Data quality: A survey of data quality dimensions

Article · August 2013

DOI: 10.1109/InfRKM.2012.6204995

Fatimah Sidi
Universiti Putra Malaysia
79 PUBLICATIONS **460** CITATIONS
SEE PROFILE

Payam Hassany Shariat Panahy
University of Texas Health Science Center at Houston
13 PUBLICATIONS **186** CITATIONS
SEE PROFILE

Lilly Suriani Affendey
Universiti Putra Malaysia
75 PUBLICATIONS **580** CITATIONS
SEE PROFILE

Marzanah A. Jabar
Universiti Putra Malaysia
63 PUBLICATIONS **417** CITATIONS
SEE PROFILE

Data Quality:A Survey of Data Quality Dimensions

[1]Fatimah Sidi, [2]Payam Hassany Shariat Panahy, [1]Lilly Suriani Affendey, [1]Marzanah A. Jabar, [1]Hamidah Ibrahim,
, [1]Aida Mustapha

Faculty of Computer Science and Information Technology
University Putra Malaysia
[1]{fatimacd, suriani, marzanah, hamidah, aida} @ fsktm.upm.edu.my
[2]Payam_shp49@yahoo.com

***Abstract*— Nowadays, activities and decisions making in an organization is based on data and information obtained from data analysis, which provides various services for constructing reliable and accurate process. As data are significant resources in all organizations the quality of data is critical for managers and operating processes to identify related performance issues. Moreover, high quality data can increase opportunity for achieving top services in an organization. However, identifying various aspects of data quality from definition, dimensions, types, strategies, techniques are essential to equip methods and processes for improving data. This paper focuses on systematic review of data quality dimensions in order to use at proposed framework which combining data mining and statistical techniques to measure dependencies among dimensions and illustrate how extracting knowledge can increase process quality.**

Keywords-Data Quality,Data Quality Dimensions,Types of Data,

I. INTRODUCTION

In order to support organisation's activity we should design the process activity appropriately since this involves data. Data is the primary foundation in operational, tactical and decisions making activities. As data are crucial resources in all organizations, business and governmental application, the quality of data is critical for managers and operating processes to identify related performance issues [1], [2], [3]. There are variety of data quality issues from definition, measurement, analysis and improvement which are essential for ensuring high data quality [4]. As the various research shows if the process's quality as well as information's inputs are not controlled, after a while, the degradation of the data quality will be obvious [2].

For improving process's quality with enhanced efficiency in production and administration, using process design is necessary for automation and management technology. It is well known that both of them present, services to business and individual user quickly and consistency [5]. Despite availability of large variety of techniques for accessing and improving data quality such as business rules, record linkage and similarity measures, due to rise difficulty and multiplicity of using these system, data quality methodology has defined and provided [2].

Data quality can provide various services for an organisation as well as nowadays, high quality data can increase opportunity to achieve top services in an organization. Likewise, lack of data quality in organizations can be multiply in the Cooperative Information System (CIS). In fact, (SIC) is an information system with capability to distribute and share general objective between interconnect different systems among of various independent organization in different geographical area as the data is basic recourses for it [6].

Some researchers have identified and tested some effecting factors on data quality inside an organization with collecting data from survey and interview with senior manager and their results show that management responsibilities such as *commitment improving data quality continually , effective communication among stakeholder* and understanding of data quality are significant elements for influencing data quality in an organization [1].

The rest of this paper discusses on data quality strategies and techniques, types of data, data quality definitions, data quality problems classification and data quality dimensions to provide, fundamental issues in this field.

II. DATA QUALITY STRATEGIES AND TECHNIQUES

There are two types of strategies that one adapted for improving data quality namely data-driven and process-driven, and each strategy employs various techniques [2]. However, improving the quality of data is the aim of each technique.

A. Data-Driven

Data-driven is strategy for improving the quality of data by modifying the data value directly. Some related improvement techniques of data-driven are: *acquisition of new data, standardization or normalization, error localization and correction, record linkage, data and schema integration, source trustworthiness,* as well as *cost optimization* [2].

