World Multiconference on Systemics, Cybernetics and Informatics

July 22-25, 2001
Orlando, Florida, USA

PROCEEDINGS

Volume II

Information Systems

Organized by IIIS
International
Institute of
Informatics
and Systemics

Member of the International
Federation of Systems Research

IFSR

Co-organized by IEEE Computer Society
(Chapter: Venezuela)

EDITORS
Nagib Callaos
Yunfa Hu
Manuel Rodriguez
Quang Ha

ISBN 980-07-7542-0

PROGRAM COMMITTEE

Chairman: Prof. William Lesso (USA)

Abe, Akinori (Japan)
Abe, Jair Minoro (Brazil)
Abi-Raad, Maurice (Australia)
Acuña Castillo, Silvia Teresita (Argentina)
Ali, Zulfiqur (United-Kingdom)
Allegra, Mario (Italy)
AlMohammad Bader, AlBdaiwi (Kuwait)
Ammenwerth, Elske (Germany)
Amos, David (France)
Andrieux, Laurent (France)
Anzaloni, Alessandro (Brazil)
Audestad, Jan Arild (Norway)
Aveledo, Marianella (Venezuela)
Badawy, Wael (Canada)
Baïna, Jamal (France)
Banathy, Bela A. (USA)
Banathy, Bela H. (USA)
Beheshti, Mohsen (USA)
Bemley, Jesse L. (USA)
Bica, Marin (Romania)
Boiko, Igor (Canada)
Bozinovski, Stevo (Macedonia)
Brankovic, Ljiljana (Australia)
Bruzzone, Agostino G. (USA)
Burge, Jamika (USA)
Buzuloiu, Vasile (Romania)
Cano Escriba, Juan Carlos (Spain)
Carretero, Jesús (Spain)
Castillo, Oscar (Mexico)
Cernuzzi, L (Paraguay)
Champarnaud, Jean-Marc (France)
Chandra, Ramesh (India)
Chen, Chung-Ming (Taiwan)
Chen, Pei-Min (Taiwan)
Cheng-Li, Liu (Taiwan)
Chiu, Yi-Jen (USA)
Cohen, Eli (USA)
Curcio, Igor (Finland)
Da Silva, Ivan Nunes (Brazil)
David, Amos (France)
Duale, Ali (USA)
Dubois, Daniel (Belgium)
Dujet, Christiane (France)
Duong, Tuan (USA)
Eastabrook, Glenn (Canada)
Elizondo, César (Mexico)
Esquivel, Susana (Argentina)
Farooqui, Aamir (USA)
Fernandez, Jorge (USA)
Ferri, Fernando (Italy)
Fong, Anthony (Hong-Kong)
Fulantelli, Giovanni (Italy)
Gammoudi, Mohamed Mohsen (Tunisia)
Ganz, Aura (USA)
Garcia-Zambrana, Antonio (Spain)
Gendron, Michael (USA)
Goldman, C. (Israel)
Gou, Bei (USA)
Gou, Hongmei (China)
Granja-Alvarez, Juan Carlos (Spain)
Grout, Ian (Ireland)
Grzymala-Busse, Jerzy (USA)
Gunay, Kibarer (Turkey)
Guo, Qinglian (Japan)
Gutierrez, Rafael (USA)
Guu, Sy-Ming (Taiwan)
Ha, Quang (Australia)
Hamid, Issam A. (Japan)
Hanratty, Timothy (USA)
Hantos, Peter (USA)
Hernandez-Encinas, Luis (Spain)
Hideo, Fujimoto (Japan)
HongSheng, Li (China)
Horvath, L. (Hungary)
Hu, Yunfa (China)
Iliopoulos, Costas (United-Kingdom)
Izworski, Andrzej (Poland)
Jacob, Lillykutty (Singapore)
Jacomet, M. (Switzerland)
Jha, Manoj (USA)
Jordan, Andrzej (Poland)
Jozwiak, Lech (The Netherlands)
Juric, Matjaz B. (Slovenia)
Kamijo, Shunsuke (Japan)
Karlin, Jay (USA)
Karras, Dimitris A. (Greece)
Kasabov, N. (New Zealand)
Kaupp, Gerd (Germany)

