CONCEPTUALIZING BIG DATA QUALITY FRAMEWORK FROM A SYSTEMATIC LITERATURE REVIEW PERSPECTIVE

***Mohamad Taha Ijab*[1]*, Ely Salwana Mat Surin*[2]*, Norshita Mat Nayan*[3]**

[1, 2, 3]Institute of Visual Informatics, National University of Malaysia, Malaysia

E-mail: taha@ukm.edu.my[1]; elysalwana@ukm.edu.my[2]; norshitaivi@ukm.edu.my[3]

DOI: https://doi.org/10.22452/mjcs.sp2019no1.2

ABSTRACT

In its effort to modernize its public service delivery, Malaysia is actively leveraging on big data analytics. To support this ambitious initiative, a new framework addressing the needs of Big Data Quality for Malaysia is imperative, as big data analytics requires high quality data in order for it to be useful. Unfortunately, a proper big data quality framework, particularly one which focuses on the specific context and needs of Malaysia's Public Sector Open Data initiative is missing. This paper thus focuses on the proposed development of the Big Data Quality Framework for Malaysia's Public Sector Open Data Initiative (MyPS-ODI). Using Systematic Literature Review (SLR) approach, we conceptualize and propose a framework for big data quality that can contribute in enabling the sharing of quality data widely, so as to increase the transparency of the Malaysian government's services, and provide people and the business community the opportunity to increase creativity and innovation in developing new products and services through high quality data. The proposed framework will benefit IS managers in government sectors and help them better understand and meet their consumers' data quality needs, as well as help them to facilitate big data analytics readiness of the public sector in the country. It will also assist in providing a high-quality platform to the citizens to get quality information from official government sources. Finally, the framework will help in saving the time and effort needed in correcting the results of data analysis due to poor data quality, by providing quality data from the data preparation stage.

Keywords: Big Data, Open Data, Data Quality, Data Quality Framework, Systematic Literature Review

1.0 INTRODUCTION

Malaysia was ranked 53rd in the Open Data Barometer in 2016 [1]. The Open Data Barometer (ODB) index is produced by the World Wide Web Foundation, and is a collaborative work of the Open Data for Development (OD4D) global partnership which aims to measure the prevalence and impact of open data initiatives around the world. The ODB is participated by 115 countries globally and the index measures the countries' open data readiness, implementation, and impacts. Based on the ODB data from 2016, Malaysia was ranked much lower than its neighboring ASEAN countries such as the Philippines, which was in 22nd position, Singapore, in 23rd position, and Indonesia, in 38th position. The reasons for this relatively poor ranking are that some data champions in the ministries and government agencies still do not share their data, or the data is not in machine-readable format. Other reasons include the unavailability of timely data, and also the fact that the data is not updated, is not free, data is not openly licensed, and finally, data is not accurate.

Hence, this research is timely and we posit that it is able to help Malaysia's long-term strategy in enhancing its open and big data quality strategy, and also in improving its future position in the Open Data Barometer rankings. For this, Malaysian Administrative Modernization and Management Planning Unit (MAMPU) has been given the task by the Malaysian Government to champion the modernization of public service delivery by leveraging big data analytics on the platform called Open Data portal. In lieu of that, this paper will focus on the conceptual development of a Big Data Quality Framework for Malaysia's Public Sector Open Data Initiative run by MAMPU. It is also aimed that the proposed framework is generic enough such that it can be applied by other countries intending to adopt a similar big data quality approach into their respective open data initiatives.

MAMPU in its strategic document called "Public Sector Open Data Analytics – Strategy, Challenges, Direction", mentioned that data readiness and data quality is one of the six critical success factors for the success of Malaysia's Public Sector Open Data initiative [2]. MAMPU also listed the ten principles of open data for the country's Public Sector Open Data. These ten open data principles are completeness, primary source, timeliness, accessibility, machine readable, non-discriminatory, use of Open Standard, permanent, licensing, and usage costs [2]. In addition to that, there is also available the quality rating of data sets based on the works of Tim Berners-Lee in 2006 [3] adopted by MAMPU. Among others, the 5-star open dataset should comply with all of these requirements: (i) data should be available on the Web, and in any format, provided the data has an open license; (ii) data should be available as machine-readable structured data (e.g., in Excel format instead of PDF or image scan); (iii) data should be available in a non-proprietary format (e.g., CSV instead of Excel) that is readable by various software; (iv) data should make use of open standards from W3C (RDF and SPARQL) and URIs to identify and link data directly to its source; and (v) the data should be linked to other providers' data to provide context and to display real-time analysis [3].

This paper is divided into five sections and organized as follows. The next section covers the previous studies on big data and data quality. In the subsequent section, the methodology used (i.e., the Systematic Literature Review) is discussed, and the framework development for the proposed big data quality framework for My-PSODI is then elaborated. It is then followed by a discussion on the proposed big data quality framework and lastly wrapped up by a conclusion.

2.0 PREVIOUS STUDIES

We are surrounded by data in our daily life. It is a major part of our life and an important element for enabling businesses and organizational processes [4], [5]. The data quality, which represents the degree to which the data characteristics fulfill certain and specific requirements have a significant impact on the businesses, the companies [6], or even in human lives. Strong [7], Wang [8] and Wand and Wang [9] argue that data quality is basically defined as usefulness and usability of the data - data that is fit for use by data consumers. Other researchers including Crosby [10] defines data quality as "conformance to requirements". The following presents discussion from literature on the similarities or differences of big data, open data and government data concepts.

