Journal of Theoretical and Applied Information Technology
30th June 2017. Vol.95. No 12
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A REVIEW OF DATA QUALITY RESEARCH IN ACHIEVING HIGH DATA QUALITY WITHIN ORGANIZATION

[1]M. IZHAM JAYA, [2]FATIMAH SIDI, [3]ISKANDAR ISHAK, [4]LILLY SURIANI AFFENDEY, [5]MARZANAH A. JABAR

[1,2,3,4] Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia

[5] Department of Information System and Software Engineering, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia

Corresponding Author: fatimah@upm.edu.my

ABSTRACT

The aim of this review is to highlight issues in data quality research and to discuss potential research opportunity to achieve high data quality within an organization. The review adopted systematic literature review method based on research articles published in journals and conference proceedings. We developed a review strategy based on specific themes such as current research area in data quality, critical dimensions in data quality, data quality management model and methodologies and data quality assessment methods. Based on the review strategy, we select relevant research articles, extract and synthesis the information to answer our research questions. The review highlights the advancement of data quality research to resemble its real world application and discuss the available gap for future research. Research area such as organizations management, data quality impact towards the organization and database related technical solutions for data quality dominated the early years of data quality research. However, since the Internet is now taking place as the new information source, the emerging of new research areas such as data quality assessment for web and big data is inevitable. This review also identifies and discusses critical data quality dimensions in organization such as data completeness, consistency, accuracy and timeliness. We also compare and highlight gaps in data quality management model and methodologies. Existing model and methodologies capabilities are restricted to the structured data type and limit its ability to assess data quality in web and big data. Finally, we uncover available methods in data quality assessment and highlight its limitation for future research. This review is important to highlight and analyse limitation of existing data quality research related to the recent needs in data quality such as unstructured data type and big data.

Keywords: *Data Quality, Data Quality Management Model, Assessment Methods, Database, Organization*

1. INTRODUCTION

Achieving high data quality has become an important element in managing data within an organization. Possessing high data quality could help an organization to formulate better business strategy and unveil business pattern for decision making. Failure in providing high data quality to the organization have brought various issues such as false decision due to incorrect data, high cost of operation and lack of customer satisfaction [1]. Moreover, the increasing numbers of data available today with unknown quality levels added challenges to optimally analyze and make use of data that are relevant to the organization.

High data quality has been defined as a data that is fit for use and able to meet the purposed set by data user [1]–[4]. This definition clearly suggested that quality of data is highly dependent to the context of data usage and synergies to the customer needs, ability to use and ability to access data. Thus, in data quality assessment and improvement process, participation of data users and other data stakeholders that are involve during data entry, data processing and data analysis is important. Various methods have been proposed to assess data quality from the context of data users and other data stakeholders including using survey, and questionnaire.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Researchers adopted surveys and questionnaire to collect requirements and expectations from data user, data entry personnel and other data stakeholder. Adoption of surveys and questionnaires is important in order to define data quality attributes and data quality dimensions to achieve high data quality within its context. Statistical methods such as correlation analysis is then used to identify correlation between attributes and classify the attributes into data quality dimensions for example data completeness, data consistency, data accuracy and data timeliness. On the other hand, findings of the analysis help researcher to identify cause of data quality problems and subsequently, suggestion on data quality improvement can be made.

A lot of progress has been established in data quality research which are not only limited to the adoption of surveys and questionnaires as mentioned before. Thus, we urge to answer questions regarding progress in data quality research through a review of data quality research articles that has been published before this. Our main intention in doing this review is to highlight potential issues in data quality research and to discuss potential unfilled research gap in data quality research especially in managing data quality within the organization. Furthermore, this review is intended to facilitate data quality implementation within the organization by discussing the strengths and weaknesses of existing data quality management model and data quality assessment methods.

