Subjective Assessment of Data Quality considering their Interdependencies and Relevance according to the Type of Information Systems

María del Pilar Angeles, Francisco Javier García-Ugalde
Facultad de Ingeniería
Universidad Nacional Autónoma de México
México, D.F.
pilarang@unam.mx, fgarciau@unam.mx

***Abstract* — The assessment of Data Quality varies according to the information systems, quality properties, quality priorities, and user experience among other factors. This paper presents a number of user stereotypes, a study of the relevance of the quality properties from experienced users mainly at the industry and an assessment method for subjective quality criteria considering their interdependencies. A Data Quality Manager prototype has been extended to suggest quality properties, their corresponding priorities and the possibility of subjective assessment of secondary quality criteria. The relevance and assessment of quality criteria according to the type of Information Systems is presented and validated.**

Keywords-data quality; subjective assessment; user stereotypes; quality framework; data cleansing.

I. INTRODUCTION

The prime motivation for the research is that when users query a database system, they get returned a set of data which is inherently presented as perfect, original, and atomic. Users have no information by which to judge its quality and whether data comes from a number of data sources or by a transformation function. We have developed a Data Quality Manager (DQM) [1], [2], [3], [4] in order to objectively assess data quality within heterogeneous databases. The DQM is composed of a generic Data Quality Reference Model, a Measurement Model, and an Assessment Model. The data quality criteria classification as primary and secondary dimensions has been also previously identified in [5]. The assessment of data quality has been classified as objective and subjective assessment in [11].

The DQM was originally designed to assess primary data quality properties such as currency, response time, volatility. The assessment has been done at different levels of granularity by considering data provenance and aggregation functions. Therefore, the DQM was unable to assess subjective data quality properties.

Further work has shown that the overall assessment of data quality depends on the quality properties chosen as quality indicators, and the priority of each quality property might change the final quality score [2], [5].

Subjective assessment of data quality is not an easy task for naive users without enough experience. Data consumers of a Decision Support System (DSS) might prefer some data against other because of the reputation that data producers have as well as the credibility and relevance of data for the task at a hand, or the level of satisfaction they have on making strategic decisions effectively from using reliable data. Furthermore, data consumers of operational systems might be more interested in timeliness, response time, and accessibility of data for an effective On-Line Transaction Processing (OLTP) than completeness or relevance of data.

Within the DQM the specification of which quality properties and the priority of those quality properties were meant to be established by expert users. However, in the case of non-expert users, the DQM has no suggestions to make.

In order to establish a data quality assessment tool that can help naive users according to the type information system and to implement the assessment of subjective quality properties to provide more information to expert users, we have established the main objectives of the present paper.

a) *Data Quality Reference Model improvement*: To identify a set of relevant quality properties according to the type of Information Systems (IS) and the role of user, named as user stereotypes, part of this work has been published in [6].
b) *Data Quality Assessment Model improvement:* To distinguish a set of data quality priorities from user prototypes, to establish their corresponding weights during the assessment process, part of this work has been published in [1], and to determine a subjective assessment method in order to incorporate secondary quality properties.
c) *Data Quality Manager Prototype enhancement:* To be able to suggest a set of ranked data quality properties to inexperienced users to assist them with the analysis of a number of data sources to query the best ad-hoc ranked data sources to support informed decision making.
d) *The implementation of a generic and flexible questionnaire* for the subjective assessment of secondary data quality criteria by the aggregation of its components.
e) *The implementation of the data quality interdependencies* identified in [6] as part of the subjective assessment method within the DQM.

The following Section is focused on previous research

concerning quality criteria classifications, and types of assessment of quality criteria. The third Section is related to the assessment of quality by considering quality properties interdependencies. The fourth Section presents a questionnaire for the subjective assessment of quality properties considering their interdependencies. The fifth Section presents the ranking of quality priorities according to the IS and role of user. The sixth Section presents a survey that was conducted to provide a ranking of quality properties according to the type of Information System from experienced users. The seventh Section is aimed to the implementation of the relevant quality properties, their corresponding ranking and the subjective assessment within the Data Quality Manager, in order to provide automatic, semi-automatic and manual assessment of data quality. The last Section concludes with main achievements and future work.

II. Previous Work

We will present a number of relevant data quality classifications and types of assessments of quality criteria.

A. Data Quality classifications

There are a number of quality criteria classifications, the difference between concepts and classifications rely mainly on user focus according to the role and experience [1], [2]. However, our proposal is focused on the implementation of assessment methods; this Section only presents the assessment oriented model explained in [22] and the data quality classification according to their measure interdependencies.

The Assessment Oriented Model: Quality criteria have been classified in an assessment-oriented model by F. Naumann in [22], where for each criterion an assessment method is identified.

The scores of objective criteria are determined by a careful analysis of data. In this classification the user, the data, and the query process are considered as sources of information quality by themselves.

Individual users determine the scores of subjective criteria based on their experience, knowledge, and focus. Therefore, subject assessment is recommended in case of experienced users.

Object-criteria and process-criteria have been utilized for an unbiased assessment of data within the DQM for any level of user experience. However, subjective criteria had not been considered within the original DQM.

The measurement dependencies classification: The measurement of a quality criterion might be part of the measurement of an aggregate one. The quality dimensions, which measurements derive from primary criteria, are identified as secondary quality properties [2], [5]. We have identified some relationships between these quality properties based on their definitions from previous research [7] [8], [9], [10], [11], [12], [13] [14], where the quality properties are only defined. However, there is no identification of interdependencies. The metrics or assessment methods identified in previous research were established with no consideration of interdependencies among subjective quality criteria. The secondary quality criteria definitions and their relationship with primary criteria are as follows:

Primary Quality Criteria: From the Data Quality Reference Model, we have identified a number of criteria, which measurement does not depend on other quality criteria, namely Primary Quality [2], some of them are presented in Table 1.

Table 1 Primary Quality Criteria

Accuracy	Format Precision
Currency	Format Flexibility
Efficient use of storage	Volatility
Response time	Representational Consistency
Availability	Concise Representation
Amount of data	Appropriateness of Format
Unbiased data	Uniqueness

Secondary Quality Criteria: This Section presents a set of secondary quality criteria, their conception and measurement are established on a primary or secondary quality property. These properties are mainly assessed by subjective methods and some of them are presented in Table 2.

Table 2 Secondary Quality Criteria

Interpretability	Completeness
Reliability	Timeliness
Reputation	Ease to use
Credibility	Accessibility
Usefulness	Cost
Added Value	

B. Types of data quality assessment

Objective assessment may use metrics with no consideration of the context application, or may use task dependent metrics, which include the organization's business rules, regulations, and constraints provided by the database administrator, to be applied to any data set [12].