2.1 The Concepts of Big Data, Open Data, and Government Data

Fundamentally, big data is linkable information that has large data volumes and complex data structures [11]. According to Gartner [12], big data is characterized as high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insights and decision making as well as in facilitating process automation. In order to mine big data, it requires the capability of extracting valuable and quality information. It is made more difficult due to the fact that we are mining data from large datasets, and with a variety of types of structured, semi-structured and non-structured data [13]. Open data is defined by Gurin [14] as accessible public data that individuals and organizations can use to develop new ventures, discover patterns or trends, make data-driven decisions, and answer complex problems. Open data includes two basic features: the data must be publicly available for anyone to use, and it must be licensed in a way that allows for its reuse. Open data should also be relatively easy to use, although there are certain gradations of openness of such data. There is general agreement that open data should be available free of charge or at minimal cost. Government data, on the other hand is usually data created and held by government agencies for public's consumption and therefore, this also makes such government data a subset of open data.

Although the concepts of big data and open data are related, however, they are not similar. Open data can make big data more useful and more democratic. Applying open data principles to big data can help solve some of the difficult issues that big data has raised [15]. The problem now is not only that government agencies and some businesses are collecting personal data; it is also that individuals do not know what is being collected and do not have access to the information about themselves. The combination of both big data and open data are compelling as they can be leveraged to transform business, government, and society. Big data gives consumers ability and power to understand, analyze, and ultimately change the world we live in if big data analytics are done properly. In the same vein, open data ensures that power will be shared and used for more democratic purposes.

2.2 Data Quality Challenges

Detail studies analyzing and researching on data quality methods for big data and open data are still lacking. Over the years, various researchers have been investigating into the data quality issue. As evidenced in the following anecdotal reviews, researchers generally still do not have any consensus on what constitutes data quality or even how to properly define data quality in a consensual manner. We consider that this situation has indeed become a critical theoretical-and-practical gap that could hamper further understandings of big data quality. In summary, big data quality faces the following challenges:

- The variety of data sources brings numerous data types and complex data structures and hence, increases the difficulty of data integration. Before this, enterprises only used the data generated from their own business systems, such as sales and inventory data. But now, the data collected and analyzed by enterprises have outdone this scope. Big data sources are very wide, including (i) data sets from the Internet and mobile Internet [16] (ii) data from the Internet of Things; (iii) data collected by various area of industries; (iv) experimental and observational data [17], such as physics experimental data, biological data, and space observation data. These sources produce many data types. One data type is unstructured data (example: documents, video, audio, etc). Second type of data is semi-structured data, including software packages/modules, spreadsheets, and financial reports. The third is structured data. The quantity of unstructured data occupies more than 80% of the total amount of data in existence. For enterprises, obtaining big data with complex structures from different sources and effectively integrating them are a daunting task [18]. There are conflicts and contradictory situations among data from different sources. For small data volumes, the data can be checked by a manual search or by programming tools to search the data, even by ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform). However, these methods are arguably useless when processing PB-level even EB-level data volumes.

- Data volume is tremendous and therefore, it is difficult to judge data quality within a reasonable amount of time. After the industrial revolution, the amount of information dominated by characters doubled every ten years, but after 1970, the amount of information doubled every three years. Today, the global amount of information can double every 2 years. It is difficult to collect, clean, integrate, and obtain the necessary high-quality data within a reasonable time frame. Furthermore, the proportion of unstructured data in big data is very high and it would take a lot of time to transform unstructured types into structured types and process the data.

- Data changes very fast and the timeliness of data is very short, and data needs higher requirements for processing technology. Due to the timeliness and changes in big data, the usefulness period of some data is very short. If organizations cannot collect the required data in real time they may obtain outdated and invalid information. Processing and analysis based on this data would produce useless or misleading conclusions, and also would lead to wrong decision-making by governments or enterprises.

- Due to lack of standards of approved data quality, research on the quality of big data has been initiated. In order to guarantee the product quality, in 1987 the International Organization for Standardization (ISO) published ISO 9000 standards. Nowadays, there are more than 100 countries and regions all over the world actively carrying out these standards. This implementation endorses mutual understanding among enterprises in domestic and international trade and brings the benefit of removing barriers. The study of data quality standards began in the 1990s, but not until 2011 did ISO published ISO 8000 data quality standards [19]. The standards need to be mature and at the same time, research on open data and big data quality has just begun.

2.3 Poor Data Quality Consequences

The basic rule of data quality is “garbage-in, garbage-out,” or GIGO where accurate results cannot be expected based on inaccurate data. This rule specifically applies in the context of big data [20]. As huge volume of data is generated, and a large variety of heterogeneous data produced, the quality of data is highly questionable. Previous studies indicated that poor quality big data is rampant in large databases and on the Internet, which causes resource wastage, poor service efficiency and significant costs in repairing the data. This has caused huge losses to organizations [9], [21]–[24]. Therefore, the importance of veracity and value of big data is increasingly being acknowledged. Studies estimate that erroneous data costs businesses in the USA almost USD600 billion dollars annually [25]. It is also found that in some organizations, data error rate is approximately between 1% to 5%, and in some more major cases, above 30% [26]. Poor data quality requires tedious data cleaning processes which involve discovering rules, detecting/checking for inconsistencies, and data repairing. All these corrective activities can cost organizations about 30% - 80% of the development time and budget. Thus, it is important that data quality is managed, ensured and verified right from the earliest stage of data preparation.

