vary from 0 to 100%, the proposed performance graph is divided into five ranges of twenty percentile points: [0,20]; [20,40]; [40,60]; [60,80] and [80,100], as shown in Figure 3.

Step 4.1: Construction of the Quality Graph

To each $X \in [0,\infty]$ should be associate a $d = [0,100\ \%]$, so that $d(X = 0) = 100\%$ and $d(X=\infty) = 0\%$. This means that, the greater the value of X, the worse the performance will be.

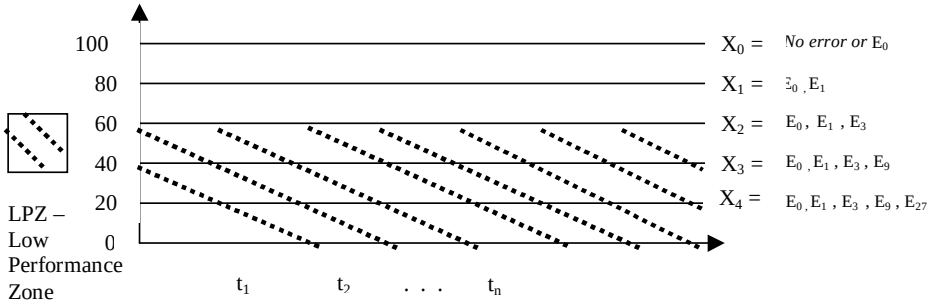

Figure 3–Performance Graph

Where t_n corresponds to the moment of the evaluation and *Ei* indicates the occurrence of *all* types of error -- *each one exactly once* -- which affect the characteristic that has weigh *i*, as discriminated by Table 4.

Table 4 – Occurrence of Errors Associated with Lines

Error	This represents the value of all types of error-occurrence, *each one exactly once*, associated with the ...
E_0	characteristics that have weight 0 (characteristic that does not have any importance)
E_1	characteristics that have weight 1 (characteristic that has little importance)
E_3	characteristics that have weight 3 (characteristic that has moderate importance)
E_9	characteristics that have weight 9 (very important characteristic)
E_{27}	characteristics that have weight 27 (essential characteristic)

The construction of the performance graph is indicated in the Table 5, where each line represents a set of values.

Table 5– Values of Each Line of the Performance Graph

Line	Correspondence to the *d* (percentile of quality)	Except for the first line, this represents that each type of error associated with the characteristics that have weights occurred one time...
X_0	100	No occurrence of errors or, each type of error associated with the characteristics that have weight occurred one time
X_1	80	0 and 1
X_2	60	0, 1 and 3
X_3	40	0, 1, 3 and 9
X_4	20	0, 1, 3, 9 and 27

Step 4.2: Calculation of d corresponding to calculated X

When the calculated ***X*** is equal to X_0, X_1, X_2, X_3 or X_4, the percentile of quality *d* corresponds exactly to one of the five main lines (100%, 80%, 60%, 40% or 20%). Then, to find the corresponding *d* for the calculated ***X***, the following described steps are necessary:

Step 4.2.1 – Calculation of the five lines

Step 4.2.1.1 - Calculation of Line X_0 that corresponds to 100 %

To calculate the line ***X_0***, which corresponds to the 100% performance percentile , it is necessary to consider all characteristics that has weight equal to 0 (zero) and to execute the next calculations:

a) Calculation of $Q_{[1..m]}$ of the $C_{[1..m]}$ considering that all types of error of the characteristics occurred only once during tu – time accumulated of utilization:

Considering *m* the quantity of characteristics with 0 weight and that ***m*** $\geq$ 1, for each C_i it calculates the Q_i (average rate of errors that affect C_i) as the following:

$$QAverage_i = \sum_{j=1}^{n} 1.p_{ij} \Big/ \sum_{j=1}^{n} p_{ij} \qquad Q_i = \frac{QAverage_i}{tu}$$

where: *n* corresponds to the quantity of error types associated with the characteristic C*i and* p_{ij} is equal to the weight of the error type e_{ij}.

b) Calculation of $X_0Partial$:

With the values of $Q_{[1..m]}$**,** the value of $X_0Partial$ is determined as:

$$X_0 Partial = \sum_{j=1}^{i} Q_j.W_j \Big/ \sum_{j=1}^{i} W_j$$

where: Q_j is the average quantity of errors that affect the characteristic C_j, W_j is the weight of the *j-th* characteristic and *i* is equal to the quantity of characteristics with weight W_J.

c) Value of X_0: $X_0 = X_0Partial$.

Next Steps - Calculation of Lines X_1, X_2, X_3 and X_4, that correspond to 80%, 60%, 40% and 20%, respectively

To calculate each one of these lines, the same approach from the previous step should be applied. Observing that: $X_n = X_{n-1} + X_nPartial$..