Cleansing techniques: In order to correct, standardize and consequently, to improve data quality, data cleansing has emerged to define and determine error types, search and identify error instances, and correct the errors. "Data cleansing is applied especially when several databases are merged. Records referring to the same entity are represented in different formats in different data sets or are represented erroneously. Thus, duplicated records will appear in the merged database. This problem is known as

merge/purge problem." [21].

According to [20] the most common methods utilized for error detection are:

a) Statistical methods through standard deviation, quartile ranges, regression analysis, etc. [18], [19].

b) Clustering that is a data mining method to classify data in groups to identify discrepancies [25].

c) Pattern recognition based methods to identify records that do not fit into a certain specific pattern [25].

d) Association rules to find dependencies between values in a record [21].

Performing Data cleansing in offline time is unacceptable for operational systems. Therefore, cleansing is often regarded as a pre-processing step for Knowledge Discovery in Databases and Data Mining systems during the Extraction Transformation and Load (ETL) process. However, it is still a very time consuming task, "The process of data cleansing is computationally expensive on very large data sets and thus it was almost impossible to do with old technology" [21].

The Parsing technique: By considering the actual data or a metadata, it is possible to determine if a given string (in this case an entire tuple or an attribute) is an element of the language defined by the grammar. Accuracy is commonly assessed by this method.

The assessment of value consistency is calculated objectively by parsing or cleansing techniques.

Sampling: Samples of data are considered appropriate for finding the score of the entire data source. This method is often used for completeness, and accuracy criteria.

Continuous assessment: In case of dynamic criteria, quality assessment is executed at regular intervals. Continuous assessment is required for timeliness, response time, and availability criteria.

Subjective assessment depends upon the user experience, the task at hand, and the use of questionnaires [7], [19].

User experience: Data quality is assessed depending on previous user experience and knowledge of the specific domain and data sources. For instance, reputation and believability are criteria suitable to be judged by user experience assessment.

User sampling: A user will assess data by analyzing several sample results. The user should be skilled enough to find appropriate and representative samples. In the case of interpretability of data, users find which attributes are more suitable for sampling than others are.

Continuous user assessment: In the case where finding representative samples of data is not possible, the user needs to analyze every data, not only samples. That is the case of relevancy or amount of data.

Contract: The assessment is performed depending on the terms of the contract of agreement between the provider and the data consumer, which is the case of price or cost of data [23].

One example of subjective assessment is the use of control matrices proposed by E. Pierce in [Pierce04], to audit the information products. The evaluation is in terms of how well they meet the consumer's needs, how well they produce information products, and how well they manage the life cycle of the data after it is produced. The information product manager shall perform the evaluation.

The columns of the control matrix utilized by E. Pierce are the list of data quality problems and the rows correspond to the quality checks or corrective process exercised during the information manufacturing process to prevent them. Each cell shall contain a rating that can have three different forms:

a) The values Yes or No, whether the quality check exists or not.

b) The category of effectiveness at error prevention, detection, or correction ranked as "low", "moderate", or "high".

c) A number to describe the overall level of assessment of the quality check its effectiveness.

From the objective assessment perspective, the original DQM prototype had implemented the parsing technique, sampling and continuous assessment. In the case of subjective assessment this proposal is aimed for use in questionnaires.

III. ASSESSMENT CONSIDERING DATA QUALITY INTERDEPENDENCIES

The present Section is aimed to describe the data quality interdependencies and how they are utilized as part of a new subjective assessment as part of a novelty of the proposal.

A. DQ interdependencies

The assessment of secondary quality criteria relies on data quality interdependencies by the scores aggregation of its components, which in turn might be a primary and/or a secondary quality criterion.

From the Data Quality Reference Model presented in [2], [5], we have identified a number of criteria whose measurement does not depend on other quality criteria as Primary Quality Criteria. Here we present very briefly the following data quality property definitions.

Accuracy: "A measure of the degree of agreement between a data value or collection of data values and a source that has been agreed as being correct." [9].

Timeliness Is the extent to which the age of data is appropriate for the task at hand [6], and is computed in terms of currency and volatility.

Currency Time interval between the latest update of a data value and the time it is used [14].

The measurement of a quality criterion might be part of the measurement of an aggregate one. The quality dimensions, whose measurements derive from primary criteria, are identified as secondary quality properties.

Completeness is the extent to which data is not missing [12], [23], it is divided by two quality dimensions coverage, and density in [11].

The interpretability dimension is the extent to which data are in appropriate language and units, and the data definitions are clear [15]. Thus, it depends on several factors: If there is any change on user needs, its representation should not be affected, this can be possible with a flexible format; The data value shall be presented consistently through the application and that the format is sufficient to represent what is needed and in the proper manner.

Reputation is the extent to which data are trusted or highly regarded in terms of their source or content [12]. Three factors shall be considered at measuring time: reputation of data should be determined by its overall quality. If authors of data provide inaccurate data then they are unreliable and their reputation shall therefore be decreased. Commonly reputation might be increased if data producers have enough experience gained across the time. For instance, when data owners produce accurate data consistently, modify data as soon as possible when mistakes are found, and they in turn recommend data producers of excellent quality data.

Accessibility is the extent to which data is accessible in terms of security [14], availability and cost.

Data might be available but inaccessible for security purposes, or data might be available but expensive.

Data is *credible* as true [12] if it is correct, complete, and consistent.

Usability is the extent to which data are used for the task at a hand with acceptable effort. In other words, users prefer data that are useful and ease to use.

Usefulness is the degree where using data provides benefit on the job performance. In other words, the extent to which users believe data are correct, relevant, complete, timely, and provide added value.

Easy to use is the degree of effort users need to apply to use data [13]. This effort is in terms of understand ability and interpretability as the resources needed to achieve the expected goals. However, it is common that users use determined data sources, due to the reputation of data producers. The measurement of usability allows users to decide on the acceptance of data, and select a specific datum, data or data source among other alternatives.

Data is *reliable* if it is considered as unbiased, good reputation [15] and credible [7].

The *value-added* is stated in terms of how easy it is to get the task completed, also named as effectiveness; how long could the task take known as efficiency; and the personal satisfaction obtained from using data [23].

B. Assessment & Measurement Model

There are some interdependencies very straight forward to compute. For instance, we have already mentioned that timeliness is the extent to which the age of data is appropriate for the task at hand [6]. Therefore it shall be computed in terms of currency and volatility and fused with an aggregation function as presented in Table 3.