3.0 METHODS

This study uses Systematic Literature Reviews (SLR) to develop the proposed framework for big data quality. The SLR is a systematic, explicit, comprehensive, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars, and practitioners [27]. We posit that the SLR as well as meta-analyses of appropriate studies can be the best form of evidence available to identify suitable elements for the framework.

The advantage of SLR is that it is a well-defined methodology. SLR also can provide information about the effects of some phenomenon across a wide range of settings and empirical methods. If studies give consistent results, systematic reviews provide evidence that the topic is robust and transferable. If the studies give inconsistent results, sources of variation can be studied. In the case of quantitative studies, it is possible to combine data using meta analytical techniques. This increases the likelihood of detecting real effects that individual smaller studies are unable to detect. The major disadvantage of SLR is that they require considerably more effort than traditional literature reviews.

Our review method was based on that used by Webster and Watson [27] as per illustrated in Figure 1. This method involves a systematic examination of selected databases using a variety of strategies including keywords and subject headings. It allows the integration of data across studies where they have similar outcome measures and the summary of findings where methods used are diverse.

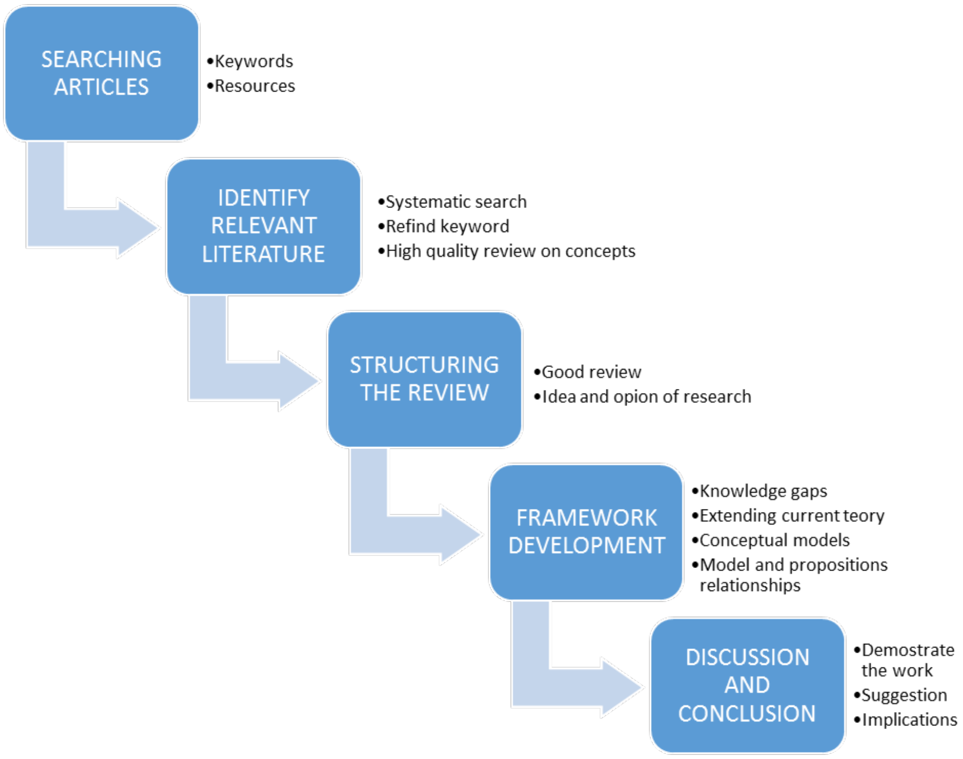

Fig. 1: Systematic Literature Review (Webster & Watson, 2002)

3.1 Searching articles

Following the review method based on Webster and Watson [27], the authors searched the databases ScienceDirect, IEEE Xplore, Scopus and AIS Electronic Library. The following terms were searched *("big data", "open data", "data quality framework" and "government sector")*, with a restriction of time period from Year 2000 to Year 2018. This manual search process led to 295 papers being captured and eventually, sixteen (16) highly relevant papers were analyzed in detail after screening the title, abstract, keywords, and contents of the articles. These sixteen papers were used as the basis for this paper's analysis and discussion. The journal papers were selected because they were known to include either empirical studies or literature surveys, and they had been used as sources for other systematic literature review related to big data. To make an introduction of the paper, the topic needs to motivate. The best way to write the introduction is providing a working definition of the key variable(s), and clearly articulate the paper's contributions. The ways of demonstrating contributions include providing a new theoretical understanding that helps to explain previously confusing results noting that little research has addressed this topic, providing calls from well-respected academics to examine this topic, bringing together previously-disparate streams of work to help shed light on a phenomenon, and suggesting important implications for practice. Searching articles with the relevant topic regarding the title of the paper is very important to make a motivation for readers. In this paper, the keyword used for searching are quality in big data. The elaboration of the keyword will give more on definitions of the key variables and set the boundaries on the topics of work like quality variables, elements involved and other similar topics. Boundaries also include issues like level(s) of analysis, temporal and contextual limitations, the scope of the review, and the implicit values [28], [29]. Other than that, some suggestions and future research should be considered during the article searching process to show how the framework should be implemented in the next phase.

3.2 Identify Relevant Literature

One way to identify relevant literature is by using a structured approach to determine the source material for the review. Journals are the main resources and give the major contributions in identifying the literature. We use leading journals, while journal databases that have a high impact journal citation ranking and quartile scores accelerate identification of relevant articles. Scanning a journal's table of contents is a useful way to pinpoint other items not caught by the keyword sieve. Conference proceedings should also be examined, especially those with a reputation for quality. Based on Robey [30], reviewing the literature has two major streams of research:

1. Go backward by reviewing the citations for the articles identified in – quality in big data, to determine the prior articles that should be considered.