The discussion provided in this review is limited only to the existing proposals in data quality research that can be directly implemented in any organization with specific purposes such as business. In doing this study, we excluded research articles that proposed measurement metrics for data quality and data quality framework without proper assessment technique. By doing this, we narrow our discussion only on research articles that proposed data quality solution that is suitable for direct implementation in organization.

Indeed, we aim to seek answers for several specific research questions:

RQ.1: What areas of research have been covered in data quality research?

RQ.2: What data quality dimensions are critically needed to be managed in data quality research?

RQ.3: How to systematically manage data quality in the organization?

RQ.4: How are the assessments of data quality being conducted in the organization?

We begin with the big landscape of data quality research as this helps readers to understand the wide areas in data quality. In any organization, data quality improvement is not only limited to the data available in databases but also included data from other areas such as websites and big data technologies. In the current landscape of data quality, data resources are varied and data quality implementation is bound to the type of data available in each data resources. Moreover, for each data resources, identification, definition and measurement of data quality dimensions plays an important role to ensuring data quality success within the organization. In RQ.2, we direct readers' attention to the identification of data quality dimensions and discuss several potential issues in critical data quality dimensions. We limit our discussion only to data quality dimensions that are mostly mentioned in literatures as critical for data quality success. Discussion on the critical data quality dimensions provides strong substance for readers to understand our discussion in RQ.3 and RQ.4. As our review focus on the data quality implementation within the organization, we further discuss ways to systematically manage data quality within the organization. The discussion includes the strengths and weaknesses of existing data quality models and methodology. As data quality management model and methodology required methods to assess data quality, we continue our discussion in RQ.4 with more attention given to the limitation of existing data quality assessment methods.

To the best of our knowledge, comparisons of available data quality management model, methodologies and assessment method specifically for the implementation within the organization has not been undertaken in previous analysis of literature. Lately, organizations are facing new challenges in managing data quality resulted from new technologies such as big data and open data. Huge data in organization can be collected from various resources and stored in different data types such as structured, unstructured and semi-structured [5], [6]. There is a need to revisit existing research proposals in data quality and thoroughly discuss their limitation in handling various types of data.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This paper is organized as follows. In Section 2, we describe the methodology used for this review. Then, in Section 3 we discuss the answers to our research questions. Section 3.1 discusses the available research areas in data quality. We explain critical data quality dimensions, namely data completeness, data consistency, data accuracy and data timeliness in Section 3.2. We then discuss data quality management models and methodologies in Section 3.3. In Section 3.4, we discuss existing data quality assessment methods with further elaboration on assessment methods being used in data quality research. Lastly, we conclude the findings and highlight research gap for future research in Section 4.

2. RESEARCH METHODOLOGY

This paper involves systematic literature review [7] of data quality research in organizations. Based on the research questions, we identified specific themes which are related to data quality research area, data quality dimensions, data quality management and data quality assessment methods. Then, we construct our review strategy as a guideline during research articles selection, information extraction and information synthesis in order to answer the research questions. We limit our review to the research articles published in journals and conference proceedings.

3. DISCUSSION

3.1 Areas of Research in Data Quality

Success of data quality implementation is determined by seven critical factors including management responsibilities, operation and assurance cost, research and development, production, distribution, personnel management and legal function [8]. Based on the critical factors mentioned earlier, a framework of critical areas in data quality has been proposed. The proposed framework highlighted areas where more attention should be given by researchers in achieving high data quality. However, the framework is too complex even though it was well described. Improvisation and simplification of the framework has been proposed later by characterizing data quality research into several research topics and methods. The improvised framework suggested that data quality research evolved within several research topics such as data quality impact, database related technical solutions for data quality, data quality in context of computer science and information technology and also data quality in curation [9]. However, the authors do not include data quality in the context of online users including social media and websites in the suggested framework. During our review, we found several works in data quality that focus on quality assessment of the web information. For example, the assessment of data quality in web using Support Vector Regression (SVR) method [10] and the measurement of data quality dimensions in online newspaper content [11]. This indicates the advancement of data quality research to support the expansion of computation era. In social media context, data quality plays an important role to influence online users decision [12]. As organization todays utilize social media for marketing and building trusted relationship with targeted customers, improving data quality is essential.