Table 3 MEASUREMENT OF TIMELINESS

Currency	**Volatility**	**Timeliness**
Cu(t)=Time Request – last update time	Vo(t)= Update frequency	T(t)= max(0,1-Cu(t)/Vo(t))

Furthermore, interpretability is assessed in terms of representation consistency, data appropriateness, data precision, and format flexibility. However, we have not established or tested any kind of correlation among them. Consequently, the questionnaire will consider these 4 criteria, it will ask user to choose from 5 possible answers to identify how consistent, appropriate, etc., is the information he or she utilizes.

The formula to assess interpretability considering the subjective criteria already mentioned is shown in Table 4.

Table 4 MEASUREMENT OF INTERPRETABILITY

Possible answer per each criterion a) Representation Consistency, b) Appropriateness, c) Precision, d) Format Flexibility	Formula to compute Interpretability based on 4 answers a, b, c and d.
Possible answer Very Rarely 0 Rarely 25 Occasionally 50 Frequently 75 Very Frequently 100	Interpretability= (a+b+c+d)(0.25)%
Answers: a, b, c, d	

In the case of reliability, this quality criterion depends on credibility, reputation and unbiased data. However, credibility is measured in terms of objective criteria such as accuracy, completeness and consistency. These last two criteria can be measured directly from data by sql queries or by asking user, the decision is up to the user. Once computed credibility in terms of accuracy, completeness and consistency (computed value a), the second step is to compute reputation which in turn is also in terms of further quality criteria (computed value b). The third step is to compute unbiased data (value c). For instance, the question to obtain "unbiased data" would be: "Is the information unbiased enough to believe the decisions made from it would be reliable?" which possible answer would be yes or no. Yes is interpreted as 0 points and Yes as 100 points.

We are considering a conservative estimation of the secundary quality criteria, the final measure will be the average of its components. Please refer to Table 5 for the corresponding formula to assess reliability.

Table 5 MEASUREMENT OF RELIABILITY

Credibility	Reputation	Unbiased	Reliability
(%TotalAccuracy+ %TotalCompleteness+ %TotalConsistency+) *0.33 OR %TotalCredibility	%TotalReputation	Question/ Answer Yes 100 No 0	(a+b+c)(0.33)
Answer: a	Answer: b	Answer: c	Total: %

Fig. 1 presents the data quality interdependencies. These dependencies can help the measurement of such quality properties.

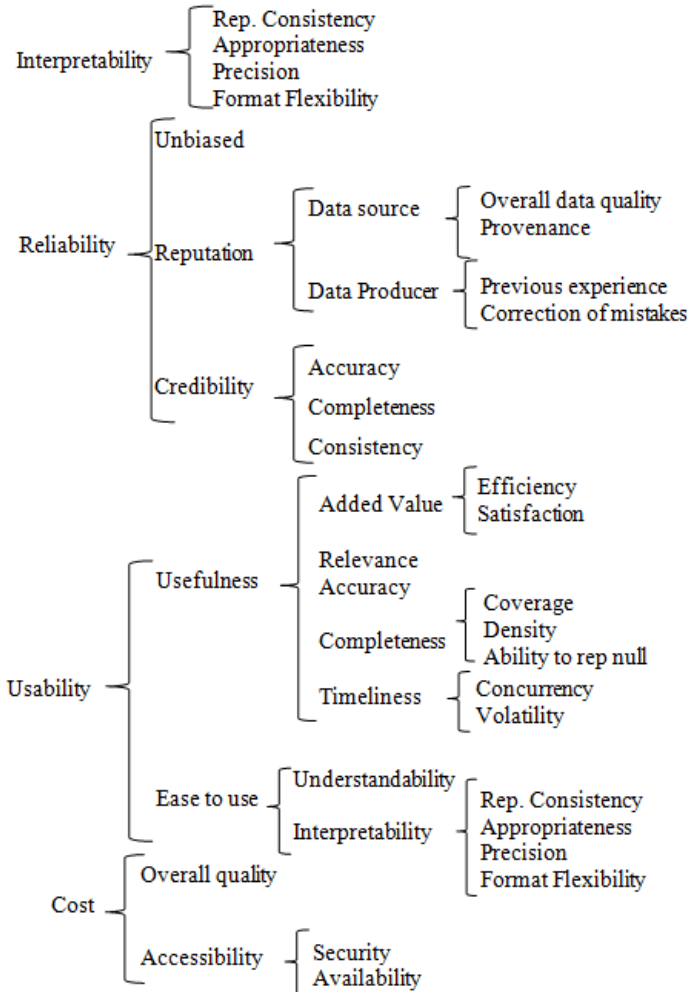


Figure 1. Data Quality Interdependencies

In order to improve the Data Quality Assessment Model, there are four scenarios:

a) *The objective assessment of primary quality criteria* had been considered within the implementation of the original Data Quality Manager for accuracy, response time, currency, uniqueness and volatility.
b) *The objective assessment of secondary quality criteria* had been implemented within the original DQM in the case of completeness, and timeliness.
c) *The subjective assessment of primary quality criteria* will be considered within the enhanced DQM in the case of the 15 properties.
d) *The subjective assessment of secondary quality criteria* will be considered in the case of the 11 properties and their interdependencies.

The last two scenarios have been implemented and will be presented in Section IV.

IV. QUESTIONNAIRE FOR THE SUBJECTIVE ASSESSMENT OF DATA QUALITY

This Section presents a questionnaire for the subjective assessment of some quality properties considering their interdependencies.

In the case of subjective quality criteria such as interpretability, credibility, reputation, representation consistency, reliability, added value, usability, usefulness, ease to use and understandability there is not a practical possibility to assess them directly through SQL-queries but by asking expert users only.

We have designed an on-line questionnaire that can adapt questions according to the most relevant quality criteria and their corresponding interdependencies.

This Section presents the questionnaires developed according to the quality interdependencies for those subjective quality criteria. In the case of expert users, the questionnaire requires what quality criteria user wants to assess according to his or her experience, otherwise the user stereotypes are presented. If the criterion depends on other criteria, then a number of questions are presented to the user in order to measure them and to assess the desired quality criteria. Fig. 2 shows the case of interpretability.

