Journal of Information Systems and Technology Management – Jistem USP
Vol. 17, 2020, e202017003
ISSN online: 1807-1775
DOI: 10.4301/S1807-1775202017003 Original Article

BIG DATA QUALITY DIMENSIONS: A SYSTEMATIC LITERATURE REVIEW

Anandhi Ramasamy http://orcid.org/0000-0003-4833-0183
Soumitra Chowdhury http://orcid.org/0000-0003-4301-478X

Linnaeus University, Växjö, Sweden

ABSTRACT

Although big data has become an integral part of businesses and society, there is still concern about the quality aspects of big data. Past research has focused on identifying various dimensions of big data. However, the research is scattered and there is a need to synthesize the ever involving phenomenon of big data. This research aims at providing a systematic literature review of the quality dimension of big data. Based on a review of 17 articles from academic research, we have presented a set of key quality dimensions of big data.

Keywords: data, big data, quality, dimensions, assessment

Manuscript first received: 2019/09/14. Manuscript accepted: 2020/02/03

Address for correspondence:
Anandhi Ramasamy, Linnaeus University, Växjö, Sweden. E-mail: ar223fx@student.lnu.se
Soumitra Chowdhury, Linnaeus University, Växjö, Sweden. E-mail: sontuchowdhury@yahoo.co.uk

INTRODUCTION

Data has become an integral part of organizations. It is considered as an economic commodity and has raised to the extent that data is even called as the most valuable resource next to oil (Economist 2017). The volume and variety of the data generated has increased multi fold. Over 2.5 quintillion bytes of data are created every single day[1]. Low-cost commodity hardware and open source platforms have made the processing of petabyte and exabyte of such big data much easier. From the aspects of variety and especially volume, the quantity of data available has increased manifold. But the follow-up question here is whether big data means better data? As Taleb (2013) notes,

> There is plenty of information. The problem – the central issue – is that the needle comes in an increasingly larger haystack.

Huge quantities of data does not automatically guarantee quality. And with larger volumes it becomes more important to focus on the quality in order to derive some meaningful insights out of the available data. In most contexts the worth of the data is determined by its 'fitness for use', this criteria of data is still the central part to determine the necessity or mandate for organizations to invest in Big data.

The foremost part of data quality management is the data quality assessment. In the data quality literature, data quality of a system is assessed by means of several dimensions. Some of the most commonly used dimensions to measure data quality of systems are Accuracy, Completeness, Validity, Uniqueness, Consistency, etc., While considering Big data, the most important characteristics of Big data, the 3 Vs – Velocity, Variety and Volume play a crucial role. The ever increasing research on big data quality dimensions are scattered and therefore, it is important to review data quality dimensions applicable in the Big data era.

FOUNDATIONS AND BOUNDARIES

Data quality researchers often focus on the several aspects of data quality management. In the area of data quality assessment, dimensions are commonly used to define and assess the quality of data. Wang & Strong (1996) defines a "data quality dimension" as a set of data quality attributes that represent a single aspect or construct of data quality.

Data Quality & Data Quality Dimensions

There is no agreed upon one definition for data quality. Wang & Strong (1996) defines data quality as data that are fit for use by data consumers. Data quality is widely used to represent a set of "characteristics" data, such as its accuracy, completeness, consistency, and timeliness (Fu & Easton, 2017). Some or many of such characteristics determine the different dimensions that data quality can be represented upon. A poor (or low) level of data quality can have a severe impact on the overall effectiveness of the corresponding data applications (Fu & Easton, 2017).