Kawai, Shigeru (Japan)
Kim, Daeyoung (USA)
Klovsky, Daniel D. (Russian)
Koide, Seiji (Japan)
Kolodnytsky, Mykola (Ukraine)
Kraiem, Naoufel (Tunisia)
Krajcar, Slavko (Croatia)
Kvasnica, Milan (Czech-Republic)
Lastra, Jose (Finland)
Lau, Vincent (Hong-Kong)
Lebby, Gary (USA)
Lee, Hsuan-Shih (Taiwan)
Lee, Jong kun (Korea-(Republic-of))
Lefevre, T (Thailand)
Liang, Jie (USA)
Long, Steven (USA)
Loutfi, Mohamed (United-Kingdom)
Lu, Ning H. (USA)
Luo, Zhi-Wei (Japan)
Maciel, Paulo (Brazil)
Martin, Curtis (USA)
Masaki, Murakami (Japan)
Meijer, Bart R. (The Netherlands)
Motus, Leo (Estonia)
Mounir Alaoui, Salim (USA)
Mow, Wai Ho (Hong-Kong)
Muravyov, Sergey (Russian)
Najarian, Kayvan (USA)
Nakashima, Hideyuki (Japan)
Naranjo, Michel (France)
Naviner, Lirida (France)
Neagoe, Victor-Emil (Romania)
Neelov, Igor (Finland)
Ng, H.S. Raymond (Hong-Kong)
Nishio, Satoru (Japan)
Ohsawa, Yukio (Japan)
Ohta, Toshizumi (Japan)
Ohyama, Yasuhiro (Japan)
Oka, Ryuichi (Japan)
Okun, Oleg (Finland)
Palade, Vasile (Romania)
Patel, Ahmed (Ireland)
Patterson, Jr. F. G. (USA)
Pelaez, Javier R. (Brazil)
Perales, Francisco José (Spain)
Petrounias, Ilias (United-Kingdom)
Pham, Tuan (Canada)
Piattini, Velthuis Mario (Spain)
Pierre, Samuel (Canada)
Postigo, José (Aegentina)
Potaturkin, Oleg (Russian)
Power, Gregory (USA)
Power, James (Ireland)
Prokop, Roman (Czech-Republic)
Protheroe, Dave (United-Kingdom)
Ramesh, B. Inampudi (India)
Refice, Mario (Italy)
Remmel, Jeff (USA)
Rishe, Naphtali (USA)
Rodríguez, Manuel (Spain)
Rosario, Dalton (USA)
Sadka, Abdul H. (United-Kingdom)
Saglietti, Francesca (Germany)
Sancho, Gómez José Luis (Spain)
Saseetharran, M. (Sasheei) (USA)
Shinonaga, Hideyuki (Japan)
Sommerer, Christa (Japan)
Succi, Giancarlo (Canada)
Surdu, John (USA)
Suzuki, Naoki (Japan)
Szabo, Raisa (USA)
Takahashi, S. (Colombia)
Tamulis, Arvydas (Lithuania)
Teixeira, Antonio Luis Jesus (Portugal)
Testorf, Markus (USA)
Tinetti, Fernando G. (Argentina)
Torres, Michel (Venezuela)
Trofimov, Vyacheslav (Russian-Federation)
Tsai, Cheng-Fa (Taiwan)
Tseng, Shian-Shyong (Taiwan)
Uchôa, Elvira (Brazil)
Vacher, Jean-Philippe (France)
Vafaie, Haleh (USA)
Van der Schaar, Mihaela (The Netherlands)
Vanhala, Jukka (Finland)
Vergez, Christophe (France)
Verna, Didier (France)
Watanabe, Tomio (Japan)
Welzl, Michael (Austria)
Whymark, Greg (Australia)
Wieczorkowska, Alicja (Poland)
Wolfram, Klein (Germany)
Wong, Kin Yeung (Hong-Kong)
Wu, Xuezhong (China)
Xenos, Michalis (Greece)
Xia, F. (Macau)
Xiong, Yingen (China)
Yamaguchi, Yoko (Japan)
Yasser, E. (Egypt)
Yigang, He (China)
Yu, Guoyang (China)
Zboril, F. (Czech Republic)
Zeeuw, G. (Holland)
Zhang, Qingying (China)
Zhang, Yiying (Canada)
Zimmermann, Kerstin (Austria)
Zong, Xuli (USA)
Zubairi, Junaid (USA)