2. Go forward by using the web site - the electronic version of the research Citation Index to identify articles citing the key articles identified in the previous steps. Determine which of these articles should be included in the review.

A systematic search should ensure the accumulation of a relatively complete census of relevant literature in big data. The nearest key word of quality in big data can gauge that the review is completed when we are not finding new concepts in the article set. Some articles will be missed out, but we can identify them by referring to the other researchers who read and write about that particular paper in their research.

3.3 Structuring the Review

A literature review is concept centric. Thus, concepts determine the organization of elements for the big data quality framework. In contrast, some authors take an author-centric approach and essentially present a summary of the relevant articles. However, this method fails to synthesize the literature. For this paper, finding from the literature is synthesized by creating a concept matrix where it describes the unit of analysis for each element involved (refer Table 1).

Data quality is demonstrated through the data quality dimension. Data quality dimension is a set of data quality attributes representing each and every single aspect of data quality. There were fifteen data quality dimensions identified by Wang and Strong [8]. These dimensions include relevancy, timeliness, completeness, believability, accuracy, objectivity, reputation, value-added, appropriate amount of data, interpretability, ease of understanding, representational consistency, concise representation, accessibility, and access security. From the total of 42 studies, we have identified 16 studies which have been recognized as the most suitable studies related to the subject matter. Two researchers, Cai and Zhu [31] and Juddoo [32] believed that quality data consists of all 6 categories. While Batini and Scannapeico [33] stated that Availability, Usability, Concordance, Presentation quality and Correctness of data will determine the quality level of the data. In 2015, Tam and Clarke [34] and Batini [35] stated that quality of data is about Availability, Concordance, Presentation quality and Correctness of data. Zhu and Gauch [36] claimed that availability, usability, presentation quality and correctness are the main factors that will determine the quality of data. This is different from Abdullah [37] perspective; they believed the quality of data is about availability, reliability, concordance and correctness. Others such as Shanks and Corbitt [38] considered that availability, usability and correctness are the main elements of big data quality. Besides, Tate and Alexander [39] argued that quality data is about availability, presentation quality and correctness. Other researchers including Lucas [24], Weiskopf and Wen [40] and Taleb [41] considered availability, concordance and correctness as part of the elements in big data. Wang and Strong [8] suggested that availability and correctness of data are parts of data quality. While Roberson [42] claimed that data quality is about reliability and correctness of the data. Lastly, Radhakrishna [43] and Kwon [44] stated that the quality of data is only dependent on one thing, which is correctness and concordance respectively.

Table 1: *Concept Matrix*

Author	**Concept Categories**					
	Availability	**Usability**	**Reliability**	**Concordance**	**Presentation quality**	**Correctness**
(Wang & Strong, 1996)	√					√
(Shanks & Corbitt, 1999)	√	√				√
(Tate & Alexander, 1999)	√				√	√
(Zhu & Gauch, 2000)	√	√			√	√
(C. Batini & Scannapeico, 2006)	√	√		√	√	√
(Lucas, 2010)	√			√		√
(Radhakrishna et al., 2012)						√
(Roberson, 2013)			√			√
(Weiskopf & Weng, 2013)	√			√		√
(Kwon et al., 2014)				√		
(Abdullah et al., 2015)	√		√	√		√
(Cai & Zhu, 2015)	√	√	√	√	√	√
(Carlo Batini et al., 2015)	√			√	√	√
(Juddoo, 2015)	√	√	√	√	√	√
(Taleb et al., 2015)	√			√		√
(Tam & Clarke, 2015)	√			√	√	√

Based on the literature, six (6) concept categories have been identified as per Table 2 to represent all dimensions that have been studied by data quality researchers. The categories are (i) Availability, (ii) Usability, (iii) Reliability, (iv) Concordance, (v) Presentation, and (vi) Correctness. Now, based on the identified six categories, we managed to identify all the unit of analysis for each category. This is reported in Table 3.

Table 2: Concept Matrix Augmented with Unit of Analysis

Author	**Concepts**																
	Availability			**Usability**		**Reliability**			**Concordance**			**Presentatio n**		**Correctness**			
Unit of analysis	**Accessibility**	**Timeliness**	**Authorization**	**Credibility**	**Clarity**	**Authenticity**	**Integrity**	**Auditability**	**Agreement**	**Consistency**	**Variation**	**Readability**	**Structure**	**Accuracy**	**Errors**	**Misleading**	**Validity**
(Wang & Strong, 1996)	√	√												√			
(Shanks & Corbitt, 1999)	√	√			√											√	√
(Tate & Alexander, 1999)			√									√		√			
(Zhu & Gauch, 2000)	√		√		√												
(C. Batini & Scannapeico, 2006)	√	√			√				√	√	√	√		√	√	√	√
(Lucas, 2010)		√							√	√	√			√			
(Radhakrishna et al., 2012)														√			√
(Roberson, 2013)							√							√			√
(Weiskopf & Weng, 2013)	√	√							√	√	√			√	√	√	√
(Kwon et al., 2014)										√							
(Abdullah et al., 2015)	√	√					√		√	√	√			√			√
(Cai & Zhu, 2015)	√	√		√			√			√		√		√			
(Carlo Batini et al., 2015)	√									√		√		√		√	
(Juddoo, 2015)	√	√		√		√			√	√	√	√	√		√		
(Taleb et al., 2015)	√	√							√	√	√			√			
(Tam & Clarke, 2015)	√	√								√	√	√		√			