Recent work suggested that data quality research can be divided into four subgroups including data quality for websites, data quality for decision support, data quality assessment and other data quality application such as software development process and medical data quality [13]. In contrast to the earlier suggested framework, data quality research area expanded to the new area such as data from various data sources including website [14] and unstructured data in big data [15].

During the review, we also found that data quality research has expanded to support recent big data technology. For example, a research has been conducted to analyse the impact of data size increment towards data quality problems [16]. As this research proved that large volume of data does not guarantee high data quality state, it gives opportunities for researchers to explore new methods and approaches to manage and improve data quality in big data environment with high data volume, high data velocity and high data variety characteristic.

The advancement in data quality research areas as we discussed in this paper resembled data quality in its real-world application. Thus, synergy between research progress in data quality and the real-world application is essential in order to deliver high quality outcome. For that reason, technical solution of data quality provided in the computer science field should strongly compliment the information system field and their needs to improve the organization [17]. Future research should be able to fill the gap between the research done in both fields with more concern should be given to the least

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

explored area such as unstructured data type from web and social media.

3.2 Critical Dimensions of Data Quality

Data quality researches have been conducted in many organization to identify data quality dimensions and its related attributes. However, as data quality dimensions are dependable to the context of usage, no agreement on standard set of dimensions that contribute to high data quality can be defined. Instead, previous research suggested that data quality dimensions can be described using multi-dimensional concept with each dimension has dependency to others in order to achieve high data quality. For example, a study has been conducted to identify relationships between data accuracy, data consistency, data completeness and data timeliness [18]. The constructed framework derived from this study can further be used to enhance quality in information system.

The ontological characteristic of data quality dimensions is important to provide better understanding of data quality as data quality dimensions is related to each other. Data quality dimensions such as accuracy, reliability, timeliness, completeness and consistency dimensions can be classified into internal and external view [19]. Each of these classifications can be divided into data-related and system-related dimensions. Alternatively, data quality dimensions can also be classified into four categories namely, intrinsic, contextual, representational and accessibility [20]. A study has been done using these data quality categories to examine data quality problems in data rich organization. As a result, it confirmed that categorization of data quality dimensions into intrinsic, contextual, representational and accessibility is valid [1].

Table 1: Context of Previous Study and Methods Used

Reference	Context of Study	Methods
[20]	Data consumers	Two-stage survey with exploratory factor analysis.
[19]	Information System Development	Analysis of representation mapping, comprehensive literature review.
[21]	Health	Model generated based on Financial Accounting Standards Board (FASB) with examples from medical domain. Evidential network is then developed and assessed.
[22]	Data producers, data custodians and data consumers.	Questionnaire and assigned dimensions into suggested category if more than 50% respondent agreed.

Data quality dimensions such as accuracy, completeness, consistency and existence is related to a group of integrity attributes [21]. Integrity attributes can be described as the intrinsic dimension of data quality and related to the ability of data to map to the data user interest [21]. In comparisons to representational consistency [20], consistency in integrity attribute has been defined from the data value perspective and not just the format or representation of the data itself. The availability of data value, no-redundancy of data and fabricated data value have been described as existence of data [21]. Integrity attribute also suggested that believability and credibility dimensions shall be rename as integrity due to the small differences between them.

Table 1 explained the context of study and methods used by each research being reviewed in this paper. We mentioned the context of these studies and methods used in order to help reader to have a clear view on how the researcher identified and defined data quality dimensions such as data completeness, consistency, accuracy and timeliness. In this review, we focus on the dimensions stated above as these dimensions being mentioned as critical data quality dimensions [19], [23]. From our review, survey method has been widely used to confirm the dimensions and reflected data quality from its context. Further explanation on data completeness, data consistency, data accuracy and data timeliness will be given in the following subsections.