Note:

Considering that all types of error of the characteristics have occurred one time in one *tu* (time of use) accumulated until the moment of the evaluation, all of $QAverage_i$ (average value of errors that affect C_i) will always be equal to 1. In the same way, the value of Q_i (average rate of errors that affects C_i) will always be equal to $1/tu$. Considering the partial value of $X_{[0..4]}$ the average of all Q_i, consequently, except X_0, the partial values for $X_{[1..4]}$ will always be equal to $1/tu$. This implies that the five main lines can be calculated as:

X_0Partial = 0; (X_0Partial is a special case, therefore all the weights are equal to 0)

$X_0 = X_0Partial \therefore \boxed{X_0 = 0}$

$X_1 = X_0 + X_1Partial \therefore X_1 = 0 + \frac{1}{tu} \therefore \boxed{X_1 = \frac{1}{tu}}$

$X_2 = X_1 + X_2Partial \therefore X_2 = \frac{1}{tu} + \frac{1}{tu} \therefore \boxed{X_2 = 2 . \frac{1}{tu}}$

$$X_3 = X_2 + X_3Partial \ \therefore \ X_3 = 2.\ \frac{1}{tu} + \frac{1}{tu} \ \therefore \qquad \boxed{X_3 = 3\ .\ \frac{1}{tu}}$$

$$X_4 = X_3 + X_4Partial \ \therefore \ X_4 = 3.\ \frac{1}{tu} + \frac{1}{tu} \ \therefore \qquad \boxed{X_4 = 4\ .\ \frac{1}{tu}}$$

In the case that the calculated X has fallen in some interval, for example, between X_0 and X_1, it is necessary to interpolate this point, so that we know which is the value that actually corresponds to *d*.

Step 5: Result

With the percentile of quality *d* calculated, the quality graph of some version of the DBP will be able to indicate the moment in which the quality index fell below the tolerable level (lower than 60%), beyond the current quality, and the quality calculated in previous evaluations. With this information, a report can be emitted which lists all the occurred errors that have taken the product quality to the Low Performance Zone – LPZ (Figure 3), where t_n represents an evaluation executed in the day *dd/mm/aa*, that reached the quality index of *x*% during a time of use in *n* minutes. This allows the conduction of investigations that can lead to corrective actions.

Considering the described model in Section **4.1**, one of the parts that constitutes the evaluation process, the next Section will describe all the necessary stages to integrally accomplish the data quality evaluation trough non-conformity proposed in this work.

5 The Process

First Stage: Determination of the Data Quality Standard Expected (DQSE) by database evaluators

In this stage the database evaluators determine the importance degree of each quality characteristic listed in Section **2**. For that, they fill out the form named "Instrument for Classification of Necessary Characteristics to Evaluate Data Quality". Then, the attributed weight of each characteristic defines the data quality expectation of the users. This stage results in the data quality standard expected - *(DQSE)*.

Second Stage: Evaluation of Database Quality, Supported by a previously defined DQSE

From the *DQSE*, it is possible to measure how much the database satisfies the ideal of established quality. To obtain the index (percentile) of quality, a registry of all the occurred non-conformities is required. That registration is possible through filling out the "Form to Register the Error or Non-Conformity Occurred". At the moment of the evaluation, it is identified which quality characteristic a non-conformity has been affecting. This way, considering the accumulated time of product utilization, and the non-conformities that degrade its quality, the *Quality Index of the Stored Data* is obtained through the evaluation model that was described in Section **4.1**.

Third Stage: Results

Beside the quality index obtained in previous evaluations, it also exhibits the current index. For that, a graph divided into percentile ranges is used, like the graph introduced in Figure 4. thief the calculated index is lower than 60%, the quality will be positioned in an area considered as low performance (LPZ – Low Performance Zone). At this moment, in addition to an alert, a report, that contains the quality characteristics and the registered errors associated with them, is also generated to the users.

Thus, it is possible, according to the data quality policy of the organization, to conduct investigative actions for the probable error sources that lead to data quality degradation.

6 Validating the Method

An evaluation experiment has been accomplished through AQUA, an automatic tool prototype. The important aspects of AQUA and that experiment are succinctly discussed below.

6.1 The AQUA Prototype

This tool executes a data quality evaluation by non-conformity. To do this, it collects the errors or non-conformities occurred during DBP utilization, in order to verify how much each error degraded data quality. The tool incorporates several modules that aim to register occurred errors during database utilization, new data quality characteristics and its respective types of error; to make the evaluator's task easy, in attributing weights to quality characteristics; and to accomplish data quality evaluation. Figure 4 shows two examples of these functionalities and the evaluation results through a performance graph.

