Relevance of Quality Criteria 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 others. This paper presents a number of user stereotypes and a study of the relevance of the quality properties from experienced users mainly at the industry. A Data Quality Manager prototype has been extended to suggest such quality properties and their corresponding priorities. The relevance of quality criteria according to the type of Information Systems is presented and validated.**

Keywords-data quality; assessment; stereotypes.

I. INTRODUCTION

The analysis of data quality requires a number of indicators as a reference for the assessment of data quality. However, this is not an easy task for naive users with not enough experience. Data consumers of a Decision Support System (DSS) might prefer some data against other because of reputation of data producers, 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).

Previous 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. Refer to [4] for further information.

We have developed a Data Quality Manager (DQM) [1], [2], [3], [4] in order to assess data quality within heterogeneous databases. However, the DQM still required the specification of which quality properties and the priority of those quality properties in order to provide a global assessment [4]. Therefore, the assessment result depends on the user experience.

The purpose of the present research was to identify a set of relevant quality properties according to the type of Information Systems (IS) and their corresponding weights that shall be established for computing the global assessment of data sources. Therefore, we have suggested a number of quality properties according to the information systems and the type of user in [5]. We have also identified the data quality properties interdependencies within the data quality assessment. Therefore, we have decided to conduct a survey to validate or analyze such proposals, and to identify a general estimation of the priorities (weights) that should be considered within a data quality assessment.

As user experience is substantial within the data quality assessment, a survey was applied to OLTP and OLAP specialists on the web.

The identification of a set of user stereotypes with their corresponding weights for the assessment of data quality according to the type of Information Systems is a novelty and the contribution of the present paper.

The following section presents a number of data quality interdependencies identified for the assessment of quality indicators in [5]. The third section analyses the data quality stereotypes to be implemented within a Data Quality Manager. The fourth section presents a survey that was conducted to provide a ranking of quality properties according to the type of Information Systems from experienced users. Such ranking has been implemented as weights during the assessment of data quality within the DQM and presented in the fifth section. The last section concludes with main achievements and future work.

II. DATA QUALITY INTERDEPENDENCIES

From the Data Quality Reference Model presented in [3], 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 is the measure of the degree of agreement between a data value o collection of data values and source agreed to be correct in [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.

Completeness is the extent to which data is not missing [11], [12], it is divided by two quality dimensions coverage, and density in [10].

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*. However, we have not established or tested any kind of correlation among them. We have identified some relationships between these quality properties based on their definitions from previous research as has been

 ISBN: 978-1-61208-185-4

referenced on each quality property definition.

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 [11]. 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 be therefore decreased. Commonly reputation might be increased if authors have enough experience gained across the time. If data owners produce accurate data consistently, modify data as soon as possible when mistakes are found, and they in turn recommend authors of quality data.

Accessibility is the extent to which data is accessible in terms of security [15], availability and cost.

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

Data is *credible* as true [11] 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 is useful and ease to use.

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

Easy to use is the degree of effort user needs to apply to use data [12], because as less effort is easier to use. This effort is in terms of understand ability and interpretability as resources needed to achieve the expected goals. However, it is common that users use determined data sources, due to the reputation of authors.

The measurement of usability allow user 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 [14] and credible [6].

The *added value* is stated in terms of how easy is to get the task complete named as effectiveness; how long could the task take known as efficiency; and the personal satisfaction obtained from using data.

Fig. 1 presents the data quality interdependencies. These dependencies can help the measurement of such quality properties. There are some interdependencies very straight forward to compute. For instance, in order to compute timeliness, currency and volatility are required to be estimated and fused with an aggregation function as presented in Table 1.