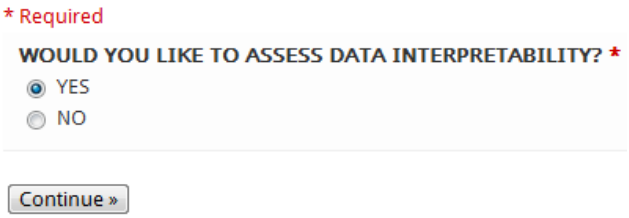


Figure 2. Asking to assess Interpretability

As we have indicated in the previous Section, interpretability relies on representation consistency, appropriateness of data, precision of data and format flexibility. Therefore, five questions are presented to user in order to measure all of them. The possible answers are: very rarely, rarely, occasionally, frequently, very frequently, whereas each of these answers are mapped to a numeric value. Fig. 3 shows the questionnaire for interpretability.

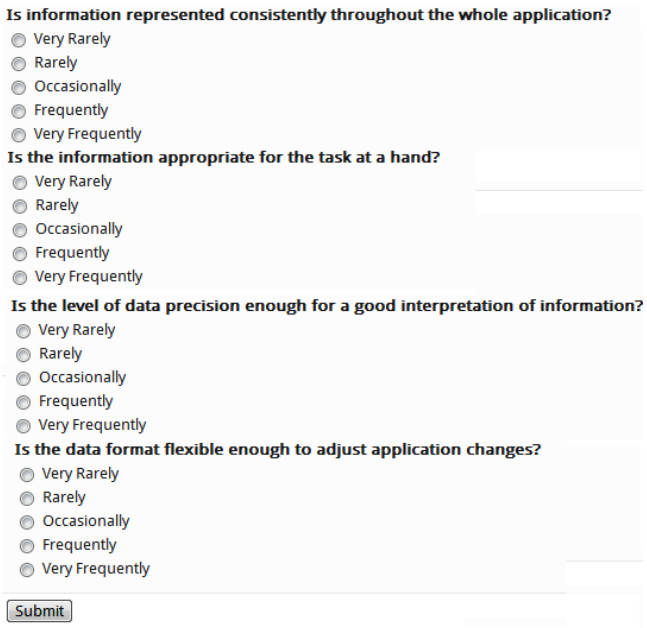


Figure 3. Interpretability Questionnaire

After the questionnaire is completed, the data quality criteria measures are computed and considered within de DQM for the assessment of data quality, or showed as bar graphs at user request. Fig. 4 shows the corresponding bar graph.

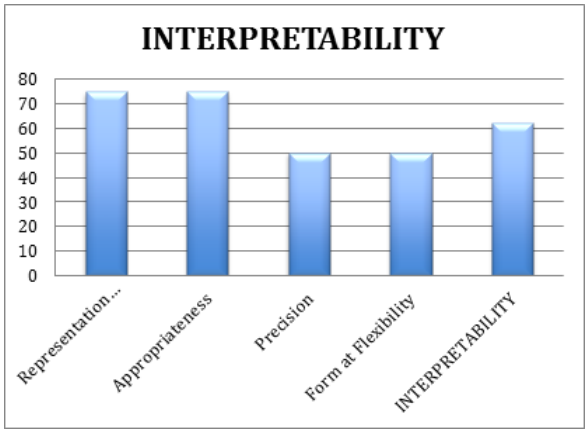


Figure 4. Interpretability assessment as percentage

The assessment of reliability is computed based on unbiased data, credibility and reputation. However credibility and reputation are secondary criteria, credibility measurement depends on completeness, accuracy, and consistency.

Reputation depends on what user trust the most: data source, data provider or both. Therefore, the questionnaire asks the user the corresponding quality properties. Fig. 5 shows the reliability questionnaire.

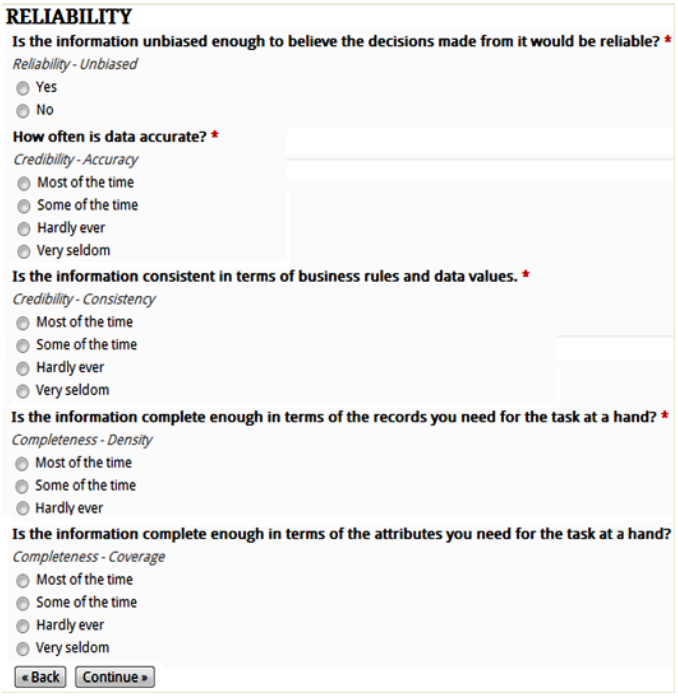


Figure 5. Reliability Questionnaire

The reliability questionnaire is composed of five questions: the first one is relative to unbiased data, the following two questions are regarding credibility of data and the remaining questions are focused on completeness. See Fig. 6 for the corresponding assessment.

Figure 6. Reliability assessment as percentage

According to his or her answers, data consumer can observe from Fig. 6 an 83% of reliability derived from completely reliable data, 80% of credible data because sometimes data is not complete or accurate and the correction of mistakes are not always on time. Due to space restrictions, the present paper is not showing completeness, accuracy and consistency measures for credibility neither measures for reputation.

V. USER PROTOTYPES

The identification and ranking of relevance for data quality properties according to the type of users and Information Systems is not straightforward. For instance, if we consider volatility as the update frequency the relevance of such quality property varies very remarkable according to the application domain, volatility is essential within operational systems, but not quite important within DSS where historical information is materialized.

An Executive Support System (ESS) is designed to help a senior management tackle and address issues and long-term trends to make strategic decisions for the business. It gathers analyses and summarizes aggregate, internal and external data to generate projections and responses to queries. Therefore, the main data quality problem on ESS relies on external data, so decisions depend on accuracy, timeliness, completeness and currency of the external data collected. Furthermore, users are interested in those quality properties that are very much related to their work role.

According to Lee and Strong [10], the responses from data collector, data custodian, and data consumer within the data production process determine data quality because of their knowledge. Data consumers require friendly and usable tools in order to deal with making decisions only rather than the IS per se. Possible inconsistencies might be derived from different data sources so making decisions regarding which external data source to trust is an issue. Response time however, is not of great relevance when the analysis is on long-term trends.