B. Process- Driven

Process-driven is another strategy that redesigns the process which is produced or modified data in order to improve its quality. *Process-driven* strategy consists of two main techniques: *process control* and *process redesign*. In fact, in the process *control* data will be check and manage among the manufacturing process, while in the *process redesign* the causes of low quality will be eliminated and

978-1-4673-1090-1/12/$31.00 ©2012 IEEE

new process will be added in other to producing high quality. Furthermore, adding an activity that can control format of data before storage is another fact in the *process redesign* [2].

However, the advantages of *Process-driven is better performing* than *Data-driven* techniques in long period, because they remove root causes of the quality problems completely. In contrast, *Data-driven* is expensive than *Process-driven* in long period but it is efficient in short period [2].

III. Types of data

Data are real world objects, with ability of storing, retrieving and elaborating through a software process and can communicate via a network [7]. Researchers have provided different classification for data in different area. As implicitly or explicitly, three types of data are described in the field of DQ [2] .Table I presents types of data based on this classification.

A second classification of data is based on considering data as a product, this model classify data in to three types. Table II shows this classification.

TABLE I. TYPE OF DATA SEE DATA AS A IMPLICITY OR EXPLICITY

Types of Data	Definition	Example
Structured data	Generalization or aggregation of items described by elementary attributes defined within a domain	Relational tables Statistical Data
Unstructured data	A generic sequence of symbols, typically coded in natural language.	Body of an Email Questionnaire with free text answering
Semi structure data	Data that have a structure with some degree of flexibility.	Mark up language, XML

TABLE II. TYPE OF DATA SEE DATA AS A PRODUCT

Types of Data	Definition
Raw data items	Smaller data unites which are used to create information and component data items
Component data items	Data is constructed from raw data items and stored temporarily until final product is manufactured
Information products	Data ,which is the consequence of performing manufacturing activity on data

Another classification of data is based on strictness to measure and to achieve data quality, which has two class specifically *elementary data* and *aggregated data*. In an organization, data which managed by operational process and represent atomic phenomena of the real world are called *elementary data*, (*e.g.,* sex, age), While data which are collected from elementary data for applying aggregation function, is called aggregated data, (*e.g.,* average income that tax payer paid in a specify city) [7].From point of view, data can be classified in different types based on their usage in variety of field (e.g., network or web).

IV. Data Quality Definitions

Data quality has different definition on different field and period. Researcher and expert made different understanding about data quality. According to quality management data quality is *appropriate for use* or *to meet user needs* or it is *quality of data to meet customer needs* [9].Also, another definition for data quality is *fitness for use*. Indeed, quality of data is critical for improvement process activity as it can be addressed in different field including management, medicine, statistics and computer science. The widespread collection of definition through data quality may give opportunity to better understand the nature of data process.

V. Data quality problems classification

Data quality problem generally can be divided in to two classes that are single-source and multi-source problem. According to some research four categories for data quality are identified which are shown as the following table.

As a result, the goal of classifying data quality problem is illustrating non-standard data and identifying exact application of data for corresponding requirements [10].

TABLE III. DATA QUALITY PROBLEMS CLASSIFICATION

Data quality problem	*Category*	*Definition*
Single -source problem	Schema level	Lack of integrity constraints, poor schema designer Uniqueness constraints Referential integrity
	Instance level	Data entry errors Misspelling Redundancy Duplicates Contradictory values
Multi-source problems	Schema level	Heterogeneous data models and schema design Naming Conflicts
	Instance level	Overlapping contradicting and inconsistence data Inconsistent aggregating Inconsistent timing

VI. DATA QUALITY DIMENSIONS AND DEFINITION

Table IV illustrate some data quality dimensions and their definition from literature. From the research perspective, there is various numbers of dimensions for Information Quality and Data quality. In fact, "*Data Quality*", "*Information System*" and "*accounting and auditing*" are three initial categories for identifying proper DQ dimensions [11]. In the field of Data Quality ,Wang [11] determined four categories that are *Intrinsic DQ, Accessibility DQ, Contextual DQ, Representational DQ* and fifteen dimensions for DQ/IQ (*e.g., objectivity, believability, reputation, value added*). Other researcher recognized extra dimensions for DQ such as data *validation, credibility, traceability, availability* for identifying. In the area of Information Systems, researcher identified different factors such as *reliability, precision, relevancy, usability, and independency*. In the accounting and auditing, researcher explained that *accuracy, timeliness* and *relevance* are three data quality dimensions. In addition, in this area some scholars explained that internal control systems need lowest cost and highest reliability which refers to some dimensions such as *accuracy*, *frequency* and *size of data* [12].