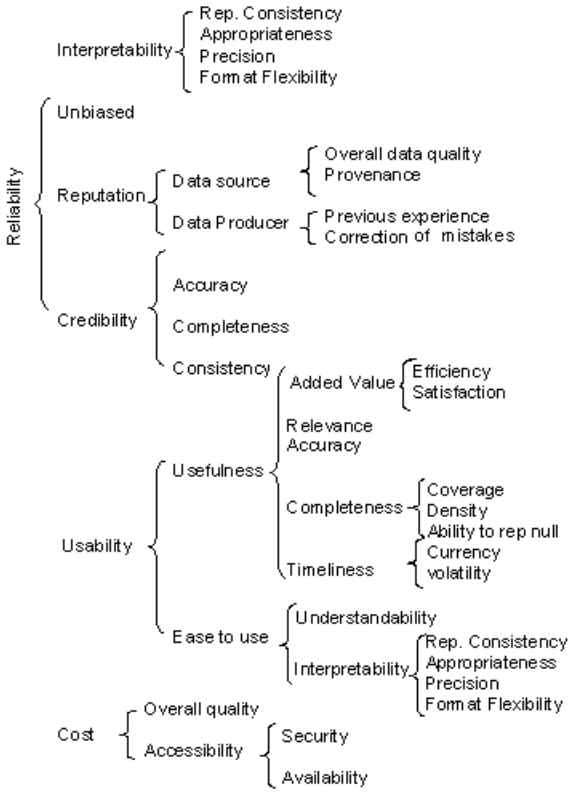

Figure 1. Data Quality Interdependencies

TABLE 1 EXAMPLE OF DATA QUALITY PROPERTIES

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

The present research has been focused mainly in quantitative data quality properties. For further information, refer to [4][5], where Measurement and Assessment model are detailed.

III. DATA QUALITY PROPERTIES ACCORDING TO IS

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 relays on external data, so decisions depend on accuracy, timeliness, completeness and currency of the external data

 ISBN: 978-1-61208-185-4

collected. Furthermore, users are interested in those quality properties that are very much related to their work role.

According to Lee and Strong [9], 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. Therefore, the role of data collector impacts 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. 2. Therefore, 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

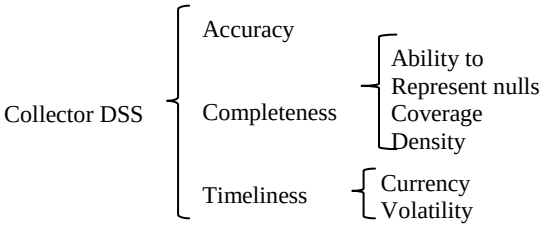

Figure 2. Quality properties for collectors within DSS

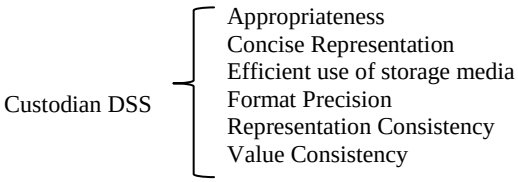

Figure 3. Quality properties for custodians within DSS

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. Refer to Fig. 3 for the relevant quality properties among data custodians within Decision Support Systems. 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.

C. Data Consumer in DSS

Data consumers are involved in retrieval of data, additional data aggregation and integration. Therefore, they 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. 4.

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. 5.

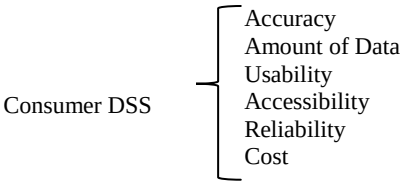

Figure 4. Quality Properties for data consumer within OLTP

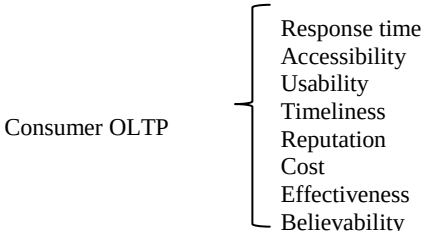

Figure 5. Quality properties for consumers within OLTP systems

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. 6.

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. 7 presents such relevant quality properties for collectors within OLTP systems.

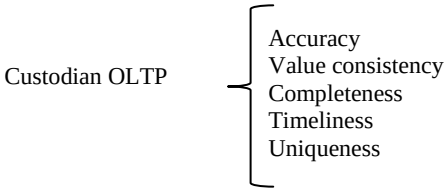

Figure 6. Quality properties for custodian within OLTP systems

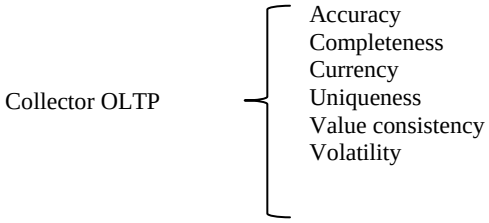

Figure 7. Quality properties for collectors within OLTP systems

A number of quality properties have been identified according to the type of users and shown in the past 6 figures. However, there is no an assessment method to be considered to provide an overall data quality score. Therefore, we require identifying the priorities of such quality properties that is the aim to be achieved in the following section.