3.2.1 Data completeness

Data is complete when all necessary value pertaining to the data exist [19], [21], [22], [24].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

However, data should also be able to represent null values as in some cases, data may have no value associated [21]. For example, employee with unmarried status would leave the spouse name field blank. In this case, the null value in spouse name cannot be considered as incomplete data. It is important to understand that data may carry null value and the existence of null value in which it is appropriate should not be consider as incomplete data. Incomplete data occur when null value assigned to data that should has value. This suggests that data quality assessment process should be able to identify the cause of null value found in dataset before data completeness can be assessed.

Recent data application make used multiple data sources to gain holistic view of application purpose. These data sources may contain the same data value but with different criticality level. Criticality level in data reflected the awareness within the organization towards data changes and null values. Data with high criticality level will be reviewed cautiously after data entry and prompt rectification process will be made if data contained null values. In contrast, data with low criticality level is not cautiously handled. Due to this problem, some data may not contain value as data is intentionally left or being truncated during data entry and data production. This gap may lead to data incompleteness problem as least important data was not given full attention [25]. The best way to solve this problem is to improve the process during data entry by including data verification process. Instead, further research can be done to improve data completeness by doing a cross reference to other sources of data such as online database and knowledge bank.

3.2.2 Data consistency

Consistent data reflects a state which the same data represent and the same value with standard representation are used throughout the system [19], [21], [22]. Data representation such as currency unit, month and year should be represented in the same manner as long as it represents the same value. However, data inconsistency problem still can occur even if the proactive action has been taken. Such example is when the currency unit for a country has been changed. The ability to make update to the entire system will help to improve data consistency.

In data integration, data are being integrated from multiple sources to uncover information and unveil insight. However, as data sources possibly can have different schema and naming convention, inconsistency issue after the integration are expected. Data variation from multiple data sources being integrated lead to the tendency that data after the integration is not consistent [25]. With regards to the latest development in big data, and the variety of data being integrated, we suggest that consistency issues should be managed in the early stage of the integration by defining data standards and data policies within the organization.

3.2.3 Data accuracy

Precision and free-of-error data are the main characteristic of data accuracy [19], [21], [22]. In order to justified the precision and the accuracy of data, comparisons towards real-world data can be made [24]. However, to make this definition more clearer, data accuracy dimension has been categorized into semantic accuracy and syntactic accuracy [26], [27]. Syntactic accuracy refers to the closeness of the value towards the element of corresponding definition domain, whereas, semantic accuracy refers to the closeness of the value towards the real world value. In data quality research, the value of data in real-world is hard to be known as it required more knowledge about the data. Without this knowledge, we cannot measure the semantic accuracy.

Data accuracy problem can happen due to many reasons. Inaccurate data can happen during data entry and due to systematic error in data production [25]. For example, untrained data entry personnel may accidentally altered data value during data entry due to insufficient experience. In this case, proper training and redesigning data entry process by including expert verification would help to improve data accuracy in the organization.

3.2.4 Data timeliness

Data timeliness referred to the age of data [20]. Conversely, data timeliness can also be defined as the attribute of datedness [21]. Datedness attribute included age and volatility as a measure of data timeliness. On the other hand, both research strongly agreed that timeliness and datedness should be measured by user in the context of application purposes. Data timeliness is very important as the most current data has more potential to be consider as high data quality [20], [28]. However, we did not agree that data timeliness can be used as a measure of high data quality as it does not expose the relevancy of that data. This is supported by other researcher that discussed data relevancy from the perspective of data timeliness [21].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Large collection of data increased the potential to discover knowledge. Large data collection can be gathered trough database integration, pool of machines and also web information. However, distributed heterogeneous data sources as the example given and the vast volume of data may lead to the timeliness issue [25]. This is due to the long time needed to access requested data within large data collection and the complexity of process to interpret data with heterogeneous format. We believe, with appropriate methods in hand, such limitations can be minimized. For example, the segregation of inactive data in a separate database allows database optimization and better performances [29].