CALIDAT: a methodology for measuring data quality.

Mario Piattini Velthuis
Grupo ALARCOS. Departamento de Informática.
Universidad de Castilla-La Mancha.
Ronda de Calatrava 7, 13071, Ciudad Real, ESPAÑA.

and

Ismael Caballero Muñoz-Reja
Grupo ALARCOS. Departamento de Informática.
Universidad de Castilla-La Mancha.
Ronda de Calatrava 7, 13071, Ciudad Real, ESPAÑA.

ABSTRACT

Organizations databases contain data which maybe useless or undesirable, making the working processes with this data become ineffective and unproductive. It needs be assured data in databases present an acceptable quality level. Below the CALIDAT methodology is presented, which is good to determine if data have the required quality level. It is based on defining dimensions to which quality is demanded to data, in taking measures about these dimensions and in analysing the results. If data have quality enough it can work with them. Otherwise they will make opportune corrective actions and measurements will be repeated until be ok.

Keywords: Data Quality, Measuring Data Quality, Parameters, Indicator, CALIDAT Methodology.

1. INTRODUCTION

It is necessary existent data in organizational databases present certain characteristics that allow the best performance in their exploitation. This circumstance is causing changes in the way of making business: it has forced to the companies to think about to manage the information like an another more assets ([5]). It also can suppose the base for knowledge creation allowing planning tactical and strategies of market that bear benefits for the organization. That is way, it is possible to affirm data have become into the raw material of the companies so when developing their activity as when making their relationships with their clients or with their suppliers. Organizations need collect, store and process a big amount of information coming from multiple sources and in different moments. But after collecting these data may happen that in their databases, also useful data, there are these other types of data:

- **Unnecessary Data:** They are those do not belong to the domain of expectations of information workers or users.
- **Redundant Data**: They are those are recollected and stored several times in the databases.
- **Expired Data:** They are those having been used, they are going to be used no more .

To eliminate this undesirable types of data, it needs detecting and locating it making a previous study on the viability and the impact so eliminating it as maintaining it in the database. This work need a judgement on data to identify this causing problems. Once identified, decisions should be taken to eradicate problems due to have it. In this article, a methodology is proposed to determine the global state of data quality through a valuation of certain data dimensions according to some certain evaluation criteria.

2. CONSIDERATIONS ABOUT DATA QUALITY.

Data quality can be found at three different levels: the quality of DBMS, the quality of database schema, and the quality of the proper data. Three line to improve the intrinsic quality of the databases are given in [9]:

1. Build richer semantic models that reflect the reality better. These models would should:
    a. Not allow data without having the appropriate semantics.
    b. Require the acquisition of data that it has not been introduced to guarantee the completion of the database.
    c. Quantify and store data quality into the data model.
2. Reinforce databases with a bigger number of restrictions to identify and discriminate against

data with problems and to connect this data with the appropriate applications.

3. Restrict the use of data to predefined processes, not allowing they are modified by any process so that they cannot be erased accidentally.

The quality that is sought to measure is that relative to the own nature of the data. The proposed methodology in this article is based on 1.C: store information about the quality of data in the same database and evaluate this information to determine the state of this quality.