6.2 An Experiment in Data Quality Evaluation

The CAM - Control of Maritime Area System, was chosen for the data quality evaluation experiment. This is a decision-support system developed by CASNAV - Center of Naval System Analysis, a military organization of the Brazilian Navy. Each one of the stages is described as follow.

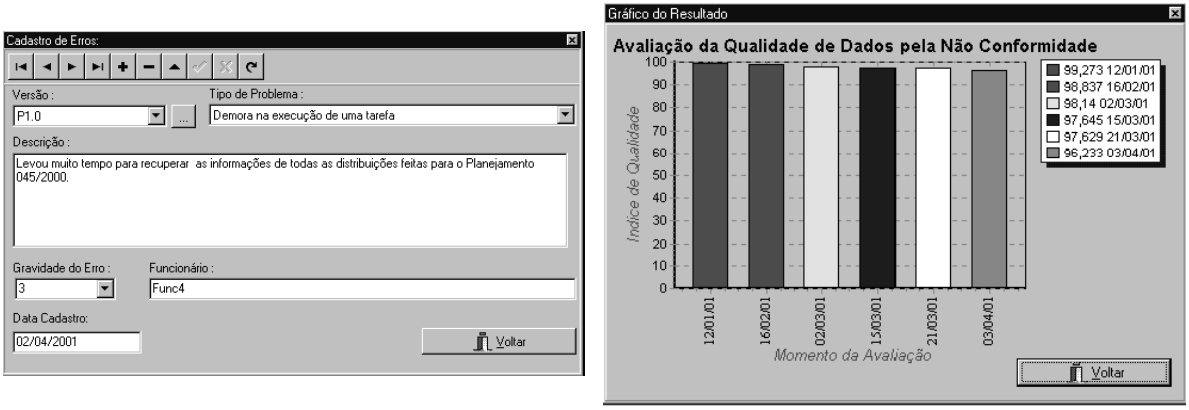

Figure 4 – Examples of AQUA's user interface

Stage 1: Registration of errors or non-conformities

Specialists of the Navy, users of CAM, registered all occurred errors during an utilization period of 4,800 minute (approximately 10 days). Some of these errors are in Table 6. This aims to obtain the set of non-conformities that affect the quality of the stored data in version P1.0 of CAM.

The first and the second column correspond to the types of occurred problems and its succinct description, respectively. The third column informs the relative weight of severity (S) of each occurred error. The fourth and fifth column identify the employee (E), with the name suppressed due to privacy reasons, which registered the error, and the occurrence date (CD).

Stage 2: Determination of Data Quality Standard Expectation (DQSE)

A specialist analyzed and attributed a weight to the set of quality characteristics (Table 7).

Stage 3: The Database Quality Evaluation

Based on the previous stages, a quality index of stored data was generated .

Stage 4: Exhibition of the result

The quality index created in stage 3 shows if occurred non-conformities in the CAM System affected the quality characteristics, consequently resulting in general data quality degradation to the system. The calculated index quality was 96,23%.

Thus, when necessary, it is possible to adopt investigative and/or corrective actions to locate probable sources of error that degraded the quality of the stored data.

The performance graph, illustrated in Figure 5, shows the result of a more recent evaluation, as well as the result of other evaluations. In this way, it is possible to follow the data quality evolution along time.

Table 6 – Occurred Non-Conformities in CAM System

Type of problem	Description	S	E	CD
Error insertion in database after abnormal situation of system operation	A problem in the electric power supply caused an interruption of a fixed area distribution and data of friendly means were lost	3	Emp1	06/01/2001
Lack of alert to indicate incorrect or non-conform data entry	It allows registration of a negative speed reaction of friendly means	3	Emp1	07/01/2001
Lack of necessary field in forms	It was not possible to register the commander's name of friendly means	1	Emp2	09/01/2001
Data stored without update for long time	The width of sweeping of the friendly means is not up-to-date as the plan in validity, thus a distribution was inadequate because of this	4	Emp3	12/01/2001
Two or more different values stored in database	The Dead Time field, which is exclusive to aircrafts, is equivalent to the Preparation field existing for all friendly means	2	Emp1	09/03/2001
Execution task delay	Execution time delay to recover the information of what all the distributions have done to the Planning 045/2000	3	Emp4	02/04/2001

Table 7 – Weight of Data Quality Characteristics of CAM

Characteristics	Weight	Characteristics	Weight
Availability of information	4	Precision of data	4
Age of data	3	Consistency	2
Opportunity	2	Easy of signaling	3
Efficiency of execution	2	Accountability	0
Relevance	2	Retrievability	3
Utility	2	Flexibility	1
Lucrativeness	3	Interoperability	2
Aid at user work	3	Access security	4
Competitiveness	4	Understandability	0
Appropriate amount of data	0	Adequacy of information	2
Accuracy	3	Uniformity (no subfactors)	3
Completeness	2	Availability of documentation	2
Coverage (depth)	3	Traceability	1
Robustness	2		