We discussed data quality dimensions in this paper as a foundation to the data quality management models and methodologies which will be discussed in the next section. Solid understanding on data quality dimensions helped us to elaborate further in data quality management models and methodologies.

3.3 Systematic Management of Data Quality in Organization

Managing data quality dimensions and improving these dimensions through a systematic process are important to ensure high data quality within the organization. For this reason, various researches have been done to propose a model and methodologies for systematic data quality management. A Total Data Quality Management (TDQM) [4] has been proposed earlier to support the concept of 'data as a product'. With this concept, high data quality can be achieved by replicating physical production of high quality product. TDQM extended Total Quality Management (TQM) framework which used in physical production. The methodologies start with the definition of information product (IP). At this stage, the IP has its own characteristics and requirements in order to achieve high quality state. Then, the information quality (IQ) metrics is developed and used to measure the IP. The measurement result is then analysed using statistical process control, pattern recognition and pareto chart. Lastly, improvement being made on the IP using Information Manufacturing Analysis Matrix based on the analysis done before. The availability of various tools for result analysis as mentioned before helps organization to implement TDQM. However, several arguments occur when comparing data production to physical production. These included the ability of data to be shared among user. Secondly, raw data may not arrive in time when needed and it is difficult to assign several quality dimensions such as believability to physical production. TDQM was designed for managing data quality in databases and current technologies including big data may limit the usage. This is due to the variety of data types available in big data. Future work is possible to redesign the framework by including other data sources in big data.

Information Integrity Methodology (IIM) [30] has been proposed later and outlined the needs to achieve information integrity by addressing the foundation of data itself. Information integrity considered data ability to meet organizations strategic objectives. However, to achieve high quality of data, a framework of information integrity should be fulfil. The framework included data policy, organization capabilities, data administration, architecture, process, validation, communication and framework compliance. On the other hand, the proposed methodologies added another phase in data quality management which is to reassure data quality after the improvement process being made.

AIM Quality (AIMQ) model [31] comprises of Product and Service Performance Model for Information Quality (PSP/IQ), IQA instrument to measure information quality and information quality gap analysis technique to improve the information quality. Questionnaire is used in this model to assess information quality. Further statistical analysis is then being used to identify information quality problem area. The used of PSP/IQ is due to the objective of achieving high quality information guided by the dimensions attributes namely, intrinsic, representational, contextual and accessibility [32].

Another example of data quality management model is Data Quality Management Maturity Model (DQMMM). The foundation of this model is to improve data structure quality and as the end result it would provide high quality of data [33]. In this model, structure of integrated databases being managed by standardizing its metadata. Standardization of database metadata can be divided into several stages such as logical, physical and mapping metadata information. Other data quality management model and methodologies mentioned before does not manage data quality during the integration of various databases across

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the organization. This model stressed the needs of data integration in order to enhance data accuracy and consistency. Furthermore, its ability to ensure high quality of data during database integration will be an added value.

Many researches done in data quality focus on structured data type compared to other types of data such as semi-structured and unstructured data. However, in this review, we found several models that are suitable for either structured, semi structured or unstructured data type. One of them is the Complete Data Quality Management (CDQM). CDQM suggest theoretical, empirical and intuitive approach to assess data quality [26]. It comprised of three stages; state reconstruction, assessment and choice of optimal improvement process. The advantage of CDQM is the flexibility of the methodology to support structured, unstructured and semi-structured type of data. However, there is no defined measurement method or calculation to measure data quality dimensions in CDQM. Thus, the implementation of CDQM in the organization is difficult. We summarize the strengths and weaknesses of the data quality model and methodology in Table 2 based on characteristics found in the literature.