[1] <https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf>

Though several researchers put forward many different types of dimensions to define and assess data quality, the six dimensions for data quality laid out by the International Data Management Association (DAMA) provides a comprehensive list of the data quality dimensions as represented in Figure 1. DAMA (2013) defines the six data quality dimensions in the following way:

- **Completeness:** The proportion of stored data against the potential of "100% complete";
- **Uniqueness:** Nothing will be recorded more than once based upon how that thing is identified;
- **Timeliness:** The degree to which data represent reality from the required point in time;
- **Validity:** Data are valid if it conforms to the syntax (format, type, range) of its definition;
- **Accuracy:** The degree to which data correctly describes the "real world" object or event being described;
- **Consistency:** The absence of difference, when comparing two or more representations of a thing against a definition.

Figure 1. Data Quality Dimensions adapted from (DAMA, 2013)

Data Quality in Big Data Systems

Modern information era produces tons of data every single second. With the advent of cost effective smart phone technology and internet, users throughout the world create so much data. With big data comes bigger responsibilities and bigger challenges. As the volume and variety grows, it becomes more challenging to control each data entry in order to ensure the data is of good quality. The variety and volume of the big data adds more challenges to the conventional data quality dimensions such as accuracy and completeness. With data coming in real time such as streaming data or Internet of things, it becomes harder to assess the quality of the data. When the quality cannot be assessed, the data becomes unreliable resulting in inaccurate decision making.

One argument put forward by studies is that the volume would compensate the quality that is as the volume becomes huge with Big data, the quality issues will get diluted. But this argument is quite selective to certain scenarios and is a dangerous precedence to consider it as true across all big data systems.

Therefore, it is important to analyze the existing the quality dimensions used widely to assess the data quality to determine whether they are still applicable to big data as well. Also in parallel, research has to be performed to identify any new dimensions required to assess the quality of big data. For structured data, data quality literature offers several contributions that propose assessment algorithms for these consolidated dimensions, but big data pose new challenges related their main characteristics: volume, velocity and variety. In particular, in order to address volume and velocity issues, it is necessary to redesign assessment methods for exploiting parallel computing scenarios and for reducing the computation space (Ardagna et al., 2018).

Scope of the review

The aim of this review study is to analyze the available scientific literature focusing on the various data quality dimensions used in the assessment of data quality in big data systems. Previous reviews (Liu et al., 2015; Anstiss, 2012) in this area have focused on the various data quality issues in big data analytics in a broader sense ranging from technical issues, business issues till ethical issues.

The scope of this review is limited to the research on data quality assessments in big data systems with focus on various data quality dimensions used in such assessments.

The review questions are as follows:

1. What are the various data quality dimensions used in the research to define a data quality assessment framework in Big data systems?
2. Are the conventional dimensions still applicable to Big data systems as well?
3. Are there new dimensions that emerge which are applicable to Big data in specific?

RESEARCH METHODOLOGY

Systematic literature review of research on data quality in Big data systems was conducted based on the guidelines laid down by Creswell & Creswell (2017) and by Kitchenham (2004).

First a literature search was carried out which started with broader topics such as ‘Data Quality Dimensions’ and got narrowed down to ‘Data Quality dimensions in Big data systems’. Data quality research discusses in detail about the various dimensions that are relevant based on the context. It was found during the search followed by the initial analysis that the research area of Data quality assessment with specific focus on Big data systems from Data quality dimensions point of view was quite interesting. Hence the title of the review was decided to be ‘Big Data Quality Dimensions - Data Quality Assessment’.

Further deeper searches (forward and backward) were done for this particular topic chosen. The searches were then filtered to arrive at 17 research papers as the final list to be reviewed. Deeper analysis he review were carried out having this title as the guideline throughout the review process. As (Creswell & Creswell, 2014) mentions choosing the title in the initial part of the review helped in keeping the focus of the review to the title, avoiding any diversions to closely related topics. The research methodology followed is briefly described in the figure below.

