International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:11, 2011

# A New Precautionary Method for Measurement and Improvement the Data Quality

Seyed Mohammad Hossein Moossavizadeh, Mehran Mohsenzadeh, Nasrin Arshadi

***Abstract*—**the data quality is a kind of complex and unstructured concept, which is concerned by information systems managers. The reason of this attention is the high amount of Expenses for maintenance and cleaning of the inefficient data. Such a data more than its expenses of lack of quality, cause wrong statistics, analysis and decisions in organizations. Therefor the managers intend to improve the quality of their information systems' data. One of the basic subjects of quality improvement is the evaluation of the amount of it. In this paper, we present a precautionary method, which with its application the data of information systems would have a better quality. Our method would cover different dimensions of data quality; therefor it has necessary integrity. The presented method has tested on three dimensions of accuracy, value-added and believability and the results confirm the improvement and integrity of this method.

***Keywords*—**data quality; precaution; information system; measurement; improvement

## I. Introduction

NOWADAYS tow concepts of "data" and "information" are using too much. We can define data, as "to being presentation phenomenon, concepts or known things formally and Appropriated for relationship, interpreter or processing by human or any auto-equipment"[2]. In addition, we can say, which data that is "fitness for use"[4, 5, 6], has a good quality grade. Most of the researcher who worked on this subject, are usually concentrating on a special data or they have used a similar definition like above definition [7].

As far as the decision accuracy depends on the accuracy (in its generic mean) of data, the topic of the data quality was proposed. This topic has its own dimensions. Every kind of incoherence and disharmony in data cause many problems in organization. The data quality is an important topic; especially in subjects "Data Warehouses", "Business Intelligence" and so on.

The lack of data quality also can cause high amount of financial burden, so always and before any phenomenon would appear and for the decrease of its consequences, the data quality situation must be under the evaluation and observation. Monitoring and improvement of the data quality is very costly [10]. The current subjects of the data quality cost millions of dollars per year. In between of 30% to 80% of data analysis is spending to data cleaning and understanding [1].In this paper, by presenting a new precautionary method, we try to improve the data quality grade. The usage of a set of precautionary principles in information systems can decrease the data quality evaluation and improvement costs explicitly. Therefore, our method has been tested on three different and challenging dimensions accuracy, value-added and believability in this paper. Results of test confirm claim exactness.

S. M. H. Moossavizadeh with Department of Computer, Science and Research Branch, Islamic Azad University, Khouzestan, Iran (email: mh.moossavizadeh@khouzestan.srbiau.ac.ir).

M. Mohsen Zadeh is with Department of Computer, Science and Research Branch, Islamic Azad University, Khouzestan, Iran (email: mohsenzadeh@srbiau.ac.ir).

N. Arshadi is with Department of Psychology Shahid Chamran University of Ahwaz, Ahwaz, Iran (e-mail: nasrinarshadi@yahoo.com).

## II. Literature Review

In the whole researches on the data quality, one of three subjects "Evaluation", "improvement", or "evaluation/ improvement" has been concerned. However, none of them point to the precaution issue [4, 8, 9]. For example TDQM, DWQ, COLDQ, CDQ, etc. Apart from which one in its goals is more successful, it must be said that none of them concern the precaution issue in the data quality.

## III. The Basic Principles Of Precautionary Data Quality Evaluation And Improvement

For the suitable data quality attainment, two kinds of cost (the data quality evaluation and improvement) should be observed. Each of discovered appearance of lack of the data quality have some reasons, which if not exist the status of data quality would be better. The main idea of this paper is the presenting a precautionary principles set, which can remove the reasons of appearance of lack of the data quality. These principles are described below. It must be taken that based on the conditions of each system, the below principles should be customized.

### *A. The primary data production principle*

One of the important reasons of incoherence between the information system and the real world is the tiny unwanted changes of data in the data extraction process. In addition, the data may be achieved from some individual sources, which could give different values of the data. For prevention from these problems, the data should be taken from the real world. For sure, it cannot be done in some places which is needed to have some arbitrator, such as data warehouses.