Having arrived to this point, two problems have to become solved:

- **what information about quality should be stored**: quality is a multidimensional concept ([3], [5], [12], [16], ...) Users examine certain characteristics of products to sum up the degree with which their necessities can be or no satisfied. The information that should be stored is the group of those characteristics that allow concreting the affordable satisfaction degree and that they come reflected in the user specifications. In the literature, these characteristics have been called **quality parameters** ([16]) and they are obtained after an analysis of user requirements. But each one of those characteristics must have associated an objective value that makes it significant in a context. To each one of these characteristics has been called **quality indicators** ([16]). Because they are precisely these quality indicators the one which should be stored in the database, they will contribute the necessary information in a certain dimension on of datum.
- **how to store the information corresponding to the indicators** or measures of quality, for a later analysis of this information. Literature offers different solutions for conceptual model: modelling data quality as entity with their corresponding attributes ([13], [14], [16], [17]) , or as attributes of other entities or of other attributes ([16]). It also could store data quality in attributes of the attributes. For example, in figure 1 a vote intention table is represented and the recollected votes by some politicians in an elections. Under each percentage (that is the datum), there are some attributes of the attribute value representing date in which datum was produced and source giving it took place.

Name	Vote Intention	Vote Pick Up
Smith	80% <30/10/99, NYT >	49% <5/11/99, Electoral Office>
Sanders	20% <30/10/99, NYT>	51% < 5/11/99, Electoral Office>

Figure 1: Data and data about data quality.

Values corresponding to proper datum and to quality attributes would be stored in the same field. In the case of a relational DBMS, this situation would or should not be occurred since First Normal Form (1FN) would not be verified in a normalized design. Problem is now to decide how to design logical model to store this information, so Data Manipulation Language can be continued used to make consultations, create reports,... At database implementation level, there are two possibilities in function of the DBMS type used:

- If a **Relational DBMS** is used, then solution goes to use **subrogates**. Those subrogated are attributes that, being only and without admitting null values, act as if went pointers to another line of another table where value of the normal attribute and value of each dimensions qualifying to this attribute will be stored. Table shown in figure 1 would have the representation of figure 2. In the source table, the fields *VoteIntention* and *Votes* could be substituted by the subrogates *IdVoteIntention* and *IdVotes*, which spread to another table where all of subrogated, the source fields and the value of the attribute of quality will be stored.

- If an **Object-Relational DBMS** is used, information can be stored in form of nested tables or of objects. Figure 3 represents the source code in SQL, for the structure that is wanted to represent.

Name	IdVoteIntention	IdVotes
Smith	NS001	NC001
Sanders	NS002	NC002

IdVotes	Votes	Date	Source
NC001	49%	05/11/99	Electoral Office
NC002	51%	05/11/99	Electoral Office

IdVoteIntention	Votes	Date	Source
NS001	80%	30/10/99	NYT
NS002	20%	30/10/99	NYT

Figure 2: Use of subrogates to solve the problem of storing quality information.

```
Create type VotesInformation as Object{
        Votes number;
        Date date;
        InformationSource varchar2(5)
}
Create table Candidates{
        Name varchar2(20);
        ExpectedVotes VotesInformation;
        RecollectedVotes VotesInformation;
}
```

Figure 3: SQL Code for object creation where to store the information quality.

3. METHODOLOGY STRUCTURE.

The main objective of CALIDAT, below described, is to offer to users a framework to determine data quality of database according to proper quality of data. What intends in this framework is, after analysing the users quality requirements, look for the most significant quality dimensions according to the users necessities, obtain values for this dimensions according to the data and to analyse the measures applying some valuation judgments.

As CALIDAT is a methodology for quality control, it is recommended using typical quality tools, such as data collection forms, histograms, Pareto's diagrams, cause /effect diagram, scatter diagrams, control graphics, analysis for stratification,...