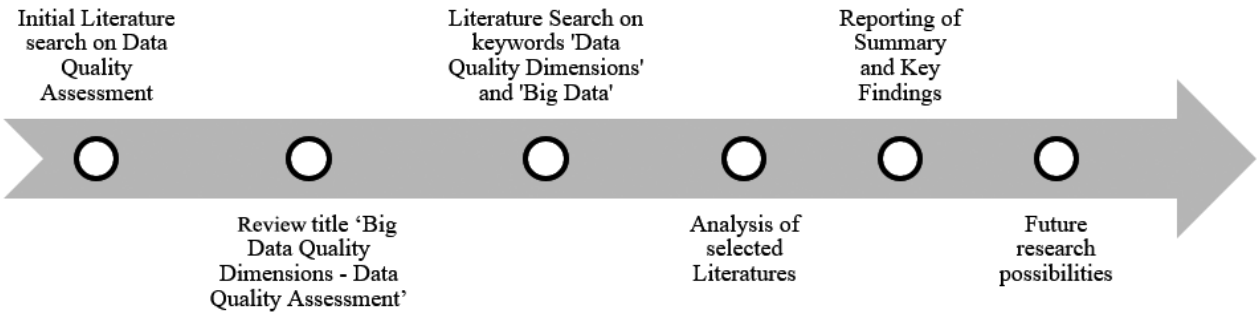

Figure 2. Research methodology

The literature search was carried out in 4 steps. First, a search was performed using keywords Data Quality' along with 'Big Data' on Scopus. Most of the relevant search results were conference proceedings published in IEEExplore. Hence the second stage of the search was carried out on IEEE using the keywords 'Data Quality Dimensions' and 'Big data' together. Third, a search was made on the reference list/bibliographies of the articles. Finally, a search was made on Google Scholar and Research Gate for the above keywords. The search yielded 105 matches.

The next step after literature search was to narrow down the research in terms of relevance, availability of research paper and contents. Only English literature which have full text available were selected. Next some search results were excluded due to relevance of context and content.

Though the topic of data quality has been researched with focus on multiple dimensions in the last couple of decades, the specific focus on Big data systems is new and hence the search and review was limited to research articles published in journals and conference proceedings. This eliminated few other research articles which just mention about the concerned topics and not in detail or just provide an overview of the said subject without much information in detail. The number of articles thus remained were 45.

Further each of the 45 research papers were read with the aim to narrow down the research papers based on relevance to the topic chosen. Half of the articles discussed about data quality in Big data systems but on a very high level of mentioning that new systems such as Big data poses new challenges to data quality. Hence only the articles which discuss in detail about the relevance of data quality in Big data were chosen. Throughout this reading, the title of the review was used as a guideline to choose only the closely related research papers which discuss about Big Data Quality Dimensions from Data Quality Assessment point of view. Figure 3 represents the methodology followed for the literature search.

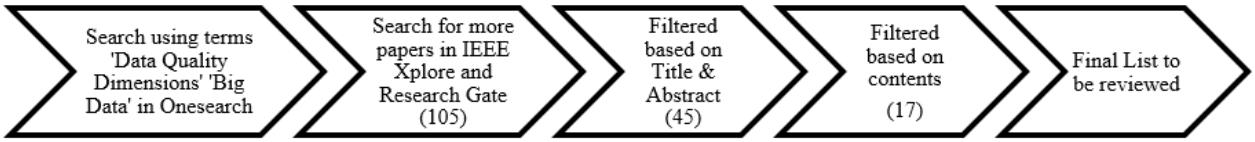

Figure 3. Methodology followed for Literature Search

LITERATURE REVIEW AND SYNTHESIS

Based on literature analysis, the articles were categorized based on the central theme of the research. Out of the papers analyzed, the most used dimensions to assess data quality are completeness, accuracy, correctness/validity, consistency and timeliness. These dimensions are however common for any data quality assessment irrespective of their relevance to Big data systems and Big data in general.