A. Data Collector in DSS

Within a Decision Support System, there are people, groups or even systems that generate, gather or save data to the information systems. Consequently, data collectors impact on accuracy, completeness, currency and timeliness of data. The quality properties identified as the most relevant within Decision Support Systems for data collectors are presented in Fig. 7.

Accuracy, completeness and timeliness shall be presented to the collector user in order to help during the assessment of data quality. Furthermore, completeness is estimated by an aggregation function of coverage, density and ability to represent nulls. Same applies for the rest of the user stereotypes.

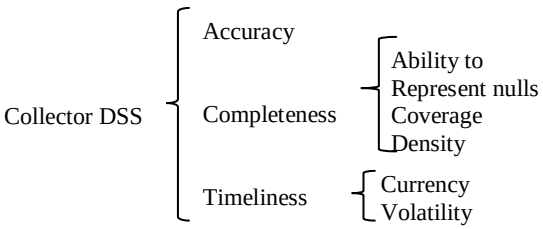


Figure 7. Quality properties for collectors within DSS

B. Data Custodian in DSS

Data custodians are people who manage computing resources for storing and processing data.

In the case of DSS, the process of extraction, transformation and load (ETL) of data within a data warehouse is mainly related to data custodians.

The ETL process is a key data quality factor; it may degrade or increase the level of quality. Therefore, custodians determine the representation of data, value consistency, format precision, appropriateness of data for the task at a hand, the efficient use of storage media.

In other words, appropriateness, concise representation, efficient use of storage media, format precision, representation consistency and value consistency shall be evaluated and presented to them in order to help them decide which data source should be utilized.

Fig. 8 is presented for the relevant quality properties among data custodians within Decision Support Systems.

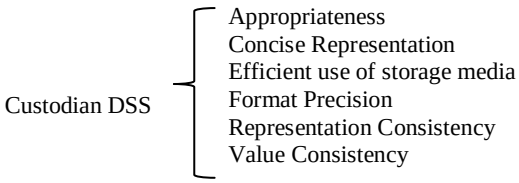


Figure 8. Quality properties for custodians within DSS

C. Data Consumer in DSS

Data consumers are involved in retrieval of data, additional data aggregation and integration. Therefore, they have an impact on accuracy, amount of data relevant for the task at a hand, usability, accessibility, reliability and cost of information in order to make decisions. An analysis on data quality properties in Data warehouses is presented in [8]; such quality properties are included in this work. Accuracy, amount of data, usability, accessibility, reliability and cost shall be considered during data quality assessment. Such quality properties are shown in Fig. 9.

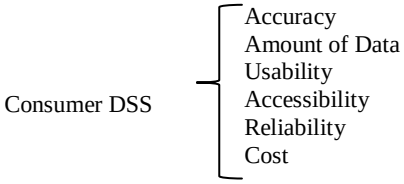
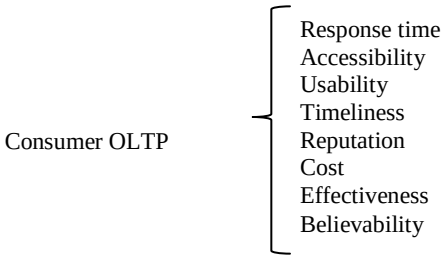


Figure 9. Quality Properties for data consumer within DSS

D. Data Consumer OLTP

As data consumers are involved in retrieval of data the quality properties usability, accessibility, believability, reputation of data sources are key factors for their job. Response time and timeliness [14] are essential within

OLTP systems. From the data consumer perspective accessibility [15] and cost are also very important. The



corresponding quality properties relevant to this role are shown in Fig. 10.

Figure 10. Quality properties for consumers within OLTP systems

E. Data Custodian in OLTP

In transactional systems, data custodians are much related to accuracy, consistency at data value level, completeness [9], timeliness [13], and uniqueness. Therefore the set of quality properties they are interested on for analysis of data quality from their perspective are shown in Fig. 11.

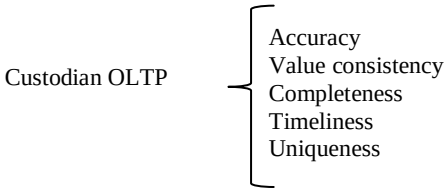


Figure 11. Quality properties for custodian within OLTP systems

F. Data Collector in OLTP

As data collectors within OLTP systems are people who generate information, this role impacts on accuracy, completeness, currency, uniqueness, value consistency and volatility of data. Fig. 12 presents such relevant quality properties for collectors within OLTP systems.

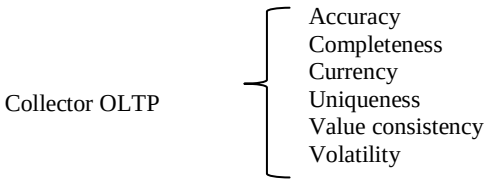


Figure 12. Quality properties for collector within OLTP systems

A number of quality properties have been identified according to the type of users and shown in the past 6 figures. However, there are no priorities assigned to such quality properties in order to assign a weight for assessment purposes.

The following Section shows an example of the online survey developed to provide such ranking.

VI. A SURVEY FOR RANKING DATA QUALITY PROPERTIES WITHIN THE USER PROTOTYPES

This Section presents a survey that was conducted to provide a ranking of quality properties according to the type of Information System from experienced users.

A. Design of Questionnaire

As user experience is substantial within the data quality assessment, a survey was applied to OLTP and OLAP specialists on the web. Therefore, we have conducted an on-line survey requesting an order of importance among the quality properties according to their corresponding experience within a specific Information System. The questionnaire requires the type of information system and what role do users play. According to these two characteristics, the questionnaire presents a set of quality properties and a percentage of relevance these properties should be assigned during quality assessment.

In order to obtain unbiased results, we have invited a number of specialists in operational and DSS information systems around the world. The following groups were invited to participate within the survey:

a) *The University Network of Contribution in Software Engineering and Databases*: To take into account a specialized academic perspective.

b) *The Professionals of Business Intelligence Group:* To retrieve all the experience from a very pragmatic perspective involved in Business Intelligence.

c) *The Very Large Database Group*: In order to obtain the perspective and experience from the databases group.

d) *The Data Quality Pro Group:* The retrieval of the relevance of data quality properties from data quality experts is part of our proposal.

e) *The Information Technology and Communications Group*: For obtaining a broad overview within the Information Technology perspective of the relevant of quality priorities from this group.

Experienced people from these groups have answered our survey, and have established which quality properties are more relevant and under which hierarchy, considering their role and the type of information system in which they are involved.

