MDQM: Mediation Data Quality Model Aligned Data Quality Model for Mediation Systems

Loubna Mimouni, Ahmed Zellou and Ali Idri
Mohammed V-Rabat University in Rabat-ENSIAS, Rabat, Morocco

Keywords: Data Quality, Mediation System, Data Quality Model.

Abstract: We are in age where data is provided from multiple heterogeneous and distributed sources hence the usefulness of data integration systems (DIS). But given that the user is obliged to filter a large volume of data to achieve the most satisfactory to his request and his information need, we can say that the approach became more qualitative than quantitative. For this purpose, the main goal of this paper is to introduce the data quality aspect concerning data retrieved from Mediation systems which is the virtual approach of data integration. The paper will be finalized by establishing an attempt of a model of data quality criterions classification in relation to Mediation system (MDQM).

1 INTRODUCTION

Progressively, systems need to integrate data coming from multiple and heterogeneous data sources, while allowing users a uniform access to data without worrying about their types and their formats. These systems are called 'data integration systems' (DIS).

In our researches, we focus more on the Mediation systems (Wiederhold, 1992) which represents a virtual approach of DIS that offer a uniform access to multiple data sources; the queries are split and directed towards various data sources through wrappers, and the results returned by the sources are combined by a mediator and finally sent to the users. Consumers who use the data retrieved from this system need to be sure that the data are sufficiently high quality.

This Quality of data (DoQ) is often defined as the ability of a collection of data to meet user requirement (Naumann and Rolker, 2002). It is therefore important to provide high data quality in mediation systems in such manner that these information systems can be effective, useful and helpful for data consumers. A model which evaluate the data quality in the environment of Mediation system is therefore necessary. For such needs, various studies, (Naumann and Rolker, 2002), (Pipino et al., 2002) and (Wang, 1998) shows that ensuring a good data quality for users is an important challenge which is related to information system success. In the case of virtual Data integration (Mediation), the problem is particularly complex since data is provided by different sources, at different levels of quality.

In this work, we present a new contribution as a classification of Data Quality Model specially designed for mediation system called MDQM. This attempt will be established through a reflection and a discussion around a state of art in the domain.

In order to prove, this paper is organized as follows. In section two we present the concept of data quality. In section three we enumerate and discuss the various approaches to modeling classification of data quality criterions already existing in literature. Section four presents an overview of mediation systems architecture which is the basic context of our researches to assess the data quality. In section five we discuss also our world view of modeling several data quality criterions that impact mainly the data retrieved by the mediation system. Finally, Section six concludes the paper by mentioning some open problems and our future visions to approach them in the coming up works.

Mimouni, L., Zellou, A. and Idri, A.
MDQM: Mediation Data Quality Model Aligned Data Quality Model for Mediation Systems.
DOI: 10.5220/0006958403190326
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 2: KEOD*, pages 319-326
ISBN: 978-989-758-330-8

2 DATA QUALITY CONCEPT

2.1 Data Quality

Data quality is a perception or an assessment of fitness to serve its purpose in a given context, i.e., the ability of a data collection to meet users' requirements (Wand and Wang, 1996). DoQ is a multidimensional and subjective concept since it is usually evaluated by means of different criteria or data quality dimensions and the selection and assessment of the DoQ dimensions that better describe users' data quality requirements mainly depend on the context of use (Pipino et al., 2002) and (Strong, D. M et al., 1997).

For this reason, in the literature, there is no general agreement on the identification of the most important data quality dimensions. Anyway, it is possible to distinguish a small set of DoQ dimensions that are considered relevant in most of the studies.

2.2 Data Quality (DoQ) vs Information Quality (InQ)

To approach the concept of DoQ and in defining the term, most of the literature uses the term Information Quality and Data Quality interchangeably, however, there is difference in meaning between data and information. According to (ISO/IEC., 2008) data is defined in the standard as "a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing".

Data can be considered as the base of information and digital knowledge and takes into account all data types, such as texts, numbers, images and sounds, whereas information is knowledge concerning objects, such as facts, events, things, process, or ideas, including concepts, that within a certain context have a particular meaning.