Finally, by using this principle, whatever must be existed, is existed and whatever is existed, is the ones that must be existed.

### *B. The trade off principle*

On one side, it must be taken into account that the amount and the expansion of components of the data potentially can be extreme and on the other side, the organization for keeping the whole data component would encounter with some limitation. Therefore, it is needed to tradeoff between the real expansion of data components and its needed expansion.

### *C. The data lifecycle extension principle*

The reason of useful data lifecycle shortage is the absence

of accurate methods for its usage. If the real role of a data in its lifecycle has not been recognized, it must be taken into account, the useful data lifecycle would be incomplete. The final place that the data could be applied is to establish new organizational standards, while these standards could be open standards and with continuous improvement. Therefore, the data usage would go on forever. All of the data can be used in the short term, middle term, long term and even strategic planning. Therefore, not only the data lifecycle never terminates, but also the important and even unimportant data in all kinds of planning can be applied. Besides spending on these data not only is possible, but also is essential.

*D. The periodic data reviews principle*

One of the necessary steps in the data lifecycle is a periodic data reviews, because the data of information system after a period may be out of date. Based on the application of data in macro planning, periodic data reviews are vital and essential.

*E. The pervasive Meta data edition principle*

We must be careful in organizational definition of a data. Firstly, because the data in organization could be known more realistic and the different parts of organization could identify the nature of the data in a single unit. According to the native interpretation of organization, the Meta data usually extracted from the data itself. In the other words, the extracted Meta data should be customized for each information system. The lack of pervasive Mata data would result in the drop of the data quality, because in the Meta data some metric of data quality is defined.

## IV. Good Point Determination For Precautionary Principles

A good point is a degree, which shows that the application of principle is sufficient. In other words, when we can by usage of a principle achieve the goals relatively or absolutely, it can be said we reach to the good point of the principle. Besides, these good points should be customized and interpreted in each special usage.

*A. The good point of the introductory data production principle*

If the values of the data directly extracted from the real data itself, and possibly could not pass from the intermediate equipment, we approached to the good point of the introductory data production principle.

*B. The good point of the tradeoff principle*

If informational components of real data exists it information system as needed, we received to the good point of the tradeoff principle.

*C. The good point of the data lifecycle extension principle*

If where it is possible, the data involved in the intra-enterprise and extra- enterprise macro decision-making, and the data lifecycle would be infinite, the principle is located in its good point. For sure, this good point is relative. It is possible that the achievement to this good point for all at the data could not be accessible.

*D. The good point of the periodic data reviews principle*

We could approach to the good point of this principle, when in the time of the data usage, we could be assure about accuracy and updated of data.

*E. The good point of the pervasive meta data edition principle*

If information system could interpret and use the data based on the realistic world without the human presence and purely with having data and Meta data, we could approach to the good point of this principle.

## V. The Presentation Of Algorithmic Methods For Access To Precautionary Principles Good Points

Alg.1. An access method for the introductory data production's good point: It is shown in figure 1.

Alg.2. An access method for the tradeoff's good point: It is shown in figure 2.

Alg.3. An access method for the data lifecycle extension's good point: The influence of data in organizational macro decision-making must be added to its lifecycle. This access method is shown in flowchart 3.

Alg.4. An access method for the periodic data reviews' good point: It is shown in flowchart 4.

Alg.5. An access method for the pervasive Meta data edition's good point: It is shown in flowchart 5.

## VI. Proof Of The Validity Of Proposed Solution

In this section, through case study, it is shown that the proposed solution will be useful for different dimensions of data quality. For this useful purpose, the proposed solution will be examined concerning three different data quality dimensions. The selected dimensions are, accuracy (a known dimension which is consensus, and the most effective quality dimension), value-added (an emerging dimension, with new results of research), and believability (a controversial and ambiguous dimension).