As described in each of data quality management model, data quality level need to be measure and assess before further analysis can be done. Following section will discussed data quality assessment method found during this review.

Table 2: Strengths and Weaknesses of Data Quality Model and Methodology

Model/ Methodology	Strengths	Weaknesses	Data Type
TDQM	Various choice of tool to analyse data quality such as statistical process control, pattern recogni-tion and pareto chart.	Data can be shared among user whereas raw material assigned to a single product. Timeliness – raw material arrived at time. Believa-bility – difficult to compare with physical products.	Structured
IIM	Reassura-nce phase helps organiza-tion to revaluate data quality after appro-priate data quality improve-ment process.	IIM required data quality policy creation and fulfilment. Thus, it takes more effort for the organiza-tion to create data quality policy.	Structured
AIMQ	Measure data quality dimen-sions in the attributes of intrinsic, represent-tational, contex-tual and accessibi-lity.	Limited tool to identify informa-tion quality problem areas.	Structured
DQMMM	Manage data quality during database integra-tion process.	Suitable only for relational database.	Structured
CDQM	Support structu-red, unstruc-tured and semi-struc-tured data type.	Unspecific. No data quality dimensions measure-ment and calculations defined in CDQM.	Structur-ed, unstruc-tured and semi-structur-ed

© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.4 Assessment of Data Quality in Organization

Several methods have been used to measure and assess data quality in previous research. For example, a data quality modelling approach during database designs to assess data quality. This research suggested the integration of data quality aspects into database design by providing quality schema generated from data quality modelling [34]. We believe that the integration will reduce cost for outlier's detection and quality improvement after database creation. However, this approach was only suitable for structured data type, especially data in a relational database.

Other methods which can be used to assess data quality are Cell Level Tagging [28], Subjective and Objective Data Quality Assessment [35], Control Matrices [36] and Semantic Reconciliation [37]. For better understanding, we explained strengths and weaknesses of each of the data quality assessment methods found during this review in Table 3. Several methods such as subjective and objective data quality assessment, control matrices and semantic reconciliation are suitable for various data types. However, in control matrices methods, the quality assessment is depended to the availability of Information Product Manager (IPM). Thus, it is not suitable for data with high volume and high velocity characteristic. On the other hand, cell level tagging is more promising choice to assess data quality as it integrated data quality requirements in database design. However, as it was design only for relational structured data type, much works is needed to extend its usage in unstructured data type.

Table 3: Comparisons on Data Quality Assessment Methods

Methods	Strengths	Weaknesses
Cell Level Tagging	Data quality being integrated into database design which to ensure high data quality being stored. This will reduces costs in scanning outliers and data quality improvement.	Approach taken in this research is only suitable for relational data. However, further research can be done to extend the method into big data area.
Subjective and Objective Data Quality Assessment	Data either structured, semi structured or unstructured can be assess in subjective and objective form. The comparisons between both gives broader view on data quality issues. This research also suggested the usage of functional form based on dimension to be measured.	Comparisons of subjective and objective result in the data quality metrics only identified data quality problems in broad meaning. Outliers' detection and quality improvement are still depending on the data stakeholders' skills and ability.
Control Matrices	Support structured, semi structured and unstructured data.	Highly depends on the availability of appointed Information Product Manager (IPM) to assess quality.
Semantic Reconciliation	Support data heterogeneity.	Small numbers of semantic heterogeneity defined in this research limit its performance.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. OPEN RESEARCH ISSUES

The advancement in data quality compliments the applications of its solution into the real world. We have seen the evolution since the early years of data quality research with much work concentrated to support the structured data type in relational database. Years ahead, we expect more consideration will be given to the unstructured data types and big data technology. Some of the models and methodologies being discussed in this paper have paved the way towards that.