2.3 Data Quality in Data Integration Systems

Research on data quality started abroad in the 1990s, and many studies proposed different definitions of data quality and division methods of quality dimensions. Particularly, there has been quite a lot of work on DoQ issues in the context of data integration scenarios, especially the use of DoQ in query formulation, processing (mediation) and optimization e.g. (Wang, 1998), (Wang and Strong, 1996), (Weikum, 1996) (Redman, 1996). (Redman, 1998), (Redman,2001). Interesting aspects in general data integration scenarios are how the quality of data from different, heterogeneous and dynamic sources, is assessed and measured and how DoQ dimensions are represented to users and applications.

Several surveys have showed the importance of data quality for end users, in particular, when dealing with heterogeneous data coming from distributed autonomous sources as mentioned before. In what follows, we propose to give an overview in chronological order of this researches which provides from different view of data quality criterions classification models used for specifying requirements and evaluating of Data quality.

3 RELATED WORKS

3.1 Previous Works in DoQ Classifications

3.1.1 Ballou and Pazer (1985)

The final sentence of a caption must end with a period. Ballou and Pazer divided data quality into four dimensions (Ballou and Pazer, 1985), as shown in Table 1. They note that the accuracy dimension is the easiest to evaluate, since this is merely the difference between the correct value and what was actually used.

They argue that the evaluation of timeliness can be carried out in a similar manner. The evaluation of the completeness dimension is also relatively straight forward, as long as the focus is on whether data are complete or not, rather than how many per cent complete some data are. On the other hand, an evaluation of consistency is a bit more complex, since this requires two or more representation schemes in order to be able to compare.

Table 1: DoQ classification by Ballou and Pazer.

DoQ Dimensions	Meaning
Accuracy	The recorded value is in conformity with the actual value
Timeliness	The recorded value is not out of date
Completeness	All values for a certain variable are recorded
Consistency	The representation of the data value is the same in all cases

3.1.2 R.Y. Wang & D.M. Strong TDQM (1996)

Richard Y. Wang & Diane M. Strong propose a consolidation of dimensions under TDQM «Total Data Quality Management» (Wang, 1998) and (Wang and Strong, 1996). This method allowed an identifica-tion of 179 initial criteria and pass from this large number to the classification of fifteen quality criteria, considered as the most important in this study.

Other aspect of this method is that it emanates from a questionnaire handed to users, what seems to be a good start to represent adequately consumers' view of point. Nevertheless, if users have a little or no knowledge in Data Quality, this could lead to a very important number of criteria, which will require a lot of efforts to synthesize the list.

In Figure 1, the authors classified data quality criterions according to four dimensions «intrinsic quality», «accessibility quality», «contextual quality» and «representation quality».

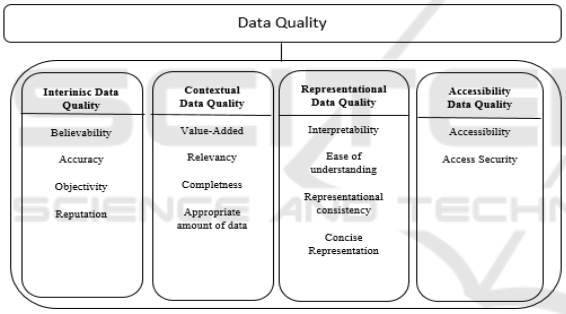

Figure 1: Total Data Quality Model classification.

3.1.3 Redman T.C (1998)

Another perspective on data quality dimensions is provided by Redman in (Redman, 1998) which categorizes data quality issues in terms of: Data view issues, Data value issues, Data presentation issues, and other issues as illustrated below in Table 2.

3.1.4 Levitin and Redman (1998)

Yet another perspective on data quality is data properties, which is provided by Levitin and Redman (Levitin, A. V., & Redman, 1998) who argue that, since data production has many similarities to processes that produce physical products, data production could be viewed as producing data products for data consumers. Thus,

Table 2: Redman.T.C Classification.