*A. Effects of the proposed solution in accuracy dimension*

*1) The primary data production principle*

*a) Preventing the inadvertent errors in data entry*

Respecting this principle will help vanishing the accidental errors in data transfer intermediates.

*b) Preventing the intentional and conscious errors in data entry*

The meaning of conscious mistakes is the changes that will inevitably occur with the awareness of the organization. Therefore complying with this principle will result the retrieved data in the information system to be a raw data. Of course, it is obvious that the necessary changes in the data are not included in this issue.

*2) The tradeoff principle*

*a) Preventing the intentional and conscious errors in data entry*

Organizations, sometime and for some reasons, would put off the keeping of all data components. This issue is an intentional and conscious one. The compromise principle, in addition to organizational restrictions, will monitor the essential components and services for analysis and accurate decisions; and will establish a balance between "what could be" (organizational restrictions) and "what should be" (data requirements).

*3) The data lifecycle extension principle*

*a) Preventing nonoccurrence between information system data and real data*

Each data lifecycle has a certain mechanism to provide the accuracy of data. With a move toward unlimited data lifecycle, the mechanism of providing data accuracy will also develop and become unlimited. Therefore using this principle will consolidate the mechanism of providing data accuracy.

*4) The periodic data reviews principle*

*a) Preventing inaccuracy of data when using them*

The main benefit of periodic reviews principle is, to prevent inaccuracy of data when using them.

*5) The pervasive metadata development principle*

*a) Preventing inaccuracy of data when entering them to the information system*

The data accuracy measurements are mainly exists in their metadata. When data enters the information system, if these measurements were mentioned more completely in the metadata, there would be more accurate control over data entry and we could obtain more confidence in the entered data accuracy.

### *B. Effects of the proposed solution in value-added dimension*

*1) The primary data production principle*

*a) Preventing the number of components in the information system to be less than the real world*

If a data in real world has a component, it "should" have it in information system as well. The lower the number of data components in the IS, is compared to its number in real world, completeness parameter will debilitate in the value-added dimension. Observance of data primary production principle will cause data and it is components to be saved in the information system just like how they are in real world.

*b) Preventing the number of "components without value" of a data in an IS, become higher than its "whole components" number*

The existence of intermediates and specially ones that are heterogeneous will higher the potential of empty data components. This problem is effective in value-added and completeness dimensions. Using the data primary production principle will remove the intermediates as long as it is possible.

*c) Preventing the data lifetime cycle to be short*

The shorter lifetime cycle of a data item, the lower it's value-added will be. This principle, will cause the data lifecycle to go into unlimited direction and show that almost all data are useable for long times and will be critical in decision makings.

*2) The tradeoff principle*

*a) Preventing incompatibility with metadata*

For information system data to match more with real data, the number and the kind of real data component should match with the assigned metadata and therefore the information system would be more believable. The trade off principle tries that if possible, the number and the kind of data component of an information system match with the real data.

*3) The data lifecycle extension principle*

*a) Preventing the unavailable values*

In the long run, with multiple referrals to real data, the probable unavailable value of information system data will be detected and will match with real data. This advantage of data lifecycle development principle, improves the data believability.

*b) Helping data accuracy*

Considering the effect of this principle on data accuracy, it is also effective in believability dimension.

*4) The periodic data reviews principle*

*a) Preventing the lack of components or the data values*

The quantity of components is effective in completeness and value-added dimensions. In the long run, the probability of data components becoming high or low is not dropped. To find out about this issue, the periodic reviews principle should be used. Furthermore, if for whatever reason, the value of one of the components is missing, the process of periodic reviews will also obtain the missing component.

*5) The pervasive metadata development principle*

*a) Preventing the lack of proper attention to data based on its nature*

One of the important instances of metadata is information relating to its nature. Data's nature shows the level of its importance for organization and it will determine how the organization should treat the data in analysis, statistics and security. The Comprehensive metadata development principle will notably help completing data's nature.