The questionnaire was designed to be briefly answered, and it was available for six months in order to allow these experts the specification of those quality priorities within the analysis of data quality. For instance, we present in Fig. 13 an example of the one developed to find out the relevance of quality properties for custodian users within OLTP systems.

B. Results of Experiments

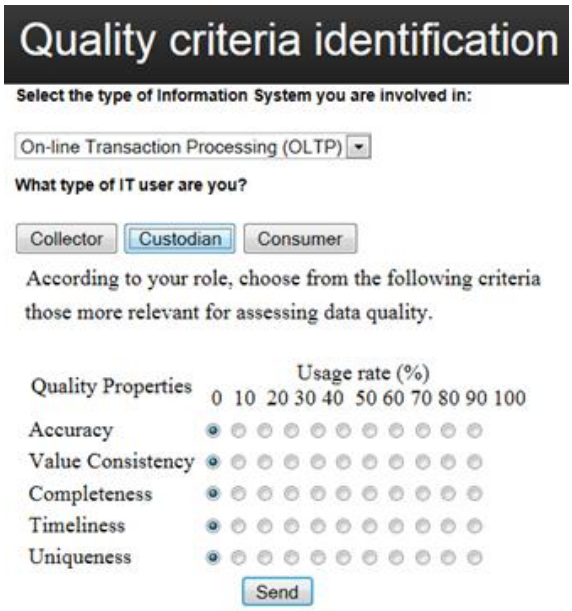


Figure 13. On-line survey for custodian users within OLTP systems

Quality is a very subjective concept; it depends on user experience, information system, business sector, among other factors. As a consequence, we have decided to collect expert opinions in order to identify the most common ranking of such quality properties they utilized during data quality assessment. The results of the on-line survey were analyzed and are presented in this Section.

There were 136 responses collected. Regarding Decision Support Systems 82 DSS specialists participated and expressed their opinion. 22 of them were user collectors, 33 data custodians, and 27 DSS consumers.

According to data collectors, accuracy, completeness, currency and timeliness are the most relevant quality properties to take into account during quality assessment on Decision Support Systems. Accuracy and completeness are equally important with 30 percent each followed by currency and timeliness with 20 percent.

Data custodians considered as first option currency and volatility of data with 30 percent, followed by accuracy, completeness, uniqueness and value consistency with 10 percent each.

Data consumers on the other hand, rely a total of 60 percent on accuracy, amount of data and usability to make their decisions regarding data quality, followed by reliability, easy to use, interpretability and relevance. Refer to Fig. 14 for the data quality prioritization within DSS.

In the case of Online Transaction Processing Systems 54 specialists have participated and answered our online survey, 19 of them were user collectors, 13 data custodians, and 22 data consumers.

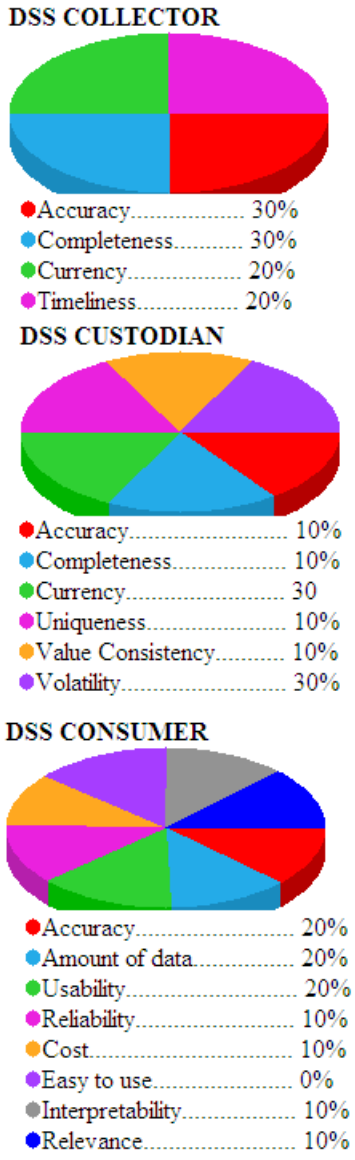


Figure 14. Data Quality prioritization within DSS

On the one hand, data collectors prefer complete data with 30 percent, accurate, and non-duplicated data with 20 percent each, followed by current and consistent information with 10 percent each. Data custodians also trust accurate, unique data with 30 percent each followed by complete data rather than timely data.

On the other hand, data consumers require fast response time, accessible, timely and usable data.

Refer to Fig. 15 for the corresponding data quality prioritization.

The final percentages are obtained through the ranking of the quality properties according to the responses collected.

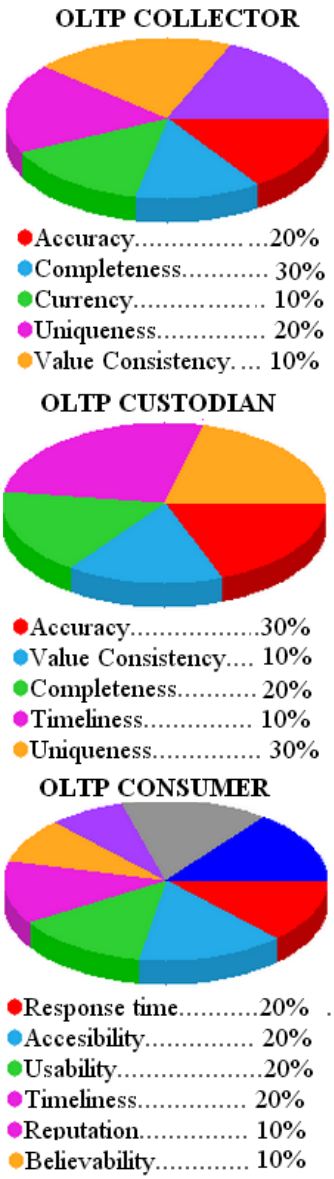


Figure15. Data Quality prioritization within OLTP systems

The present research is looking forward to having more responses in the future by incorporating more specialist groups that allow being more precise with the outcomes and also to test the effectiveness of the stereotypes presented.

In the case of data quality interdependencies, we have taken a conservative approach by computing the average of the components in order to obtain the overall quality measure of an specific secondary data quality property, we have no identified any correlation among them, this correlation is part of our future work.

The following Section presents the Data Quality Manager we have improved by taking into consideration the subjective assessment from experts by the specification of such quality priorities within a weighted matrix during the assessment process carried out by the execution of ranking and scaling methods.

VII. Data Quality Manager Improvement

We have developed a Data Quality Manager as a prototype for the assessment of data quality within heterogeneous databases in [1], [2], [3], [4], in the case of quantitative or primary data quality criteria, the assessment is performed by SQL queries or by the validation of implementation of integrity constraints.