CALIDAT is composed by a total of four well differentiated and independent phases. Each one of these phases can be formed by some activities. Advisable is to follow phases in a serial order, but there will be occasions in those that can be jumped some by not contemplating them the measurements objectives. Phases, drawn in figure 4, are following ones:

1. **Phase 1: Objectives and Measures Identification**. This is an analysis phase, where starting from users data quality requirements, a set of work products would be obtained after completing each one of following activities:
   - 1.1. **Determine the measurements objective**. It is to determine reasons for which data quality level wants to be determined.
   - 1.2. **Determine the quality parameters and indicators**. Starting from the users requirements, the most significative parameters and indicators are identified to delimit data quality problems.
   - 1.3. **Locate the data to value**. This activity is divided into the following subactivities:
     - 1.3.1. **Determine amount of data should be valued**. It would be to decide if to determine the data quality is necessary to take all the data or it would be enough with taking a sample of them and then to extrapolate the results to overall.
     - 1.3.2. **Locate data in the database.** It is sought to indicate the exact place where logical and /or physically data are. If a relational database was using, it was where the data is, it would be of identifying the table and the attribute where they are.
     - 1.3.3. **Choose the moment in which the data valuation should be made.** It can happen data quality state is truly significant it is occurred in a certain moment. It is to define moment so measure of quality is the appropriate one.
     - 1.3.4. **Identify data sources.** To be able to compare a store data with real one, it needs know data source to ask it which was the original datum.
   - 1.4. **Quality Criteria Definition.** It is to establish valuation criteria to judge how good is a datum and to define evaluation criteria to determine how good is the overall group of data. An example of valuation criterium for individual data can be the ownership or no to a defaulted values range for the quality indicator. An example of evaluation criteria to value the global quality of a data set can be to determine a threshold (percentage) such if it is not overcome (there is not a percentage of data verifying the valuation criterium) the population is not accepted like valid for its low quality.
2. **Phase 2: Quality Structure Creation.** This is a design phase, where objective is to make system have a storing structure to save data quality values in which later will be collected for the quality attributes. Depending on used DBMS, one of the two solutions mentioned before will be the right one. In the case of using a relational DBMS, this solution consist of adding tables to store valuations that are made later on the quality (Figure 2); while if a ORDBMS is had, objects or nested tables can be created (Figure 3). In function of the number of times that it has been analysed a database, one of this situations can be presented:
   - 2.1. **There is not even database:** A one will have to be designed adding it directly the quality aspects considered as the most appropriate according to the utilized DBMS (subrogated or nested tables).
   - 2.2. **There is database but that does not support for data quality**. Logic model will have to be modified for supporting subrogates or nested tables.
   - 2.3. **Database already has data quality structure previously created.** In this situation there are possible two options:
     - 2.3.1. A study on an already analysed attribute will be made:
       - 2.3.1.1. Quality indicator is the same one that the one of the previous time, with no change will have to made.

2.3.1.2. A new quality indicator is required, then it will need spread the subrogated one to a new table for a new quality indicator value.

2.3.2. A new study will be made about an another attribute not yet analysed: changes will have to be made as necessary as if was the situation 2.2. In any of these circumstances in which it needs modifying the database, it must be realized that all the processes managing those data can be affected, so it is recommended modifying them or creating triggers, if the DBMS gives them support to those changes.

3. **Phase 3: Quality Attributes Measurements.** Once data system has a structure to store the quality dimensions measures, this phase consists on recollecting values for the quality indicators. It could be necessary for some quality dimensions know the real datum and compare it with the stored one. Depending on data quantity and on required quality level it could need measure the indicators of all the data or select for sampling only a part of that whole. Anyway these mensurations will be stored in tables containing subrogates or in objects. If new analysis are required, new versions of indicator can be generated to compare the evolution of the process of cleaning or to erase the values that previously had been stored.
4. **Phase 4: Analysis and Evaluation of the values of quality attributes.** In this phase, individual values of last phase have taken measures will be judged according to the valuation criteria to determine the degree of kindness of a datum. After this, and according to the number of data with quality and realizing established evaluation criteria, data about data quality will be judged if those data have or no the desired degree of accuracy. If so, data become certified as valid for the application. Otherwise they are discarded as invalid, proceeding later as better as it suits: correction of existent data or collect new data. Anyway data will be evaluated again while they does not pass the evaluation criterium.