Dimensions	Criterions
Data view	Relevancy
	Granularity
	Level of detail
Data value	Accuracy
	Consistency
	Currency
	Completeness
Data presentation	The appropriateness of the format
	ease of interpretation
Others	Privacy
	Security
	Ownership

Levitin and Redman discuss how thirteen basic properties of organizational resources translate into properties for data.

3.1.5 Wekium Model (1999)

Weikum (Weikum, 1999) developed a data quality criteria classification: he distinguishes oriented quality criteria process, system, and data. These three categories may be directly matched with queries processing steps for MBIS (information system based on mediation).

Table 3: QOD criteria classification MBIS.

Category	Criteria
Oriented System	Reliability
	Disponibility
	Integrity
	Security
	Performance
	Auditability
Oriented Process	Security properties
	Liveness properties
Oriented Data	Exactitude
	Exhaustiveness
	Opportunity
	Credibility
	Effectiveness Cost
	Latency

3.1.6 Shanks and Corbitt (1999)

Building on existing literature, the author in (Shanks and Corbitt, 1999) presents a data quality classification, which is based in semiotic theory (the use of symbols for conveying information) and includes both product- and service-oriented aspects of data quality. This is shown in the table below.

Table 4: Data Quality classification by Shanks and Corbitt.

Semiotic level	Meaning	Goal of the representation	
Syntactic	The form of symbols rather than their meaning	Correct and consistent	
Semantic	The meaning of symbols	Complete and accurate at particular points in time	
Pragmatic	The usage of symbols	Useful and usable	
Social	The understanding of the meaning of symbols (different stakeholders)	A shared understanding of the representation is achieved	

3.1.7 Hogan and Wagner (2002)

The authors in (Hogan, W. R., 1997) provide a model for examining data accuracy by assessing correctness and completeness. The approach they use enhances the concordance studies not only by examining data in the clinical record but by prospectively constructing a gold standard so that the patient and care provider can be used as information sources. In essence, this approach goes further than concordance to ensure that the record is a correct representation of the state of the patient.

3.1.8 ISO/IEC 25012 (2008)

ISO standard for data quality, namely ISO/IEC 25012 (Table 1), which defines a general data quality model for data retained in a structured format within a computer system and aims to support the implementation of system's life cycle processes, such as those defined in ISO/IEC 15288 (ISO/IEC., 2008).

This data quality model categorizes quality attributes into fifteen characteristics considering two points of view: Inherent and system dependent. Each characteristic can be considered in a specific context of use. Each characteristic is of equal importance.

Table 5: ISO/IED 25012 classification Model.

Characteristics	**Data Quality**	
	Inherent	**System dependent**
Accuracy	x	
Completeness	x	
Credibility	x	
Currentness	x	
Accessibility	x	x
Compliance	x	x
Confidentiality	x	x
Efficiency	x	x
Precision	x	x
Traceability		x
Understandability		x
Availability		x
Portability		x

As shown in table above, Inherent data quality refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions.

System dependent data quality refers to the degree to which data quality is reached and preserved within a computer system when data is used under specified conditions; this lead us to say that the domain in which data are used and exploited impact strongly the manner of assessing its quality.

3.1.9 Haug et al. (2009)

The authors define three data quality categories: intrinsic, accessibility and usefulness. The authors argue that “representational data quality” as mentioned above in the TDQM (Wang, 1998) can be

perceived as a form of “accessibility data quality” instead of a category of its own. Otherwise, the authors also argue for the intrinsic dimensions defined by (Haug and Arlbjørn, 2011).

- **Discussion**

The mentioned classifications above, were undertaken with different goals in mind and according to different context. Most projects have avoided the difficult issue of quality assessment or have only touched it briefly. This leads us to clearly deduce that the area of data quality is widespread debate and may differ from an application domain to another, hence the interest to enroll in a specific context.