Base on the ISO standard, quality means *the totality of the characteristics of an entity that bear on its ability to satisfy stated and implied needs* [13].

A Data Quality Dimension is a characteristic or part of information for classifying information and data requirements. In fact, it offers a way for measuring and managing data quality as well as information [14].

So, primary step for understanding data quality dimension can help us to improve it. Analyser and developer use dimension and taxonomy of separate data via using data quality tools for creating and manipulating the information in order to improve information and its process.

TABLE IV. TABLE DATA QUALITY DIMENSIONS

Dimension	***Definition***
Timeliness	The extent to which age of the data is appropriated for the task at hand [15]. Timeliness refers only to the delay between a change of a real world state and the resulting modification of the information system state [2, 11]. Timeliness has two components: age and volatility. Age or currency is a measure of how old the information is, based on how long age it was recorded. Volatility is a measure of information instability the frequency of change of the value for an entity attribute [2, 16].
Currency	Currency is the degree to which a datum is up-to-date. A datum value is up-to-date if it is correct is spite of possible discrepancies caused by time-related changes to the correct value [2, 17]. Currency describes when the information was entered in the sources and/or the data warehouse. Volatility describes the time period for which information is valid in the real world [2, 18].
Types of Data	**Definition**
Consistency	The extent to which data is presented in the same format and compatible with previous data [15]. Refer to the violation of semantic rules defined over the set of data [2].
Accuracy	Data are accurate when data values stored in the database correspond to real-world values [2, 19]. The extent which data is correct, reliable and certified [15]. Accuracy is a measure of the proximity of a data value, v, to some other value, v', that is considered correct [2, 17]. A measure of the correction of the data (which requires an authoritative source of reference to be identified and accessible [14].
Completeness	The ability of an information system to represent every meaningful state of the represented real world system [2, 11]. The extent to which data are of sufficient breadth, depth and scope for the task at hand [15]. The degree to which values are present in a data collection [2, 17]. Percentage of the real-world information entered in the sources and/or the data warehouse [2, 18]. Information having all having all required parts of an entity's information present [2, 16]. Ratio between the number of non-null values in a source and the size of the universal relation [2, 20]. All values that are supposed to be collected as per a collection theory [2, 21].
Accessibility	Extent to which information is available, or easily and quickly retrievable [15].
Duplication	A measure of unwanted duplication existing within or across systems for a particular field, record, or data set [14].
Data specification	A measure of the existence, completeness, quality and documentation of data standards, data models, business rules .meta data and reference data [14].
Presentation Quality	A measure of how information is presented to and collected from does how utilize it. Format and appearance support appropriate use of information [14].
Consistent Representation	To extend to which data is presented in the same format [22].
Reputation	To extent to which information is highly regarded in terms of source or content [15].
Safety	It is the capability of the function to achieve acceptable levels of risk of harm to people, process, property or the environment [13].
Appropriate amount of data	To extend to which data volume of data is appropriate for the task at hand [22].