### *C. Effects of the proposed solution in believability dimension*

*1) The primary data production principle*

*a) Preventing data loss*

The data primary production principle for valid or invalid reasons will cause no component of data to lack value, because in real world if we assume one of the data components to have no value, it means that we have rejected its origin.

*b) Preventing lack of compatibility with intellectual ability and organizational rules*

Data should be in a defined and acceptable range (of intellectual and organizational rules) to be more believable. The practical solution to this goal is, that more efforts put in extracting data components values immediate from real data. This solution is the same as data primary production principle.

*c) Preventing inaccuracy*

The data primary production principle will result in increased probability of accuracy and succeeding that, the believability of extracted data components.

*2) The tradeoff principle*

*a) Preventing incompatibility with metadata*

For information system data to match more with real data the number and the kind of real data component should match with the assigned metadata and therefore the information system would be more believable. The compromise principle tries that if possible, the number and the kind of data component of an information system match with the real data.

*3) The data lifecycle extension principle*

*a) Preventing the unavailable values*

In the long run, with multiple referrals to real data, the probable unavailable value of information system data will be detected and will match with real data. This advantage of data lifecycle development principle, improves the data believability.

*b) Helping data accuracy:*

Considering the effect of this principle on data accuracy, it is also effective in believability dimension.

*4) Periodic reviews principle*

*a) Preventing data inaccuracy with new experiences*

To keep the data believability, it is essential that periodically we match it with new experiences.

*b) Preventing data incompatibility with specified metadata*

With change in real data, the information system data won't match the specified metadata. Execution of periodic reviews principle will rematches data with its metadata.

*5) The pervasive metadata development principle*

*a) Preventing apparent inaccuracy between data and metadata*

Errors and defects of metadata, will render the data item "unbelievable". This issue clearly expresses the essentiality of comprehensive development and improvement of metadata.

## VII. Conclusion And Summary

There have been several methods for measuring and improving the data quality and none of them has specifically and clearly discussed the precautionary topic. This problem will force organizations to pay for data quality improvement. In this paper, with a comprehensive look, we tried to present a collective of precautionary principles that prevent disruptive factors to data quality. To verify and evaluate the performance of this method, a case study has been performed on the three dimensions of data quality (accuracy, value-added, believability). These dimensions were chosen in a way that the proposed method would be examined with heterogeneous dimensions.

The proposed method here can comply with any of measuring and improving data quality methodologies.

## Acknowledgment

This research is indebted to our merciful God and Imam Mahdi (pbuh). Special thanks for my Devoted and Beloved wife 'dear Zahra' and my lovely daughter 'dear Mohaddeseh', because of their support of my researches. Additionally thanks for my dear parents.

## References

[1] J. Patrtabian, The data quality measuring using by data mining (MSc Thesis), Departmen t of computer, Science and Research branch, Islamic Azad University, Khouzestan, Iran, Winter 2009.
[2] S. M. T. Rohani Rankouhi, Foundamental concepts of DataBase, Jelveh Publications, First Edition, 2002.
[3] H. Seraj, A new framework for measuring of the data value-added, (MSc Thesis), Departmen t of computer, Science and Research branch, Islamic Azad University, Khouzestan, Iran, Summer 2010.
[4] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41, 3, Article 16 (July 2009), 52 pages.
[5] Wang, R. 1998. A product perspective on Total Data Quality Management. Comm. ACM 41, 2.
[6] Juran, J. M.; Gryna, Jr, F. M., "Quality Planning and Analysis", 2nd ed., McGraw-Hill, New York, 1980.
[7] Luebbers, Dominik; Grimmer, Udo; Jarke, Matthias. "Systematic Development of Data Mining-Based Data Quality Tools". Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003.
[8] Batini, C. and Scannapieco, M. 2006. Data Quality: Concepts, Methodologies and Techniques. Springer Verlag.
[9] Pipino, L., Lee, Y. and Wang, R. 2002. Data Quality Assessment. Commun. ACM 45, 4.
[10] Wang, R., Ziad, Mostapha; W. Lee, Yang. "Data Quality". Springer: Kluwer Academic Publishers, 2002.