- **Our Motivation**

The quantity of data handled by information systems (IS) is increasing worldwide. Particularly in the case of Mediation systems, where each data source is independent and change frequently which may affect the quality of data and consequently to the satisfaction of the end user.

We choose to study and assess the quality of data retrieved from the virtual data integration or so-called 'the mediation’ also because we noticed that despite the importance of this domain and the importance of the end user satisfaction. Until now, no study has conducted a quality assessment of data retrieved by this type of system. This motivation will make a good starting point as a problematic for our work around.

In the next section, we take first the concept of mediation up in order to have an idea about our context of our research.

4 MEDIATION SYSTEM

In what follow, we present in the first Sub Section the concept of mediation in general and in the second sub section, we illustrates the process of mediation.

4.1 The Concept of Mediation

The mediation concept dating back to 70’s, Levy defines mediation as a transparent access service to a huge number of autonomous, heterogeneous, dynamics and distributed information sources (Levy A. Y et al , 1996). Also, (A.Zellou, 2008) defines mediation systems in as an intermediate tool between a user or application, and a set of information sources; this tool provides a transparent access to sources by a unique interface and query language.

We define the mediator more precisely as a system that offers a common query interface to a set of heterogeneous data sources which can (1) Accepts the participation of different data sources (2) Contains information about the contents of the data sources; (3) Integrates the different data sources by means of a unifying, global or mediated schema; (4) Receives queries from users that are expressed in the language of the global schema; (5) Collects data from sources upon request; at query time;(6) In order to answer global queries, it sends appropriate queries to the sources; (7) Combines the answers received from the sources to build up the final answer to the user.

Mediation information has several advantages. It provides a uniform, unique and multi-source access through a unique interface. It produces an integrated answer by exploiting relations between sources. It offers an independence between applications users and information sources in order to permit evolution of applications and take into account sources autonomy (INRIA, 2001). For us, an ideal mediation system would sbe useful anytime and anywhere.

Intuitively, a Mediator cannot directly assess queries which is directed to it, because it doesn’t contain any form of data (Hadi, Zellou and Bounabat, 2013). These data are stored in a distributed way and in independent sources. Mediator has only an abstract view of data stored in these sources (Zellou, 2008).

Different integration systems based on mediators have been proposed in literature. We find mainly: MOMIS (Beneventano and Bergamaschi, 2004), TSIMMIS (Garcia-Molina et al.1994), HERMES (S. & Brink A. & Emery R. & Lu J. J. & Rajput A. & Rogers T. J. & Ross R. & Ward C. .,1995), Information Manifold (Weld, 1997), Internet Softbot (Etzioni and Weld, 1994), Infomaster (Genesereth M. R. et al. , 1997), OBSERVER (Mena et al. , 1996) , PICSEL (Rousset et al.,2002) and WASSIT (Zellou, 2008).

These systems above mentioned, were distinguishable by the manner of mapping between the global schema and the schema of data sources. Abstraction of all this variety, we choose to present hereunder, the processing of mediation as a conventional global architecture of mediation with three layers.

5 OUR CONTRIBUTION: THE CLASSIFICATION MODEL FOR THE MEDIATION SYSTEM-MDQM

In previous work (Mimouni et al., 2015) we have initially listed exhaustively all of criteria that assess the quality of data. Also, we have proposed a clear and simple definition for each criterion as a recap chart. This attempt leads us to classify, discuss it and thereafter to deduce the most impacting quality criteria concerning the data retrieved from Mediation systems in this paper as a new contribution, seeing that it is

Mediation systems in this paper as a new contribution, seeing that it is our field of research. Consequently, these criterions were sort by some dimensions as a classification according to this domain.

The principle quality key is that end users define quality. End-users typically provide subjective opinions such as 'I need the data to be right, and I need to be able extract it ‘etc., Those Subjective opinions must be translated into objective criteria on the data.

5.1 Mediation Data Quality Model MDQM

In the similar vein, the purpose of this sub section is to create a classification model for the mediation system by classifying data quality criteria which impacts principally mediation system according to six dimensions (Accessibility to data; Manipulation of data, Representation of data, Value-added of data, Usability of data, data sources) as shown below (Table 6).