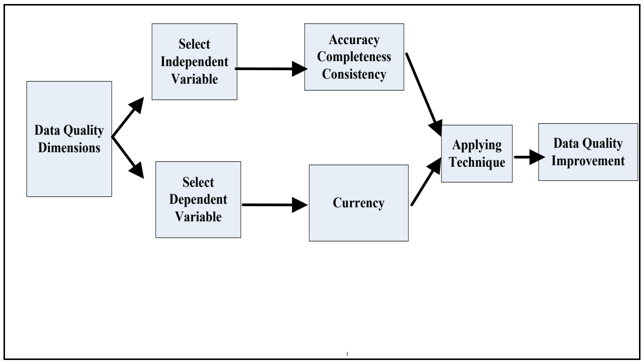
Types of Data	Definition
Security	Extent to which access to information is restricted appropriately to maintain its security [15].
Believability	Extent to which information is regarded as true and credible [15].
Understandability	Extent to which data are clear without ambiguity and easily comprehended [15]. To extend to which data is easily comprehended [22].
Objectively (Objectivity)	Extent to which information is unbiased, unprejudiced and impartial [15].
Relevancy	Extent to which information is applicable and helpful for the task at hand [15].
Effectiveness	It is the capability of the function to enable to users to achieve specified goals with accuracy and completeness in a specified context of use [2].
Interpretability	To extend to which data is appropriate languages, symbols, and units and the definition are the clear [22].
Ease of Manipulation	To extend to which data is easy to manipulate and apply to different same format [22].
Free-of –error	To extend to which data is correct and reliable [22].
Ease of Use and maintainability	A measure of the degree to which data can be accessed and used and the degree to which data can be updated, maintained, and managed [14].
Useability	To extent to which information is clear and easily used [23].
Reliability	Extent to which information is correct and reliable [15]. It is the capability of the function to maintain a specified level of performance when used on specified condition [13].
Amount of data	To extent to which the quantity or volume of available data is appropriate [15].
Freshness	Freshness represents a family of quality factors which each one representing some freshness aspect and having on its metrics [24].
Value added	To extent to which information is beneficial, provides advantages from its use [15].
Learn ability	It means the capability of the function to enable to user to learn it [13].
Data Decay	A measure of the rate of negative change to data [14].
Concise	Extent to which information is compactly represented without being overwhelming (i.e. brief in presentation, yet complete and to the point) [15].
Consistency and Synchronization	A measure of the equivalence of information used in various data stores, applications, and systems, and the processes for making data equivalent [14].
Data integrity fundamentals	A measure of the existence, validity, structure, content, and other basic characteristics of the data [14].
Navigation	Extent to which data are easily found and linked to [23].
Useful	Extent to which information is applicable and helpful for the task at hand [15].
Efficiency	Extent to which data are able to quickly meet the information needs for the task at hand [15].
Availability	Extent to which information is physically accessible [23].
Data Coverage	A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest [14].
Transactability	A measure of the degree to which data will produce the desired business transaction or outcome [14].
Timeliness and Availability	A measure of the degree to which data are current and available for use as specified and in the time frame in which they are expected [14].

VII. DISCUSSION

Most people think the quality of data is depended only to its accuracy and they do not consider and analyze other significant dimensions for achieving higher quality. Indeed, quality of data is more than considering one dimension so, the issue of dimensions' dependencies is essential to improve process quality in different domain and applications. Nevertheless, without knowing the existing relations between data quality dimensions, knowledge discovery cannot be effective and comprehensive for decision making process. From previous work found out, not only dimensions can be strongly related to each other but also, data quality can be supported via the effective dependencies [25].In fact, select appropriate dimensions with identifying correlation among them can create high quality data. In order to discover dependencies among more commonly referenced dimensions consist of accuracy, currency, consistency and completeness, we proposed framework which combining data mining and statistical techniques to measure dependencies among dimensions and illustrate how extracting knowledge can increase process quality. So, based on our hypothesis if there is a correlation between completeness, consistency and accuracy dimensions which are considered independent variable and then, consider currency correlation as dependent variable among them, improvement in data quality will be happened. Also, cause of some difficulties on currency dimension the policy is required.

Fig.1 illustrate proposed framework for evaluating the effect of independent dimensions on dependent dimensions.

FIG.1 FRAMEWORK



Data Quality Dimensions
Select Independent Variable
Select Dependent Variable
Accuracy Completeness Consistency
Currency
Applying Technique
Data Quality Improvement

So, the aim of the proposed framework is discovering the dependency structure for the assessed data quality dimensions.