An improvement of such prototype consisted in the implementation of the data quality stereotypes to be suggested to inexperienced users to assist them with the analysis of a number of data sources to identify and query the best ranked data sources and make informed decisions.

The stereotypes implemented are the result of the experiments conducted through the analysis of the results obtained from the online survey and briefly explained in the previous Section.

A. Suggestion of priorities for quality priorities according to the information systems

This Section presents very briefly the improvement of the DQM prototype for the assessment of data quality by suggesting a set of quality properties and their priorities to naive users. In the case of experienced users they still allowed to indicate explicitly their preferences.

For instance, Fig. 16 shows the DQM main menu and the selection of data quality assessment within Online Transaction Processing System conducted by non-expert custodian user.

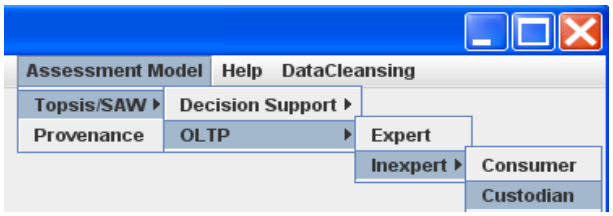


Figure 16. DQM main menu for selecting IS and type of user

Fig. 17 shows how the DQM prototype suggest a non-expert custodian user the most relevant quality properties within an OLTP system, such as accuracy, value consistency, completeness, timeliness, uniqueness and their corresponding priorities according to our expert users. For instance, accuracy and uniqueness are the most relevant quality properties with 30%, then completeness with 20%, finally, timeliness, and value consistency with 10%.

The following Section is aimed to briefly summarize the assessment of data quality. For further information, please refer to [2], [3].

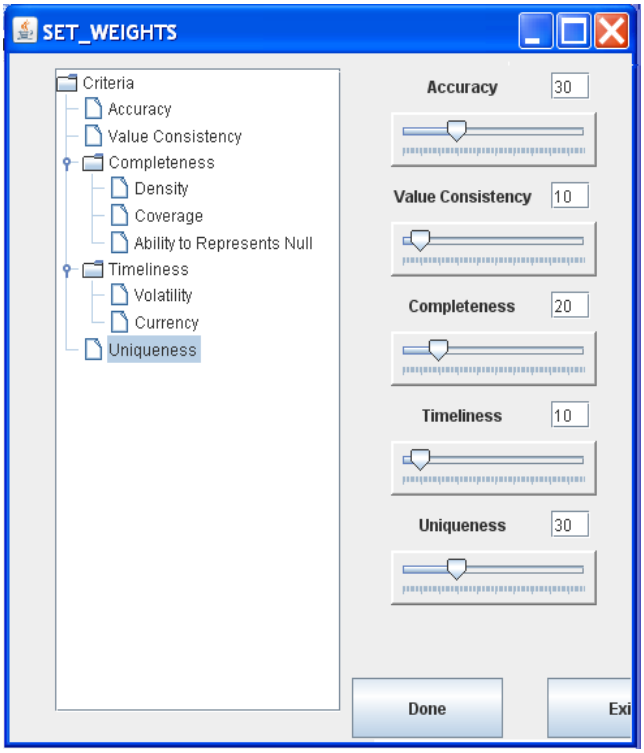


Figure 17. DSS most relevant quality properties

B. *Assessment of Data Quality*

Having all the quality properties prioritized, the weights are normalized. The next step within The DQM prototype is the selection of the data sources, scaling, and ranking methods.

In order to select data sources from a scroll pane, the prototype retrieves from the metadata all the data sources involved in the federation of interest.

The scaling method is selected by pressing its corresponding radio button and the ranking of data sources is executed by pressing the buttons TOPSIS or SAW.

Fig. 18 shows the assessment of data quality properties with Norma and SAW methods respectively.

Please refer to [2], [3] for further information. The overall quality is presented in descendent order in a Text area.

There are three ranked data sources shown in Fig. 18, which were obtained from the TPCC benchmark [24] named TPCCA, TPCCB and TPCCD, where TPCCD contains the best overall data quality.

We have validated the DQM prototype against the specification of the model, and we have verified that the Data Quality Manager (DQM) can provide appropriate information about the qualitative nature of the data been returned from the data sources.

Section VIII presents conclusion and future work.

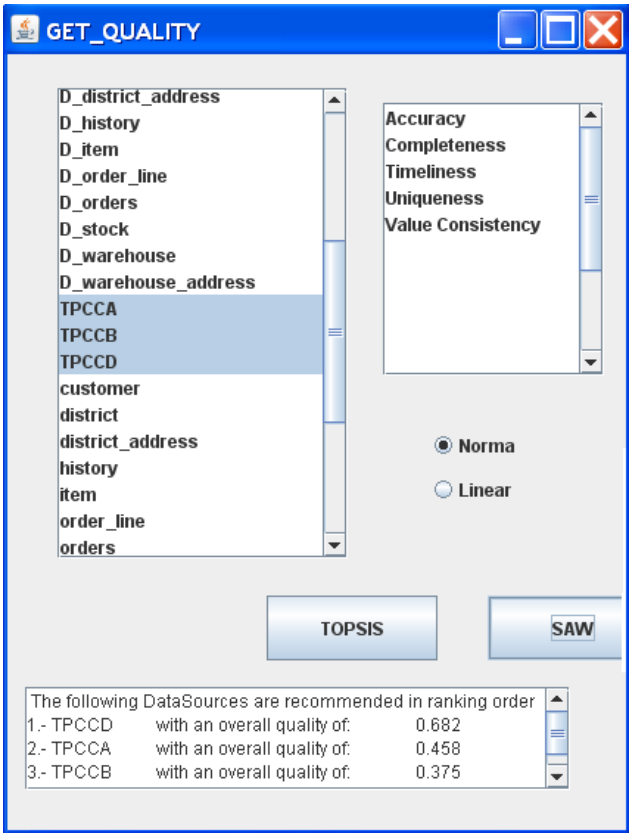


Figure 18. Ranking of data sources according to their quality

VIII. CONCLUSION AND FUTURE WORK

Nowadays data quality has been considered one of the most important factors for making business, especially in terms of profitability and competitive advantage. There are several factors that can affect data quality. Previous research has been conducted in order to propose a data quality assessment framework on the bases of a Reference Model, Measurement Model, and an Assessment Model. These models have been implemented within a Data Quality Manager Prototype, such original implementation was focused only on two out of four scenarios, considering objective assessment of primary criteria, very few secondary criteria where addressed, mainly focused on expert users. The purpose of the present research has been to establish an extended assessment framework and a data quality assessment tool that can help non-expert users according to the type information system and to implement the assessment of subjective quality properties to provide more information to expert users.