A study conducted by Shanks and Corbitt [38] produced an emiotic-based framework for data quality. Their study proposed 4 levels and a total of 11 quality dimensions where the four levels are syntactic, semantic, pragmatic and social. The data dimensions conjectured from the study are well-defined, comprehensive, unambiguous, reputable, meaningful, correct, timely, concise, easily accessed, easily understood, and awareness of bias. Meanwhile, the study by Tate and Alexander [39] discovered six data quality dimensions: authority, accuracy, objectivity, currency, coverage/intended audience, and interaction/transaction features of data. On the other hand, Zhu and Gauch [36] discussed six quality dimensions and they are authority, popularity, currency, availability, information-to-noise ratio, and cohesiveness. In another research, Batini and Scannapeico [33] listed the following data quality dimensions: timeliness, consistency, accuracy, correctness, completeness, currency, volatility, accessibility, objectivity, believability, reputation, value-added, relevancy, and ease of understanding.

Lucas, in 2010 [24], proposed the data quality dimensions of timeliness, consistency, accuracy, completeness, and relevancy. Radhakrishna [43] argued that the main element of data quality is correctness, which covers accuracy and data validity while Roberson [42] claimed that big data quality is about accuracy and data validity. It is also about integrity of data which is about reliability of data. Weiskopf & Weng [40] stated that data quality is very dependent on the accessibility and timeliness, agreement, consistency and variation of data. The quality of data for these researchers was also about accuracy of data, level of errors, misleading data and the validity of data. It can be concluded that these writers wanted to talk about data correctness. In contrast to the other researchers, Kwon [44] only examined the consistency of the data in data quality. Abdullah [37] identified the data quality dimensions completeness, validity, timeliness, of accuracy, integrity, consistency, and accessibility. Cai and Zhu [31] posited that data quality is determined based on many factors. Among the factors are accessibility, timeliness, credibility, integrity, consistency, readability and accuracy. A study by Batini [35] identifed the data quality dimensions of accessibility, consistency, accuracy, completeness, redundancy, readability, and trust. Adding to the literature, Juddoo [32] highlighted the data quality dimensions as appropriate amount of data, believability, accessibility, completeness, ease of manipulation, free-of-error, consistent representation, interpretability, objectivity, timeliness, understandability, relevancy, reputation, security, and value-added.

Taleb [41] argued that data quality has two dimensions, which are intrinsic and contextual dimensions. Intrinsic dimensions include accuracy, timeliness, consistency, and completeness, while the contextual dimensions are reputation, accessibility, value-added, believability, relevancy, and quantity. Finally, The Australian Bureau of Statistics (ABS) Data Quality Framework, according to Tam and Clarke [34] perceived that data quality comprises of seven (7) dimensions, namely, Relevance, Timeliness, Accuracy, Coherence, Interpretability, Accessibility, and Institutional Environment, which reflects a broad and inclusive approach to quality definition and assessments.

4.0 CONCEPTUAL FRAMEWORK DEVELOPMENT AND DISCUSSION

After structuring the review, the process of framework development begins. A review should identify the elements, processes, characteristics and the critical knowledge gaps, and motivate researchers to study the quality in big data. That is, writing a review not only requires an examination of past research, but it also means making a chart for future research. Highlighting the discrepancy between what we know and what we need to know alerts other scholars to opportunities for a key contribution in this paper. The roadmap is accomplished by developing a conceptual framework and using a traditional approach (i.e., the SLR methodology), which is the focus of this paper. The review articles may draw from the variance and process research to develop conceptual frameworks to guide future research. A few conceptual frameworks from previous research in data quality were collected and analyzed to capture relationships between variables. Therefore, justification for these relationships represents a crucial part of the theory-development process in this paper. From that we will know which concepts will give a high impact on what we search for. Some comparisons were done to make the result comprehensive, which also included a theoretical explanation concerning verification, and gave an example from practice.

At the moment, comprehensive studies on analysis and research of data quality standards and quality assessment methods for big data are still limited [31]. Previous studies by many researchers such as [31], [33], [41], [45], [46] argued that data quality is a multidimensional, and multifaceted concept. There is lack of consensus among the various studies on the number of dimensions, and on their definitions or metrics of data quality. In fact, there are around 200 terms that have been identified on data quality elements, and there is lack of agreement in their nature, their definitions or even measures [23]. As a solution to this problem, Taleb [41] proposed that the definition of data quality should be seen as "domain aware" or the definition should be defined by the data owners and data users themselves.

Furthermore, according to Cai and Zhu [31] and Saha and Srivastava [47], data quality usually depends not only on its own features but also on the business environment using the data, such as who produces the data, the processes surrounding the preparation of data, and the data users themselves (i.e., what specific purpose the data is used for). Thus, Lucas [24] argued that the principle of 'the one who provides the data is the one who is responsible for quality' could be applied. Somehow, the issue with this approach is that user's requirements which are considered to be important are not taken into account. This perspective is supported by Cai and Zhu [31], who claimed that data quality standards are regularly developed from the perspective of data producers instead of the data consumers. Therefore, in this paper, we are developing the Big Data Quality Framework by combining the perspectives of the data producers, data drivers, data experts and also the data consumers themselves. With this, the definition, dimension, elements, and measures of data quality will be more comprehensive and holistic.