4. CONCLUSIONS

Day by day, more and more companies are realizing of importance of maintaining quality in their databases. These data are the assets that will allow them to obtain information, by markets analysts, or by automatic methods (data mining) to potentialize their business activity. If this information is well managed, organizations can use it to make decisions that redound in important benefits. But to take this type of decisions, it is needed databases have data verifying users quality requirements. To complete this objective is needed value the existent data quality. CALIDAT is a contribution to this task. It proposes a series of structured phases for obtaining of results allowing knowledge workers analysing accuracy degree of their raw materials. They are a total of four phases in those the objectives and measures are indicated that should be identified, the modifications which is necessary to make to the databases and the possibilities when measure and value the data. The results of the application of this methodology, we expect, suppose an advance in the use of the resources of information of the companies toward the global quality of the same one.

5.ACKNOWDGMENTS

This research is part of the CALIDAT project, developed in collaboration with CRONOS IBÉRICA, S.A. and Conserjería de Educación de la Comunidad de Madrid. (09/0013/1999).

6.BIBLIOGRAPHY AND REFERENCES

[1] Ballou, D., Wang, R., Pazer, H., y Tayi, G.K., (1994). *Modelling Information Manufacturing Systems to Determine Information Product Quality*. Management Science, 44(4), 1998 pp.462-484

[2] De Miguel A., Piattini, M., y Marcos E. (1999) *Diseño de Bases de Datos Relacionales*. Ed. Ra-Ma

[3] English, L. (1999) *Improving Data Warehouses and Business Information Quality.* John Wiley & Sons.

[4] Genero, M., Calero, C., y Piattini, M. (1999) *Asegurar la calidad de las Bases de Datos... un reto para el año 2000*. Cuore N°2, Julio 1999, pp 28-35

[5] Huang, K.T., Lee, Y., y Wang, R. (1999) *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River

[6] IEEE Std 830-1993. *IEEE Recommended Practice for Software Requirements Specifications.*

[7] ISO (1998) *Software Product Evaluation – Quality Characteristics and Guidelines for their Use.* ISO/IEC Standard 9126, Geneva

[8] Madnick, S., y Wang, R. (1992). *Introduction to TDQM research.* Total Data Quality Management (TDQM) Research Program, MIT Sloan School of Management, Cambridge, MA. *http://web.mit.edu/tdqm/www/papers/92/92-01.html*

[9] Orman, L., Storey, V., y Wang, R. (1994) *Systems Approaches to Improving Data Quality.*, August 1994. *http://web.mit.edu/tdqm/www/papers/94/94-05.html*

[10] Orr, K.(1998) *Data Quality and System Theory*. Communication of the ACM, 41 (2), pp 66-71.

[11] Pierce, E. (1997) *Modeling Database Error Rates*. Web de Data Quality. September 1997. *http://www.dataquality.com*

[12] Redman, T.C.(1996) *Data Quality for the Information Age*. Artech House Publishers, Boston.

[13] Storey, V., y Wang, R. (1994) *Modeling Quality Requirements in Conceptual Database Design. http://web.mit.edu/tdqm/www/papers/94/94-02 .html*

[14] Wand, Y., y Wang, R. (1994) *Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM (CACM)*, 39, (11), pp 86-95

[15] Wang, R., Kon, H., y Madnick S. (1993) *Data Quality Requirements Analisys and Modeling.* Published in the *Ninth International Conference of Data Engineering* Vienna, Austria. Pp 670 – 677

[16] Wang, R., Reddy, M.P., y Kon, H. (1992) *Toward Quality data: An Atribute-based approach. Journal of Decision Support Systems* (DSS) 13 pp 349-372.