5.2 Semantic Categorization of MDQM Dimensions

In this sub section, we attempt to classify a previous dimension that we have suggest previously (in sub section A) grouped by their semantic meaning and we propose our own Model of Data quality classification model called “WWHL”, according to the four categorization which are;

- **What:** The content of this data and its usefulness to the users.
- **HOW:** The manner and the profile allow to access this data.

Table 6: MDQM Classification.

Criteria	Dimensions					
	Accessibility to DATA	Manipulation of DATA	Representation of DATA	Value-added of DATA	Usability of DATA	DATA SOURCES
Confidentiality	x					
Security	x					
Portability		x				
Relevancy	x					
Flexibility		x				
Traceability	x					
Availability						x
Completeness				x		
Recoverability				x		
Consistency			x			
Concise Representation			x			
Consistent Representation			x			
Exactitude				x		
Precision				x		
Accuracy				x		
Currency				x		
Validity					x	
Verifiability		x				
Credibility						x
Reputation						x
Objectivity						x
Attractiveness						x
Readability					x	
Efficiency/ Effectiveness				x		
Interpretability					x	
Amount of DATA			x			
Documentation						x
Organization			x			
Specialization				x		
Novelty	x					
Reliability					x	
Interactive		x				
Freshness				x		

- **WHERE:** Refers to data sources
- **LIKE:** In what form and what could we do with this data.

This proposed model is illustrated as shown below;

Table 7: The semantic categorization of dimensions.

Categorization	Dimension	Criteria
WHAT	Usability of DATA	Validity
		Understanbility
		Interpretability
		Readability
		Reliability
		Applicability
	Value-added of DATA	Completeness
		Currency
		Exactitude
		Freshness
		Specialization
		Precision
		Correctness
		Accuracy
		Recoverability
		Efficiency/ Effectiveness
HOW	Accessibility to DATA	Confidentiality
		Security
		Traceability
		Relevancy
		Novelty
WHERE	DATA SOURCES	Availability
		Credibility
		Attractiveness
		Objectivity
		Reputation
		Documentation
LIKE	Representation of DATA	Consistency
		Consistent Representation
		Concise Representation
		Amount of DATA
		Organization
	Manipulation of DATA	Portability
		Flexibility
		Verifiability
		Interactive

6 CONCLUSIONS

This paper has discussed the data quality in the context of virtual data integration which is mediation systems and offered an essay of modeling the classification data quality criterions in the same context. Given the importance attributed to quality factors in our research, it is obvious that proceeding to a selection of those most relevant is extremely important. Especially, as the literature contains a very large number of criteria of which the majority one is described and quoted in the previous section of this present paper.

These factors remain an essential asset to proceed to a qualitative data analysis. Thus, the selection would be a start vision of any process associated to the selection of Data Quality in an information mediation system. As prospects of this work, we propose a study which is going to allow measuring the impact and the importance of every criterion of the data quality with regard to the mediation.

REFERENCES

A. Zellou, D. C. (2008) A solution of XQuery queries in LAV approach in a mediation system.", *CARI. EMI, Rabat, Morrocco.*

Ballou, D. P. and Pazer, H. L. (1985) 'Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems', Management Science. *INFORMS, 31(2), pp. 150–162*.

Beneventano, D. and Bergamaschi, S. (2004) 'The MOMIS Methodology for Integrating Heterogeneous Data Sources', in Building the Information Society. *Boston, MA: Springer US, pp. 19–24*.

Etzioni, O. and Weld, D. (1994) 'A softbot-based interface to the Internet', *Communications of the ACM, 37(7), pp. 72–76*.

Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V. and Widom, J. (1994) 'The TSIMMIS Approach to Mediation: Data Models and Languages 1'.

Genesereth M. R. & Keller A. M. & Duschka O. M. (1997) Infomaster: an information integration system. In *Proceedings of SIGMOD 97. New-York*.