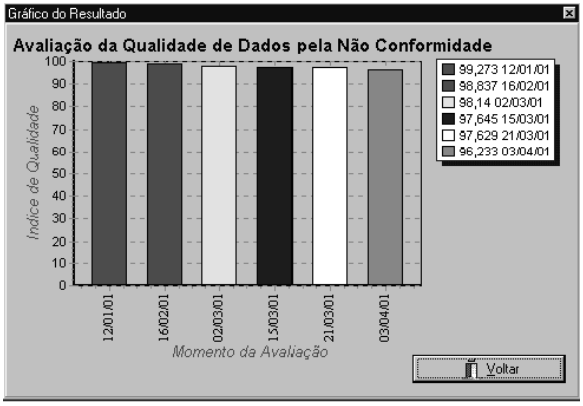

Figure 5– The CAM System performance graph

7 Conclusion

This work discusses how data quality supports the decision-making process of an organization to obtain competitive advantages. We also discuss the role of quality characteristics as a base reference element to the product quality.

The quality characteristics applied in this work were collected and defined from literature, and to reach their necessary adequacy, a field research was conducted consulting specialists from academic, military and enterprise institutions.

Thus, it supports the developed mathematical model. This model was used during the data quality evaluation experiment with the CAM System, through the use of the AQUA prototype.

Besides, the systematization of a data quality evaluation by non-conformity, and the definition of the necessary procedure to obtain a data quality index, should be emphasized as further contributions.

Finally, as future perspectives, we suggest applying fuzzy theory, taking into consideration its ability to capture human's imprecision knowledge, to identify user expectation about data quality,.

Regarding the AQUA prototype, we suggest the automation of error or non-conformity registration, automatic capture of *tu*, and association with a version control system.

References

1. Baldwin, A. A., Bowen, P. L., 1999, "Data Quality: A Review and Case Study", *Journal of Information Systems*.
2. Willshire, J. M., Meyen, D., 1997, "A Process for Improving Data Quality", *Data Quality Journal*, v. 3, n. 1 (Set).
3. ISO 9000 – Quality management systems – Fundamentals and vocabulary, ISO/IEC.
4. Rocha, A. R. C. da, 1983, *Um Modelo para Avaliação da Qualidade de Especificações*. Tese de Doutorado, PUC-RJ, Rio de Janeiro, Brasil.
5. Strong, D., Wang, R.Y., Guarascio, M . L., 1994, "Beyond Accuracy: What Data Quality Means to Data Consumers". *TDQM Research Program*, USA, October
6. Adelman, S., Marco, D., Moss, L., 2001, "When we are cleaning data from a relational source, and some values", *DM Review Online*, March 2001.
7. Wand, Y., Wang, R.Y., 1996, "Anchoring Data Quality Dimensions in Ontological Foundations", *Communications of the ACM*, v. 39, n. 11 (Nov), pp. 86-95.
8. Strong, D. M., Miller, S. M., 1995, "Exceptions and Exception Handling", pp. 206-233.
9. Wang, R.Y., Reddy, M P., Kon, H. B., 1995, "Toward quality data: An attribute-based approach", *Decision Support Systems*, v. 13, n. 3 (Mar), pp. 349-372.
10. Huh, Y. U., Keller, F. R., Redman, T. C., Watkins, A. R., 1990, "Data Quality", *Information & Software Technology*, v. 32, n. 8 (Oct), pp. 559-565.
11. Wang, R. Y., Kon, H. B., Mandick, S. E., 1993, "Data Quality Requirements Analysis and Modeling". In: *The Proceeding of the 9th International Conference on Data Engineering*, pp. 670-677, Vienna.
12. Montgomery, D. C., 1996, Introduction to Statistical Quality Control, 3 ed. USA.
13. ISO 9126, 1991, Information technology - Software product evaluation- Quality characteristics and guidelines for their use, ISO/IEC.
14. Pinho, S.F.C, 2001, Avaliação da Qualidade de Dados pela Não Conformidade. Tese de Mestrado, COPPE/UFRJ, Rio de Janeiro, Brasil.
15. Belasco, J. A., Stayer, R. C., 1994, O vôo do Búfalo: Decolando para a Excelência, aprendendo a deixar os Empregados Assumirem a Direção, Rio de Janeiro, Editora Campus, in [BELCHIOR, 1997].
16. Kume, H., 1985, *Statistical Methods for Quality Improvement*, Japan.