IV. IDENTIFYING DATA QUALITY PRIORITIES ACCORDING TO EXPERT USERS

A. Design of Questionnaire

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. The questionnaire was designed to be briefly answered. For instance, we present in Fig. 8 an example of the one developed to find out the relevance of quality properties for custodian users within OLTP systems.

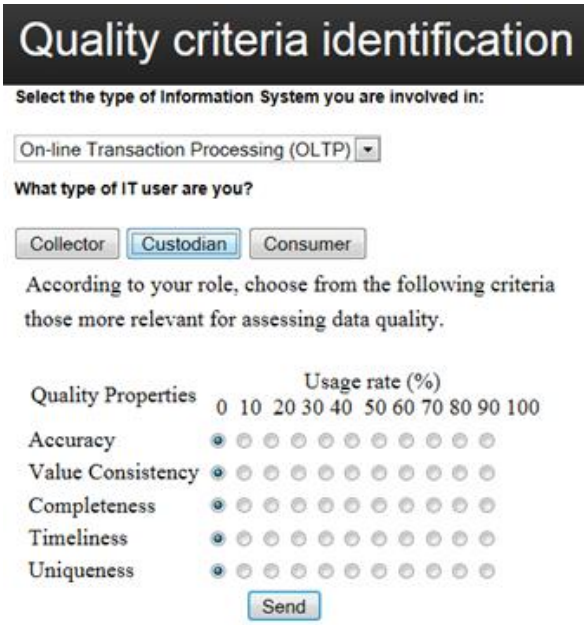

Figure 8 Questionnaire for custodian users within OLTP systems

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: the University Network of Contribution in Software Engineering and Databases, the Professionals of Business Intelligence Group, the Very Large Database Group, the Data Quality Pro Group, and the Information Technology and Communications Group.

The on-line survey was opened for six months in order to allow experts the specification of those quality priorities within the analysis of data quality.

B. Results of Experiments

Quality is a very subjective conception, depending on user experience, information system, and business sector among others. Therefore we have decided to collect opinions and to identify the most common ranking of such quality properties during data quality assessment. The results of the on-line survey were analyzed and are presented in this section. There were collected 136 responses.

Concerning Decision Support Systems there were 82 responses, 22 from user collectors, 33 from data custodians, and 27 from DSS consumers.

In the case of data collectors, they take into consideration accuracy and completeness followed by currency and timeliness. Data custodians prefer to consider as the first option the value consistency of data followed by an efficient use of storage media, the appropriateness of data, concise representation and format precision. Data

 ISBN: 978-1-61208-185-4

consumers on the other hand relay their decisions on sufficient amount of usable and accurate data. Refer to Fig. 9 for the data quality prioritization within DSS.

Regarding operational systems there were 54 responses, 19 from user collectors, 13 from data custodians, and 22 from data consumers.

Data collectors trust the most on accurate, complete and non duplicated data, followed by current and consistent information. Furthermore, data custodians also prefer accurate, unique and consistent rather than timely data. However, data consumers require fast response time, accessible, timely and usable data. Refer to Fig. 10 for the corresponding data quality prioritization.