Based on the synthesis of the papers and adaptation from Cai and Zhu [31], the data quality elements that are proposed for Malaysia's Big Data Quality Framework for My-PSODI are: Accessibility, Timeliness, Authorization, Credibility, Clarity, Authenticity, Integrity, Auditability, Agreement, Consistency, Variation, Readability, Structure, Accuracy, Error, Misleading and Validity. These data elements are then grouped into six data quality dimensions of Availability, Usability, Reliability, Concordance, Presentation, and Correctness, as shown in Figure 2.

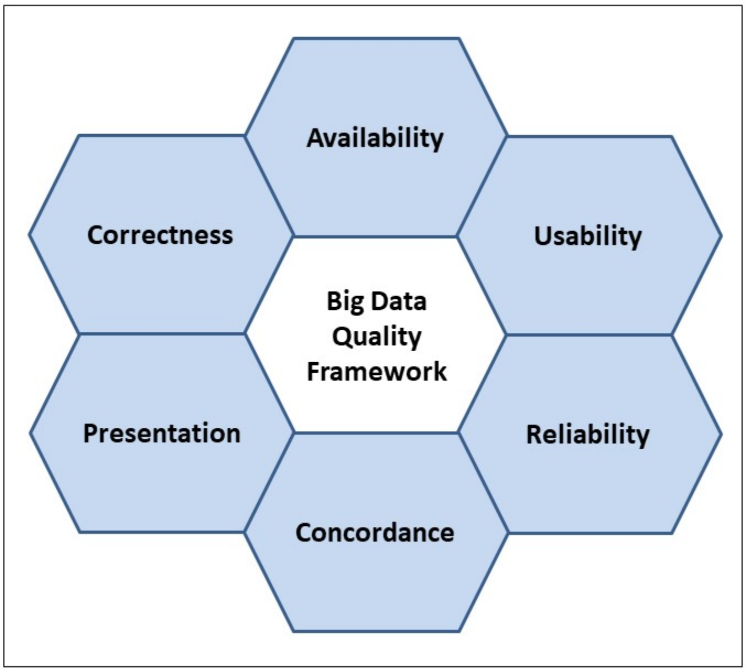

Fig. 2: The Proposed Big Data Quality Framework

In the context of Malaysia's Public Sector Open Data Initiative, hypothetically the framework can be applied by MAMPU in actual practice. For example, the indicator of whether the data is of highest quality or lowest quality will be measured, assessed, compared and given by the data champions and data owners themselves based on a certain metrics tool. Additionally, upon collating all the data, the data champions or data owners would be able to provide a certain ranking on their dataset according to a radar plot that we are also proposing, of which DQ Level 0 will be at the lowest level while DQ Level 6 will be the highest data quality level. For example, in Figure 3, following the data quality audits run by the data champions of several datasets (i.e., say Dataset 1 to Dataset 3), the data champion would state that the Availability of their Dataset 1 is at DQ Level 5, the Usability of their Dataset 1 is at DQ Level 6, the Reliability of their Dataset 1 is at DQ Level 6, the Concordance of their Dataset 1 is at DQ Level 6, the Presentation of their Dataset 1 is at DQ Level 5 and the Correctness of their Dataset 1 is at DQ Level 6. This will easily provide a snapshot of Dataset 1 which indicates that it is of high quality overall. On the other hand, Dataset 2 would be relatively at a lower data quality rank as it scored 3 for Availability, 2 for Usability, 2 for Reliability, 5 for Concordance, 2 for Presentation, and 2 Correctness. Whereas for Dataset 3, the dataset quality would be of moderate data quality as it scored 6 for Availability, 5 for Usability, 4 for Reliability, 4 for Concordance, 5 for Presentation, and 3 Correctness, respectively.