Out of the 17 research studies analyzed, 15 of them focus on data quality dimensions in Big data systems. 8 of the studies discuss about only one or many of the dimensions that are common to data quality assessment such as completeness, accuracy, correctness/validity, consistency and timeliness. Case studies and surveys form the most prominent mode of research study. Also most of the research is published in conferences and 5 of them are journal articles. 10 of the 17 research studies were conference proceedings published in one of the various conferences held by IEEE.

The list of the research articles reviewed along with the data quality dimensions they focus on are represented in Table 1. A summary of the research is illustrated below followed by the key findings that emerged out of the review.

Table 1. Classification of big data quality (classification style is adapted from Laranjeiro et al., 2015).

Research Work	Structure	Terms used in relation to dimensions
An Hybrid Approach to Quality Evaluation Across Big Data Value Chain (Serhani et al., 2016)	2 Categories, 4 Dimensions	***Contextual:*** Accuracy ***Intrinsic:*** Timeliness, Completeness, Consistency
An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics (Woodall et al., 2014)	1 Dimension	Completeness
Big Data Pre-Processing: Closing the Data Quality Enforcement Loop (Taleb & Serhani, 2017)	3 Dimensions	Accuracy, Completeness, Consistency

Table 1. Cont.

Research Work	Structure	Terms used in relation to dimensions
Big Data Quality: A Quality Dimensions Evaluation (Taleb et al., 2016)	3 Dimensions	Accuracy, Completeness, Consistency
Big Data Quality: A Survey (Taleb et al., 2018)	4 Categories 18 Dimensions	***Intrinsic:*** Accuracy, Timeliness, Consistency, Completeness ***Contextual:*** Reputation, Relevancy, Accessibility, Quantity, Value-added, Believability ***Representational:*** Interpretability, Representational, Conciseness of representation, Consistency, Manipulability, Ease of understanding ***Accessibility:*** Access, Security
Big Data Validation Case Study (Xie et al., 2017)	4 Dimensions	Validity, Completeness, Consistency, Accuracy
Big Data, Big Data Quality Problem (Becker et al., 2015)	7 Dimensions	Accuracy, Precision, Completeness, Consistency, Timeliness, Lineage/ Pedigree and Relevance
Context-aware data quality assessment for big data (Ardagna et al., 2018)	7 Dimensions	Accuracy, Completeness, Consistency, Distinctness, Precision, Timeliness, Volume
Data quality assessment: The Hybrid Approach (Woodall et al., 2013)	2 Dimensions	Completeness, Accuracy
Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications (Hazen et al., 2014)	2 Categories, 4 Dimensions	***Contextual:*** Accuracy ***Intrinsic:*** Timeliness, Completeness, Consistency
Data quality in big data processing: Issues, solutions and open problems (Zhang et al., 2017)	4 Dimensions	Availability, Usability, Reliability, Relevance
Data Quality Issues in Big Data (Rao et al., 2015)	5 Dimensions	Accuracy, Confidentiality, Completeness, Volume, Timeliness
Data quality management, data usage experience and acquisition intention of big data analytics (Kwon et al., 2014)	2 Dimensions	Consistency, Completeness

Table 1. Cont.

Research Work	Structure	Terms used in relation to dimensions
From Data Quality to Big Data Quality (Batini et al., 2015)	7 Clusters 17 Dimensions	***Accuracy:*** Correctness, Validity and Precision ***Completeness****:* Pertinence, Relevance ***Redundancy:*** Minimality, Compactness and Conciseness ***Readability****:* Comprehensibility, Clarity and Simplicity ***Accessibility****:* Availability ***Consistency****:* Cohesion and Coherence ***Trust****:* Believability, Reliability and Reputation
My (Fair) Big Data (Catarci et al., 2017)	3 Dimensions	Consistency, Accuracy, Confidentiality
The Challenges of Data Quality and Data Quality Assessment in the Big Data Era (Cai & Zhu, 2015)	5 Dimensions	Availability, Usability, Reliability, Relevance, and Presentation quality
Big Data Quality Metrics for Sentiment Analysis Approaches (El Alaoui, Gahi & Messoussi 2019)	11 dimensions	Real-time analyzability, accuracy, completeness, uniqueness, transformation, conformity, normalization, referential integrity, consistency, credibility, freshness

The earliest research on the area of big data quality with focus on the data quality dimensions appeared on late 2013. This infers that the research on Data Quality of Big data systems is still in the nascent stages. Most of the research on quality of Big data acknowledges the role of the conventional dimensions in assessing data quality of big data. The most common dimensions mentioned across the research on big data quality can be classified as in Figure 4.