Data quality dimensions are still entangled to the context of usage. As what we point out earlier, discussions and research in data quality dimensions was still in the context of well-structured data. Based on our review, some data quality dimensions have been discussed as critical in achieving high data quality; particularly, data completeness, data consistency, data accuracy and data timeliness. In the context of big data technology, the measurement and assessment methods of these critical dimensions could be different as big data provide massive volume of data with high data velocity and high data variety characteristic. We have discussed in details about the dimensions and put forward our suggestions on future research possibilities within the paragraph.

Data quality management models, methodologies and data quality assessment methods are the essential deliverables in data quality research. Additionally, data quality can be managed systematically in the organization by adopting suitable data quality management model and data quality assessment methods. We have discussed several available models, methodologies and assessment methods in this review and yet more research can be done to fill in the gap highlighted in this review such as the quality assessment of unstructured data type. Moreover, as we are now in the era of big data technology, models, methodologies and data quality assessment methods that can support unstructured data are crucially needed.

5. CONCLUSION

In this review, all the questions related to data quality research areas, critical data quality dimensions, systematic data quality management and data quality assessment methods have been successfully answered. As new technology become available, data in organizations is no longer limited to what are stored in the database. Various data sources such as website and social media have become important to organizations in marketing and building relationship with their targeted customers. In this review, we discussed on the expansions of research area in data quality to support the needs in new technology era. We also identified four critical data quality dimensions including data completeness, data consistency, data accuracy and data timeliness. These critical data quality dimensions are important and should be given highest priority in managing data quality within the organization. In order to systematically manage data quality within the organization, existing data quality management model and methodologies such as TDQM, IIM and AIMQ can be adopted. Adopting existing data quality management model and methodologies may require enhancement to support various data sources especially in unstructured data. Data quality assessment is one of the important activities in systematic management of data quality. In this review, we compared strengths and weaknesses of available methods for data quality assessment within the organization.

Consequently, this review suggests future consideration for data quality research in organization as an open research issues. Future research direction in organization context should put more effort in managing data quality of unstructured data from website, social media and big data.

6. ACKNOWLEDGEMENT

The work reported here is funded by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS 03-12-10-999FR). This support is gratefully acknowledge.

REFRENCES:

[1] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data Quality in Context," *Communications of the ACM*, vol. 40, no. 5, pp. 103–110, May 1997.

[2] Y. W. Lee and D. M. Strong, "Knowing-Why About Data Processes and Data Quality," *Journal of Management Information Systems*, vol. 20, no. 3, pp. 13–39, Jan. 2003.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] A. V. Levitin and T. C. Redman, "Data as a resource: properties, implications, and prescriptions," *Sloan Management Review*, vol. 40, pp. 89–101, 1998.

[4] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, no. 2, pp. 58–65, Feb. 1998.

[5] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 134–142, May 2016.

[6] S. Sadiq and M. Indulska, "Open data: Quality over quantity," *International Journal of Information Management*, vol. 37, no. 3, pp. 150–154, Jun. 2017.

[7] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009.

[8] R. Y. Wang, V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, no. 4, pp. 623–640, 1995.

[9] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and Framework for Data and Information Quality Research," *Journal of Data and Information Quality*, vol. 1, no. 1, pp. 1–22, Jun. 2009.

[10] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Automatic Assessment of Document Quality in Web Collaborative Digital Libraries," *Journal of Data and Information Quality*, vol. 2, no. 3, pp. 1–30, Dec. 2011.

[11] S. B. Rajakumari, "Data Quality Mining in Electronic News Paper," *Indian Journal of Science and Technology*, vol. 7(S5), no. June, pp. 47–50, 2014.

[12] J. V. Chen, B. Su, and A. E. Widjaja, "Facebook C2C social commerce: A study of online impulse buying," *Decision Support Systems*, vol. 83, pp. 57–69, Mar. 2016.

[13] Y. Xiao, L. Y. Y. Lu, J. S. Liu, and Z. Zhou, "Knowledge diffusion path analysis of data quality literature: A main path analysis," *Journal of Informetrics*, vol. 8, no. 3, pp. 594–605, Jul. 2014.