VIII. CONCLUSION

From the perspective research, many scholars have identified various methodology and framework for assessing and improving data quality through different techniques and strategies on the data quality dimensions [2].They illustrated definitions for dimensions and identified more important data quality dimensions [2], [11], [12], [22]. Existing survey identified forty data quality dimension since 1985 till 2009. Since, some dimensions such as timeliness, currency, accuracy and completeness are more referenced than others, the result of this survey will be used to find correlations among data quality dimension based on proposed framework with combining data mining and statistical techniques for measuring dependencies among them and illustrate how process quality will be increased via the extracting knowledge. Specifically, our future work would be to evaluate dependency among mentioned data quality dimensions for improving process quality.

REFERENCES

[1] S. W. Tee, P.L. Bowen, P. Doyle, F.H. Rohde, "Factors influencing organizations to improve data quality in their information systems," Accounting & Finance, vol. 47, pp. 335-355, 2007.

[2] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, "Methodologies for data quality assessment and improvement," ACM Computing Surveys (CSUR), vol. 41, p. 16, 2009.

[3] W. Eckerson, "Data Warehousing Special Report: Data quality and the bottom line," Applications Development Trends May, 2002.

[4] Y.Y.R. Wang, R.Y. Wang, M. Ziad, Y.W. Lee, Data quality vol. 23: Springer, 2001.

[5] F. Casati, M.C. Shan, M. Sayal, "Investigating business processes," ed: Google Patents, 2009.

[6] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, C. Batini, "Managing data quality in cooperative information systems," *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE,* pp. 486-502, 2002.

[7] C. Batini and M. Scannapieca, Data quality: Concepts, methodologies and techniques: Springer-Verlag New York Inc, 2006.

[8] V. Peralta, "Data quality evaluation in data integration systems," Université de Versailles (chair) Raúl RUGGIA Professor, Universidad de la República, Uruguay, 2008.

[9] F. G. Alizamini, M.M. Pedram, M. Alishahi, K. Badie, "Data quality improvement using fuzzy association rules," 2010, pp. V1-468-V1-472.

[10] Y. Man, L. Wei, H. Gang, G. Juntao, "A noval data quality controlling and assessing model based on rules," 2010, pp. 29-32.

[11] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," Communications of the ACM, vol. 39, pp. 86-95, 1996.

[12] KQ. Wang, SR. Tong, L. Roucoules, B. Eynard, "Analysis of data quality and information quality problems in digital manufacturing," 2008, pp. 439-443.

[13] M. Heravizadeh, J. Mendling, M. Rosemann, "Dimensions of business processes quality (QoBP)," 2009, pp. 80-91.

[14] D. McGilvray, Executing data quality projects: Ten steps to quality data and trusted information: Morgan Kaufmann, 2008.

[15] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," Journal of management information systems, vol. 12, pp. 5-33, 1996.

[16] M. Bovee, R.P. Srivastava, B. Mak, "A conceptual framework and belief-function approach to assessing overall information quality," International journal of intelligent systems, vol. 18, pp. 51-74, 2003.

[17] T. C. Redman, Data quality for the information age: Artech House, 1996.

[18] M. Jarke, Fundamentals of data warehouses: Springer Verlag, 2003.

[19] D. P. Ballou and H. L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," Management science, pp. 150-162, 1985.

[20] F. Naumann, Quality-driven query answering for integrated information systems vol. 2261: Springer Verlag, 2002.

[21] L. Liu and L. Chi, "Evolutionary data quality," 2002.

[22] L. L. Pipino, Y.W. Lee, R.Y. Wang, "Data quality assessment," Communications of the ACM, vol. 45, pp. 211-218, 2002.

[23] S. Knight and J. Burn, "Developing a framework for assessing information quality on the world wide web," Informing Science: International Journal of an Emerging Transdiscipline, vol. 8, pp. 159-172, 2005.

[24] V. Peralta, "Data quality evaluation in data integration systems," Université de Versailles (chair) Raúl RUGGIA Professor, Universidad de la República, Uruguay, 2008.

[25] D. Barone, *et al.*, "Dependency discovery in data quality," 2010, pp. 53-67.