While dimensions such as accuracy, consistency, integrity and completeness which all relate to whether the data is reliable are the most commonly represented dimensions, the readability and structure of the data is more prominently appropriate for big data systems. Some of the big data feeds consist of semi-structured and unstructured data. Hence in order for this data to become of any use to users, the readability is a quite important dimension to be considered.

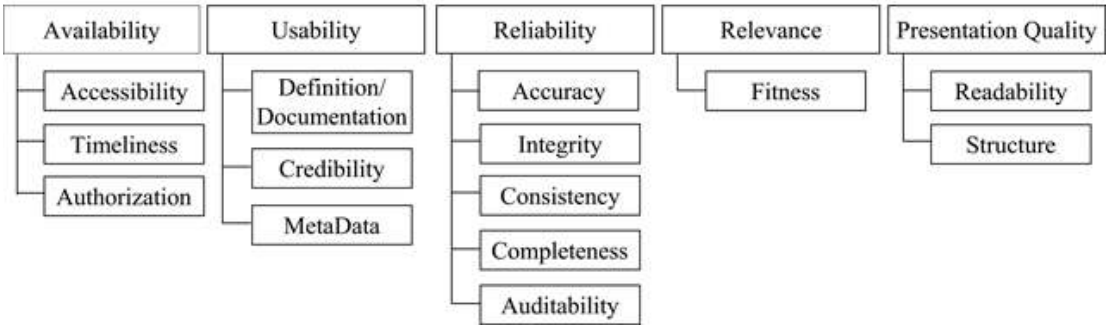

Figure 4. A two-layer big data quality standard for assessment adapted from Batini et al. (2015)

In addition to the above conventional data quality dimensions, interesting dimensions specifically focusing on Big data based on the specific characterisstics of Variety and Volume emerge in the some studies.

El Alaoui, Gahi & Messoussi (2019) highlights the importance of real-time analyzability of big data since it if often required to analyze the big data in real time. The more time spent while storing the data, the bigger is the negative effect on the results if the requirment is to understand market trends. Becker et al., (2015) mentions a key dimension in relation to Big data which is Pedigree or Lineage. Though lineage is an intrinsic part of any data quality management, it occupies a special position in Big data systems. Catarci et al., (2017) considers Confidentiality as a key quality dimension to achieve data integrity and informed policy making based on accurate and valid data.

Cai & Zhu (2015) discusses about Credibility as a dimension which determines whether the data is usable. In the big data world, lots of data are available. Even so much data is available in public domain for free which raises the question of whether the data is credible. Batini et al., (2015) provides newer views on the dimensions Redundancy, Readability and Cohesion as applicable to Big data.

Finally, Ardagna et al., (2018) discusses about a well-known and yet important dimension to be considered in big data quality assessment, Volume. Since big data means huge volumes, it is tedious to perform assessment on each data entry and quite often the assessment follows a sampling approach. It is therefore important to determine the optimum volume to be sampled in order to ensure the assessment is appropriate.

Table 2 provides an overview of key quality dimensions of big data in the modern information era.

Table 2. Key quality dimensions related to big data.