[14] A. Marotta and A. Delgado, "Data Quality Management in Web Warehouses using BPM," in *ICIQ 2016*, 2016, p. 18:1-18:10.

[15] M. Helfert and Mouzhi Ge, "Big Data Quality - Towards an Explanation Model in a Smart City Context," in *ICIQ 2016*, 2016, p. 2:1-2:8.

[16] P. Woodall, A. Borek, J. Gao, M. Oberhofer, and Andy Koronios, "An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics," in *Proceedings of the 19th International Conference on Information Quality (ICIQ-2014)*, 2014, pp. 24–33.

[17] S. Sadiq, N. Yeganeh, and M. Indulska, "20 years of data quality research: themes, trends and synergies," in *Proceedings of the Twenty-Second Australasian Database Conference*, 2011, vol. 115, pp. 153–162.

[18] P. Hassany, S. Panahy, F. Sidi, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "A Framework to Construct Data Quality Dimensions Relationships," *Indian Journal of Science and Technology*, vol. 6, no. 5, pp. 4421–4431, 2013.

[19] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, vol. 39, no. 11, pp. 86–95, Nov. 1996.

[20] R. Wang and D. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

[21] M. Bovee, R. P. Srivastava, and B. Mak, "A conceptual framework and belief-function approach to assessing overall information quality," *International Journal of Intelligent Systems*, vol. 18, no. 1, pp. 51–74, Jan. 2003.

[22] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: product and service performance," *Communications of the ACM*, vol. 45, no. 4ve, pp. 184–192, Apr. 2002.

[23] D. P. Ballou and H. L. Pazer, "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems," *Management Science*, vol. 31, pp. 150–163, 1985.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[24] V. Jayawardene, S. Sadiq, and M. Indulska, "An Analysis of Data Quality Dimensions," *ITEE Technical Report No. 2013-01*, vol. 1, pp. 1–32, 2013.

[25] D. M. Strong, Y. W. Lee, and R. Y. Wang, "10 potholes in the road to information quality," *Computer*, vol. 30, no. 8, pp. 38–46, 1997.

[26] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–52, Jul. 2009.

[27] C. Batini and M. Scannapieca, *Data Quality*. Springer Berlin Heidelberg, 2006.

[28] R. Y. Wang, M. P. Reddy, and H. B. Kon, "Toward quality data: An attribute-based approach," *Decision Support Systems*, vol. 13, no. 3–4, pp. 349–372, Mar. 1995.

[29] S. Gi Lee, B. Lee, and H. Jeong, "A Study on the Problem Analysis and Improvement Plan of the Data Quality Management System of National R&D Data," *Indian Journal of Science and Technology*, vol. 8, no. 23, Sep. 2015.

[30] S. Ruschka-Taylor, "Transforming enterprise information integrity," 2004.

[31] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: a methodology for information quality assessment," *Information & Management*, vol. 40, no. 2, pp. 133–146, Dec. 2002.

[32] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," in *2012 International Conference on Information Retrieval & Knowledge Management*, 2012, pp. 300–304.

[33] K. S. Ryu, J. S. Park, and J. H. Park, "A Data Quality Management Maturity Model," *ETRI Journal*, vol. 28, no. 2, pp. 191–204, Apr. 2006.

[34] R. Y. Wang, H. B. Kon, and S. E. Madnick, "Data Quality Requirements Analysis and Modeling," in *Ninth International Conference on Data Engineering*, 1993, no. April.

[35] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, p. 211, Apr. 2002.

[36] E. Pierce, "Assessing data quality with control matrices," *Communications of the ACM*, vol. 47, no. 2, pp. 82–86, 2004.

[37] S. Madnick and H. Zhu, "Improving data quality through effective use of data semantics," *Data & Knowledge Engineering*, vol. 59, no. 2, pp. 460–475, Nov. 2006.