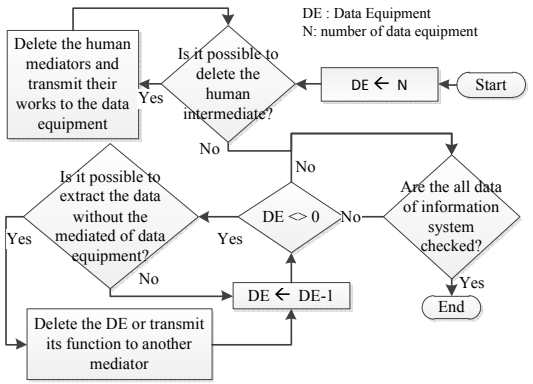


Fig. 1 A flowchart for the procedure of reaching the desired point in the primary data production principle

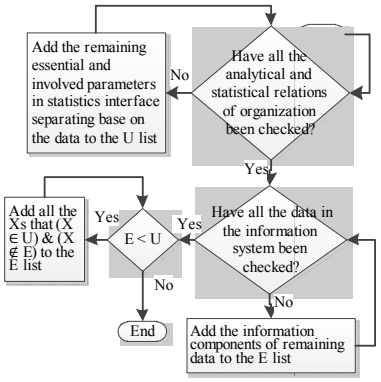


Fig. 2 A flowchart for the procedure of reaching the desired point in the trade off principle

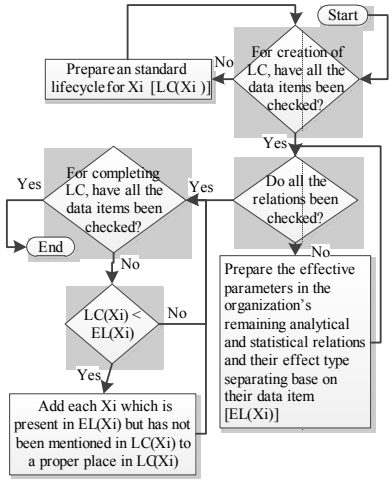


Fig. 3 A Flowchart for the procedure of reaching the desired point in the data lifecycle extension principle

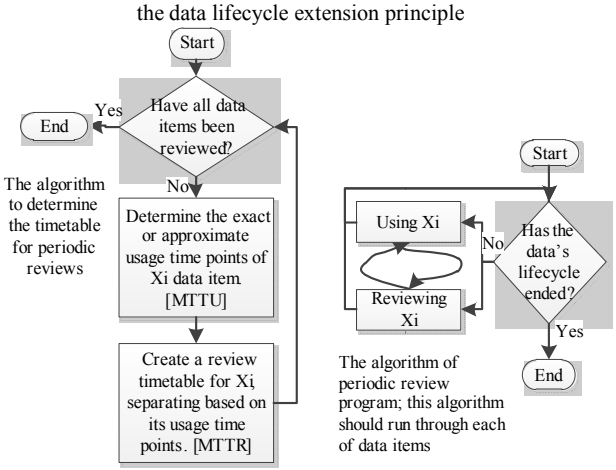


Fig. 4 Two flowchart for the procedure of reaching the desired point in the periodic data reviews principle


Fig. 5 The best order for use/review scheduling

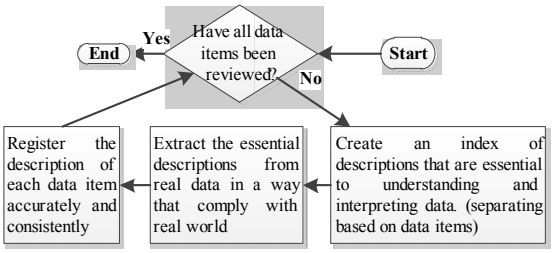


Fig. 6 A flowchart for the procedure of reaching the desired point in the comprehensive metadata development principle