The Reference Model has been extended by identifying a set of relevant quality properties according to the type of Information Systems (IS) and the role of user, named as user stereotypes [6].

The Measurement Model has been extended in order to provide specific metrics for subjective quality criteria

considering the interdependencies among them.

The Assessment Model has been extended by identifying a set of data quality priorities from user prototypes, to establish their corresponding weights during the assessment process [1].

The Assessment Model has been enhanced by identifying a subjective assessment method to incorporate secondary quality properties.

The new Reference, Measurement and Assessment Models have been implemented within the original Data Quality Manager Prototype. Therefore, the DQM is now able to suggest a set of ranked data quality properties to inexperienced users to assist them with the analysis of a number of data sources to query the best ad-hoc ranked data sources to support informed decision making.

The Data Quality Manager is able now to assess automatically, semi automatically or manually. However, more information from specialists is required in order to corroborate the prioritization and testing of the effectiveness of the stereotypes identified. This further feedback from the specialists is part of future work.

There are some quality properties whose measurement and assessment methods are suitable to be enhanced as is the case of accuracy and uniqueness. The incorporation or data mining techniques is also part of future work.

In the case of data quality interdependencies, we have taken a conservative approach by computing the average of the components in order to obtain the overall quality measure of an specific secondary data quality property, we have no identified any correlation among them, this correlation is part of our future work.

The present research is part of an effort to improve data quality in order to help users to make business by taking informed decisions. Consequently, after performing subjective and objective data quality assessments, comparing the results of the assessments, is important to identify discrepancies, and to determine root causes of errors for determining and taking necessary actions for improvement.

REFERENCES

[1] P. Angeles and F. Garcia-Ugalde, “Relevance of quality criteria according to the type of information systems”, Proc. International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 12), Mar. 2012, pp. 175-181, ISBN: 978-1-61208-185-4.

[2] P. Angeles and L.M. MacKinnon, “Managing Data Quality of integrated data with Known Provenance”, International Journal of Information Quality, ISSN (Online): 1751-0465, ISSN (Print): 1751-0457, Vol.2 No. 3, 2011.

[3] P. Angeles and F. Garcia-Ugalde, “Assessing data quality of integrated data by quality aggregation of its ancestors”, Computación y Sistemas, Vol. 13 No. 4, ISSN 1405-5546.

[4] P. Angeles and F. Garcia-Ugalde, “Assessing quality of derived non atomic data by considering conflict resolution function”, Proc. International Conference on Advances in Databases (DBKDA 09), Mar. 2009, pp. 81-86, IEEE Computer Society, ISBN: 978-0-7695-3550-0 doi>10.1109/DBKDA.2009

[5] P. Angeles and F. Garcia-Ugalde, “A data quality practical approach, International Journal On Advances in Software, Vol. 2, No. 3, 2009, pp. 259-274, ISSN 1942-2628.

[6] P. Angeles and F. Garcia-Ugalde, “User stereotypes for the analysis of data quality according to the type of information systems”, Proc. International Association for Scientific Knowledge (IASK 08), E-Activity and Leading Technologies, pp. 207-212, Dec. 2008.

[7] Ballou, D.P., Wang, R.Y., Pazer, H., and Tayi, G.K., “Modeling information manufacturing systems to determine information product quality”, Management Science, Apr. 1998, pp. 462–484.

[8] C.Cappiello, C.Francalanci, B.Pernici, P.Plebani, and M.Scannapieco, "Data quality assurance in cooperative information systems: a multi-dimension quality certificate", Proc. International Workshop "Data Quality in Cooperative Information Systems" (DQCIS 2003), pp. 64 -70 Siena, Italy, 2003.

[9] M. Jarke, M.A. Jeusfeld, C. Quix, and P.Vassiliadis, “Architecture and quality in data warehouses an extended repository approach”, Journal on Information Systems, Vol. 24", no.3, pp. 229-253, 1999.

[10] Lee Y. and Strong D. “Knowing-Why about data processes and data quality”, Journal of Management Information Systems, Vol. 20, No. 3, pp. 13 – 39. 2004.

[11] F. Naumann, J. Freytag, and U. Lesser, "Completeness of information sources", Workshop on Data Quality in Cooperative Information Systems (DQCIS2003), pp. 583-615, Cambridge, Mass., 2003.

[12] L. Pipino, W.L. Yang, and R. Wang, "Data quality assessment", Communications of the ACM, Vol. 44 no. 4e, pp.211-218, 2002.

[13] Redman “Data Quality for the Information Age”, Boston MA., London: Bartech House, 1996.

[14] D.M. Strong, W.L. Yang, and R.Y. Wang, "Data quality in context", Communications of the ACM, vol. 40, no. 5, pp. 103-110, 1997.

[15] R.Y. Wang, M.P. Reedy, and A. Gupta, "An object-oriented implementation of quality data products”. Workshop on Information Technology Systems, 1993.

[16] Wang R. Y., and Strong D.M. “Beyond accuracy: What data quality means to data consumers”, Journal of Management of Information Systems, vol. 12, no 4 1996, pp. 5 -33.

[17] R. Wang, "A product perspective on total data quality management", Communications of the ACM, vol. 41, no. 2, pp.58-65, 1998.

[18] Barnett, V., and Lewis, T.: 1984, Outliers in statistical data, John Wiley & Sons, New York.

[19] R.K. Bock, and W. Krischer, ”The data Analysis brief book” Springer 1998.

[20] Buchheit R. B., “Vacuum: automated procedures for assessing and cleansing civil infrastructure data”, PhD Thesis, May 2002

[21] Maletic, J. I., and Marcus, “Data cleansing: Beyond integrity checking”. Proc. Conference on Information Quality (IQ2000) (Massachusetts Institute of Technology, October 2000), pp. 200-209.

[22] F. Naumann, and C. Roker C., "Assessment methods for information quality criteria", Proc. International Conference on Information Quality (IQ2000), Cambridge, Mass., 2000.

[23] F. Naumann, "Quality-Driven query answering for integrated information systems", Lecture Notes in Computer Sciences LNCS 2261, Springer Verlag, Heidelberg, 2002.

[24] TPCH, TPC Benchmark ™ H, Standard Specification Revision 2.3.0 Transaction Processing Performance Council www.tpc.org.info (retrieved: December 2012).

[25] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, 2nd ed.The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, 2006. ISBN 1-55860-901-6.