[17] Wang, R., Strong, D., y Guarascio, L. (1994b) *Beyond Accuracy: What Data Quality Means to Data Consumer. http://web.mit.edu/tdqm/www/papers/94/94-10.html*

[18] Wang, R., Strong, D., y Guarascio, L., (1994a) *Data Consumers' Perspectives of Data Quality. Information Systems Research* (ISR) *http://web.mit.edu/tdqm/www/papers/94/94-01.html*

[19] Willshire, M. J. *A process for Improving Data Quality.* September 1997. *http://www.dataquality.com*

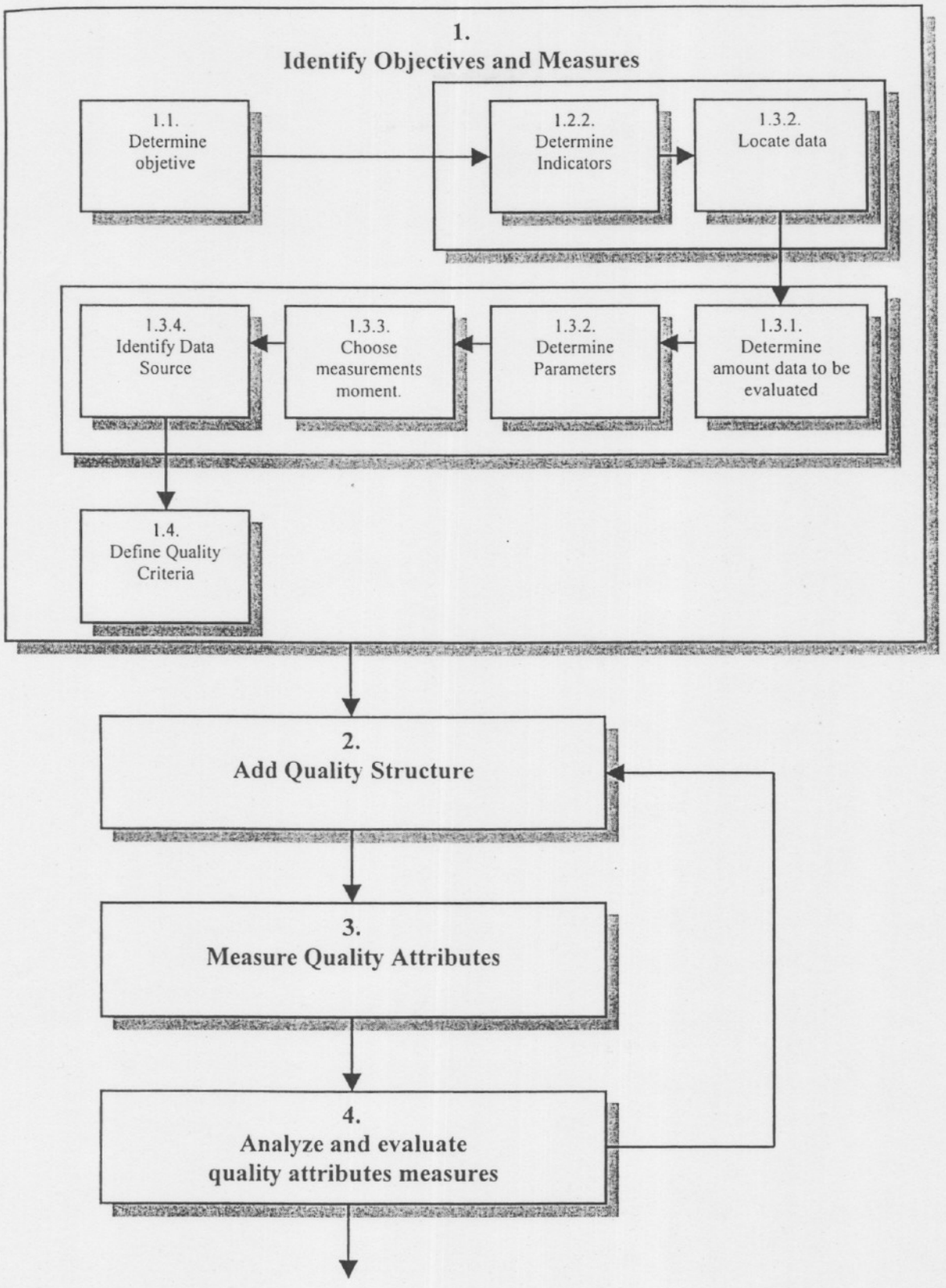

Figure 4: CALIDAT Methodology