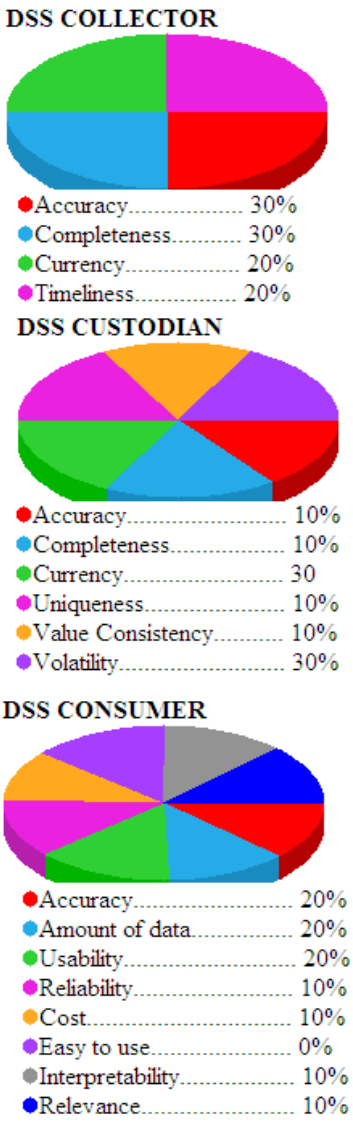

Figure 9. Data Quality prioritization within DSS

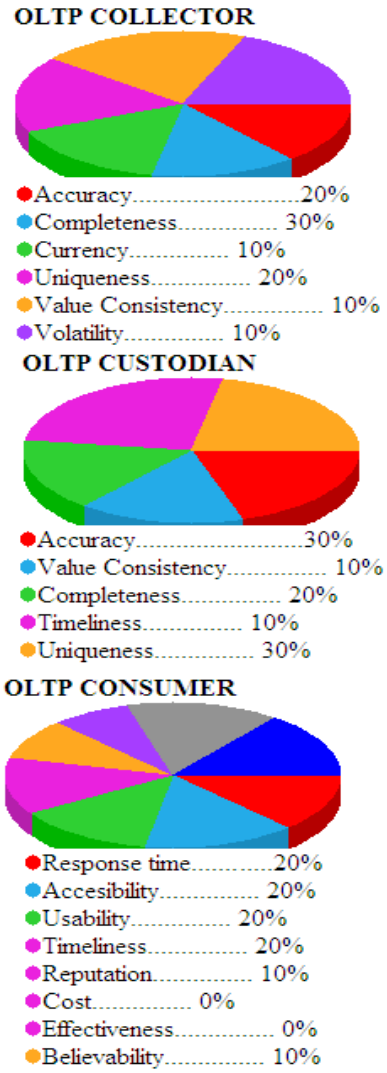

Figure 10. Data Quality prioritization within OLTP systems

The final percentages are obtained through the ranking of the quality properties according to the responses collected. However, the present research is looking forward to have more responses in the future by incorporating more specialists groups that allow being more precise with the outcomes and also to test the effectiveness of the stereotypes presented.

V. 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].

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

 ISBN: 978-1-61208-185-4

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. 11 shows the DQM main menu and the selection of data quality assessment within Online Transaction Processing System conducted by inexpert custodian user.

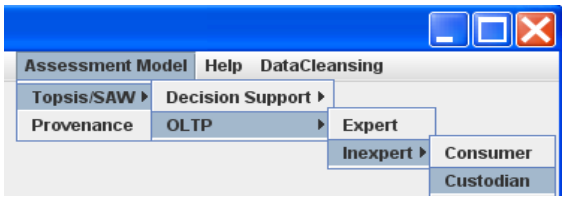

Figure 11. DQM main menu for selecting IS and type of user

The DQM prototype presents in Fig. 12 the most relevant quality properties within a DSS and their corresponding percentages. For instance, accuracy and uniqueness are the most relevant quality properties with 30%, then completeness with 20%, and Timeliness, Uniqueness with 10%.

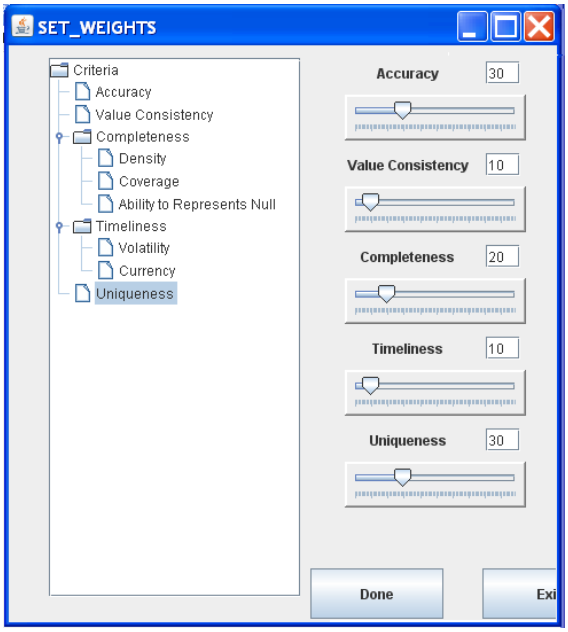

Figure 12. DSS most relevant quality properties

B. Assessment of Data Quality

Fig. 13 shows the assessment of data quality properties of three data sources obtained from the TPCC benchmark [16] named TPCCA, TPCCB and TPCCD, where TPCCD contains the best overall data quality.