Hadi, W., Zellou, A. and Bounabat, B. (2013) 'Fulvis: New approach for selecting views to materialize in hybrid information integration', in *2013 5th International Conference on Computer Science and Information Technology. IEEE, pp. 248–255*.

Haug, A. and Arlbjørn, J. S. (2011) 'Barriers to master data quality', *Journal of Enterprise Information Management, 24(3), pp. 288–303*.

Hogan, W. R., & W. M. M. (1997) 'Accuracy of data in computer-based patient records.', *Journal of the American Medical Informatics Association, 4(5), p. 342–355*.

INRIA-Rocquencourt, Projet CARAVEL: Système de Médiation d'Information. Rapport d'activité, thème 3A. (2001).

ISO/IEC. (2008) 'ISO/IEC 25012: 2008, Software Engineering–Software Product Quality Requirements and Evaluation (SQuaRE)–Data Quality Model.'

Levy A. Y. & Rajaraman A. & Ordille J (1996) 'Query answering algorithms for information agents. In proceedings of', the *13th National Conference on Artificial Intelligence (AAAI-96)., p. 40–47*.

Levitin, A. V., & Redman, T. C. (1998) 'Data as a resource: properties, implications, and prescriptions.', *MIT Sloan, Management, p. 40(1), 89*.

Mena, E. (1996) 'OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies'.

Mimouni, L., Zellou, A. and Idri, A. (2015) 'Quality of Data in mediation systems', in *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA). IEEE, pp. 1–5*.

Naumann, F. and Rolker, C. (2002) 'Quality-Driven Query Answering for Integrated Information Systems. *LNCS 2261, Springer'*.

Pipino, L. L., Lee, Y. W., Wang, R. Y. and Yang, R. Y. *(2002) Data Quality Assessment*.

Redman, T. C. (1996) 'Data quality for the information age', Artech House computer science library, (Artech House. Boston.).

Redman, T. C. (1998) 'The impact of poor data quality on the typical enterprise.', *Communications of the ACM, p. 41(2), 79-82*.

Redman, T. C. (2001) Data quality: the field guide. *Digital press*.

Rousset, M.-C., Bidault, A., Froidevaux, C., Gagliardi, H., Goasdoué, F., Reynaud, C. and Safar, B. (2002) 'Construction de médiateurs pour intégrer des sources d'information multiples ethétéro enes : le projet PICSEL'.

S. & Brink A. & Emery R. & Lu J. J. & Rajput A. & Rogers T. J. & Ross R. & Ward C. (1995) 'Subrahmanian V.S. & Adali HERMES: A heterogeneous reasoning and mediator system. *Technical Report.', Univ. of Maryland*.

Shanks, G. and Corbitt, B. (1999) Understanding Data Quality: Social and Cultural Aspects.

Strong, D. M., Lee, Y. W. and Wang, R. Y. (1997) '10 potholes in the road to information quality', *Computer, 30(8), pp. 38–46.*

Wand, Y. and Wang, R. Y. (1996) 'Anchoring data quality dimensions in ontological foundations', *Communications of the ACM, 39(11), pp. 86–95*.

Wang, R. Y. (1998) Quality Management A Product Perspective on Total Data, *Communications of the ACM*.

Wang, R. Y. and Strong, D. M. (1996) 'Beyond Accuracy: What Data Quality Means to Data Consumers', Journal of Management Information Systems. *Routledge, 12(4), pp. 5–33*.

Weikum, G. (1999) Towards Guaranteed Quality and Dependability of Information Services.

Weld, F. M. (1997) 'Efficiently executing information gathering plans.', in In *15th International Joint Conference on Artificial Intelligence. Nagoya. Japan., p. 785–791*.

Wiederhold, G. (1992) 'Mediators in the architecture of future information systems', *Computer, 25(3), pp. 38–49.*

Zellou, A. (2008) Contribution to the LAV rewriting in the context of WASSIT, a resources integration framework (Doctoral dissertation, Ph. D. dissertation).