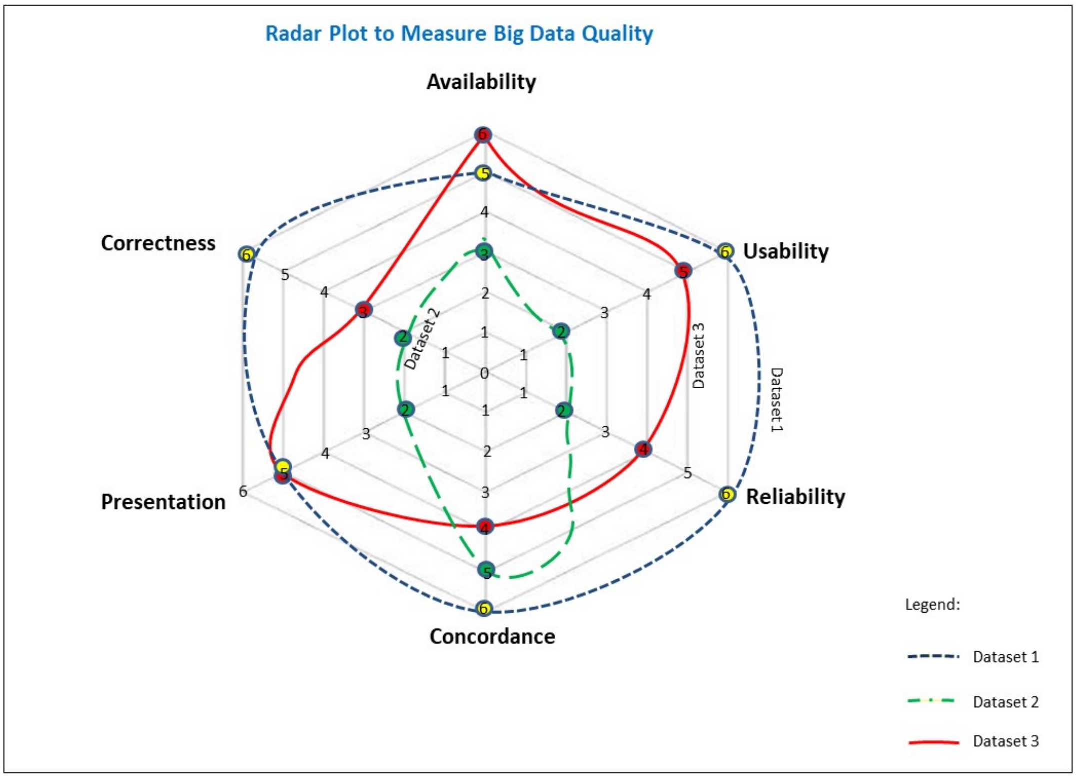

Fig. 3: Radar Plot to Measure the Big Data Quality.

5.0 CONCLUSION

Big Data and Open Data are the current ICT initiatives undertaken by the Malaysian government, and it is fully supported by the Prime Minister's Office of Malaysia. The aim of transforming public service delivery using big data analytics and open data emphasizes on leveraging data to enhance outcomes and to lower costs. One of the ways to do this is by proliferating open data among government ministries and agencies. Other steps include encouraging cross-agency data sharing and leveraging big data analytics (BDA). The Public Sector Open Data initiative allows the creation of new products by the government based on the big data and is aimed to enhance the service delivery and service quality which takes into account the needs of the country's citizens. Employing Systematic Literature Review (SLR) process, this paper proposes the Big Data Quality Framework that includes the data elements of Accessibility, Timeliness, Authorization, Credibility, Clarity, Accuracy, Authenticity, Integrity, Consistency, Completeness, Auditability, Fitness for Use, Readability, and Structure. These data elements are then grouped into six relevant data quality dimensions of Availability, Usability, Reliability, Concordance, Presentation, and Correctness accordingly. Advantage-wise, the framework will help save the time and effort in correcting the results of data analysis due to poor data quality by providing quality data from the data preparation stage. This research is also important as the proposed Big Data Quality Framework will facilitate several other benefits such as better data analytics and meaningful decision making for the data owners and data users, and subsequently help in materializing the country's strategic national vision, namely the 11th Malaysia Plan and the Vision 2020. Another benefit of the proposed framework is that it is generic in nature, making it appropriate to be used or adopted by other countries in their own quests on open data initiatives and big data quality. However, as there is a lack of empirical testing of the proposed framework in real practice, the actual practicality of the framework is not yet demonstrated. Yet, this is one of the potential research undertakings that can be pursued in the future in further enhancing and developing the proposed Dig Data Quality Framework for Malaysia's Public Sector Open Data Initiative.

ACKNOWLEDGEMENT

This project was funded by the National University of Malaysia's research grant KRA-2018-012.

REFERENCES

[1] World Wide Web Foundation, “Open Data Barometer,” 2018.

[2] MAMPU, “Garis Panduan Analitis Data Raya Sektor Awam – Program Kesedaran Dasar dan Garis Panduan ICT Sektor Awam,” 2016. [Online]. Available: http://www.mampu.gov.my/ms/penerbitan-mampu/send/89-program-kesedaran-dasar-dan-garis-panduan-ict-sektor-awam/215-7-taklimat-7-gp-drsa.

[3] T. Berners-lee *et al.*, “Tabulator : Exploring and Analyzing linked data on the Semantic Web,” in *Proceedings of the 3rd international semantic web user interaction workshop*, 2006, p. 159.

[4] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, “A Survey on Data Quality : Classifying Poor Data,” 2015.

[5] F. M. Yunus *et al.*, “Proposed Data Quality Evaluation Method for a Transportation Agency,” *Open Int. J. Informatics*, vol. 5, no. 2, pp. 52–63, 2017.

[6] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz, “Requirements for Data Quality Metrics,” vol. 9, no. 2, 2018.

[7] D. M. Strong, Y. W. Lee, and R. Y. Wang, “Data quality in context,” *Commun. ACM*, vol. 40, no. 5, pp. 103–110, 1997.

[8] R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *J. Manag. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[9] Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations.,” *Commun. ACM*, vol. 39, no. 11, pp. 86–95, 1996.

[10] P. B. Crosby, *Quality is free: The art of making quality certain*. Signet, 1980.

[11] M. J. Khoury and J. P. A. Ioannidis, “Big data meets public health,” *Science (80-.).*, vol. 346, no. 6213, pp. 1054–1055, 2014.

[12] Gartner, “Big Data,” *Gartner*, 2012. [Online]. Available: https://www.gartner.com/it-glossary/big-data/.

[13] H. A. A. Hafez, “Mining Big Data in Telecommunications Industry: Challenges, Techniques, and Revenue Opportunity,” *Int. J. Comput. Electr. Autom. Control Inf. Eng*, vol. 10, no. 1, pp. 183–190, 2016.

[14] J. Gurin, “Big Data and Open Data: What’s What and Why Does It Matter?,” *The Guardian*, 2014. [Online]. Available: https://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government.

[15] S. Saxena, I. Muhammad, and I. Muhammad, “The impact of open government data on accountability and transparency The impact of open government data on accountability and transparency data,” 2018.