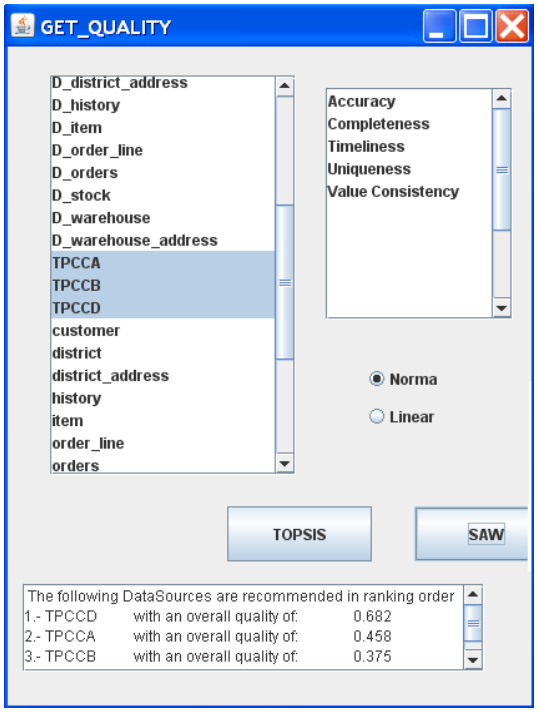

Figure 13. Ranking of data sources according to their quality

VI. CONCLUSIONS AND FUTURE WORK

This document has presented how users might prefer some quality properties during the assessment of data quality according to their role within specific information systems.

We have previously identified data quality interdependency in [5] and now by an on-line survey, the prioritization of such quality properties to the Information Systems.

A number of user stereotypes have been suggested by a Data Quality Manager prototype that is meant to help naive users within the data quality analysis.

The DQM allows the identification of which quality criteria shall be used based on the application domain and the type of users. Furthermore, the user stereotypes presented correspond to data consumer, data collectors, and data custodians. However, more information from specialists is required in order to corroborate the prioritization and testing of the effectiveness of the stereotypes identified is part of future work.

REFERENCES

[1] P. Angeles and L.M. MacKinnon, “Detection and Resolution of Data Inconsistencies, and Data Integration using Data

Quality Criteria", Proceedings of QUATIC 2004: Conference for Quality in Information and Communications Technology, Instituto Portugues da Qualidade, ed., pp. 87-94., ISBN 972-763-069-3, Porto, Portugal, 2004.

[2] P. Angeles and L.M. MacKinnon, "Tracking Data Provenance with a Shared Metadata", Proceedings of PREP 2005: Postgraduate Research Conference in Electronics, Photonics, Communications and Networks, and Computing Science, pp. 120-121, Lancaster England, U.K., 2005.

[3] P. Angeles and L.M. MacKinnon, "Quality Measurement and Assessment Models Including Data Provenance to Grade Data Sources", International Conference on Computer Science and Information Systems" at the Athens Institute for Education and Research, pp. 101-118, Athens, Greece, 2005.

[4] P. Angeles and L.M. MacKinnon, "Management of Data Quality when Integrating Data with Known Provenance", Herriot-Watt University, pp. 1-199 Edinburgh, UK, April 2007.

[5] M.P. Angeles and F. Garcia-Ugalde, "User Stereotypes for the Analysis of Data Quality According to the Type of Information Systems", International Association for Scientific Knowledge, E-Activity and Leading Technologies, pp. 207-212, Madrid Spain, December 2008.

[6] Ballou, D.P., Wang, R.Y., Pazer, H. and Tayi, G.K. "Modeling information manufacturing systems to determine information product quality". Management Science 44, 4 April,1998, 462–484.

[7] C.Cappiello, C.Francalanci, B.Pernici, P.Plebani, and M.Scannapieco, "Data Quality Assurance in Cooperative Information Systems: a Multi-dimension Quality Certificate". Proceedings of the ICDT 2003 International Workshop "Data Quality in Cooperative Information Systems" (DQCIS 2003), pp. 64 -70 Siena, Italy, 2003

[8] 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,
URL "citeseer.ist.psu.edu/jarke99architecture.html", retrieved: December, 2011.)

[9] 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.

[10] 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.

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

[12] Redman "Data Quality for the Information Age", Boston, MA., London : Bartech House, 1996.

[13] 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.

[14] R.Y. Wang, M.P. Reedy, and A. Gupta, "An Object-Oriented Implementation of Quality Data Products". Workshop on Information Technology Systems, O.1993.

[15] 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.

[16] TPCH, TPC Benchmark ™ H, Standard Specification Revision 2.3.0 Transaction Processing Performance Council www.tpc.org.info 2006 (retrieved: December, 2011.)

 ISBN: 978-1-61208-185-4