[16] L. Jianzhong and L. Xianmin, “An important aspect of big data: data usability,” *J. Comput. Res. Dev.*, vol. 6, no. 006, 2013.

[17] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, “Addressing Big Data Issues in Scientific Data Infrastructure,” in *Procedures of the 2013 International Conference on Collaboration Technologies and Systems*, 2013, pp. 48–55.

[18] D. McGilvray, *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. California: Morgan Kaufmann, 2008.

[19] J. L. Wang, H. Li, and Q. Wang, *Research on ISO 8000 Series Standards for Data Quality*. Standard Science 12, 2010.

[20] NIST, "NIST Big Data Interoperability Framework: Volume 1, Definitions," 2015. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf .

[21] P. Patel and B. Jena, "The Opportunities , Challenges and Quality Assessment of Data in Big Data Age," vol. 8, no. Iii, pp. 251–258, 2018.

[22] M. Ghasemaghaei, S. Ebrahimi, and K. Hassanein, "Journal of Strategic Information Systems Data analytics competency for improving fi rm decision making performance," *J. Strateg. Inf. Syst.*, vol. 27, no. 1, pp. 101–113, 2018.

[23] M. Chen and M. Song, "Survey on Data Quality," pp. 1009–1013, 2012.

[24] A. Lucas, "Corporate data quality management: From theory to practice," in *Information Systems and Technologies (CISTI) 5th Iberian Conference*, 2010, pp. 1–7.

[25] World Wide Web Foundation, "Open Data Barometer," 2017.

[26] W. Fan and F. Geerts, "Foundations of Data Management, Morgan & Claypool," *Synth. Lect. Data Manag.*, vol. 4, no. 5, pp. 1–217, 2012.

[27] J. Webster and R. T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Q.*, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[28] S. B. Bacharach, "Organizational Theories: Some Criteria for Evaluation," *Acad. Manag. Rev.*, vol. 14, no. 4, pp. 496–515, 1989.

[29] D. A. Whetten, "What Constitutes a Theoretical Contribution?," *Acad. Manag. Rev.*, vol. 14, no. 4, pp. 490–495, 1989.

[30] M.-C. Boudreau, D. Robey, and G. M. Rose, "Information Technology and organizational learning: a review and assessment of research," *Account. Manag. Inf. Technol.*, vol. 10, no. November, pp. 125–155, 2000.

[31] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Sci. J.*, vol. 14, no. 0, p. 2, 2015.

[32] S. Juddoo, "Overview of data quality challenges in the context of Big Data," in *Computing, Communication and Security (ICCCS), 2015 International Conference*, 2015, pp. 1–9.

[33] C. Batini and M. Scannapeico, *Data Quality: Concepts, Methodologies & Techniques*. Verlag Berlin: Springer, 2006.

[34] S.-M. Tam and F. Clarke, "Big Data, Statistical Inference and Official Statistics," Australia, 2015.

[35] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, "From Data Quality to Big Data Quality," *J. Database Manag.*, vol. 26, no. 1, pp. 60–82, 2015.

[36] X. Zhu and S. Gauch, "Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000.

[37] N. Abdullah, S. A. Ismail, S. Sophiayati, and S. M. Sam, "Data quality in big data: A review," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. Specialissue3, pp. 16–27, 2015.

[38] G. Shanks and B. Corbitt, "Understanding Data Quality : Social and Cultural Aspects," *10th Australas. Conf. Inf. Syst.*, no. 1998, pp. 785–797, 1999.

[39] M. A. Tate and J. E. Alexander, *Web wisdom: How to evaluate and create information quality on the Web*. CRC Press, 1999.

[40] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *J. Am. Med. Informatics Assoc.*, vol. 20, no. 1, pp. 144–151, 2013.

[41] I. Taleb, R. Dssouli, and M. A. Serhani, "Big data pre-processing: a quality framework," in *Big Data (BigData Congress) 2015 IEEE International Congress*, 2015, pp. 191–198.

[42] N. Roberson, "Big Data: It's About the Quality, Not The Quantity," *Business 2 Community*, 2013. [Online]. Available: https://www.business2community.com/big-data/big-data-quality-quantity-0670201. [Accessed: 01-Oct-2017].

[43] R. Radhakrishna, D. Tobin, M. Brennan, and J. Thomson, "Ensuring Data Quality in Extension Research and Evaluation Studies.," *J. Ext.*, vol. 50, no. 3, p. 3, 2012.

[44] O. Kwon, N. Lee, and B. Shin, "International Journal of Information Management Data quality management , data usage experience and acquisition intention of big data analytics," *Int. J. Inf. Manage.*, vol. 34, no. 3, pp. 387–394, 2014.

[45] R. Ijab, M. T., Ahmad, A., & Abdul Kadir, "Challenge of Data Quality: Towards A Big Data Quality Framework, IMPACT: Technologies for Society's Well-Being," *Universiti Kebangsaan Malaysia (UKM)*, p. 44, 2016.

[46] M. T. Ijab, A. Ahmad, R. A. Kadir, and S. Hamid, "Towards Big Data Quality Framework for Malaysia ' s Public Sector Open Data Initiative," in *International Visual Informatics Conference*, 2017, vol. 1, pp. 79–87.

[47] B. Saha and D. Srivastava, "Data Quality : The other Face of Big Data," in *Data Engineering (ICDE), 2014 IEEE 30th International Conference*, 2014, pp. 1294–1297.