Data Quality Dimension	Definition
Accessibility	Accessibility and availability are related to the ability of the user to access data from his or her culture, physical status/functions, and technologies available (Batini et al., 2015)
Cohesion	Consistency, cohesion and coherence refer to the capability of data to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules and other formalisms (Batini et al., 2015).
Confidentiality	This quality dimension determines whether right data is in the right hands. Is the data secure? (Catarci et al., 2017)
Credibility	Data come from specialized organizations of a country, field, or industry. Experts or specialists regularly audit and check the correctness of the data content. Data exist in the range of known or acceptable values (Cai & Zhu, 2015).
Pedigree/Lineage	This dimension helps in knowing the source of the data so that any inconsistency is corrected in the source and not in any other instances

Table 2. Cont.

Data Quality Dimension	Definition
Readability	Also represented as clarity, simplicity ease of understanding, interpretability, comprehensibility, this dimension refers to ease of understanding of data by users (Batini et al., 2015).
Real-time analyzability	Sometimes data have to be analyzed in real-time. The time spent for storage could impact the results' quality (El Alaoui, Gahi & Messoussi (2019)
Redundancy	Redundancy, minimality, compactness and conciseness refer to the capability of representing the reality of interest with the minimal use of informative resources. (Batini et al., 2015).
Trust	Trust, including believability, reliability and reputation, catching how much data derive from an authoritative source (Batini et al., 2015)
Volume	This quality dimension provides the percentage of values contained in the analyzed Data Object with respect to the source from which it is extracted (Ardagna et al., 2018).

CONCLUSIONS

Most of the studies are published in conferences and very less data about this specialization is found in journals and still fewer in standard books. The review was focused on analyzing the literature and to find answers relevant to the research questions scoped in.

Based on the available studies, the review findings are as follows:

- Some of the conventional data quality dimensions are still relevant to big data. Primary among them are accuracy, completeness, consistency, uniqueness and timeliness;
- Some studies have found new dimensions that are relevant to Big data such as trust, confidentiality, etc.;
- Data quality definition and assessment in Big Data mainly depend on the data types, data sources and applications. (Ardagna et al., 2018) In particular, in order to address volume and velocity issues, it is necessary to redesign assessment methods for exploiting parallel computing scenarios and for reducing the computation space (Ardagna et al., 2018);
- Contextual data quality dimensions need to be defined and refined based on the user requirements of big data;
- Dimensions such as credibility and confidentiality are quite integral to the assessment of quality and hence cannot be ignored while building a data quality assessment framework.

In spite of the buzz with dig data everywhere, data quality assessment specializing on Big data is not widely studied yet. There are quite a few studies focusing on data quality mentioning that Big data might introduce new challenges in the assessment framework. But concrete studies have been done by very few authors. As presented in the review data quality dimensions prove crucial in the data quality assessment and it is imperative to define and examine new dimensions related to big data.

Most of the studies highlighted in the review consider the most commonly occurring form of big data which is textual data. However, big data occurs in several forms and with varied characteristics. More research focus is required in these areas for better understanding and new dimensions formulation. Big data feeds such as audio and video feeds, social media data, weather data, etc. will provide variety to the picture and might require special dimensions depending on the context. Further research in these unstructured data feeds are quite essential.

Internet of things provides another new dimension to how data is viewed in the current setup. The impact of IoT on data quality and vice versa needs focused research. Further new quality dimensions related to IoT needs to be defined.

Machine learning provides a viable option in data standardization, research on the usage of machine learning to observe data patterns and determine or assess data quality is another possible area to be studied in the future.

REFERENCES

Anstiss, S. (2012). Understanding data quality issues in dynamic organisational environments–a literature review. In *Proceedings of the 23rd Australasian Conference on Information Systems 2012* (pp. 1-10). ACIS. http://dro.deakin.edu.au/eserv/DU:30049090/anstiss-understandingdata-2012.pdf

Ardagna, D., Cappiello, C., Samá, W., & Vitali, M. (2018). Context-aware data quality assessment for big data. *Future Generation Computer Systems, 89*, 548-562. https://re.public.polimi.it/retrieve/handle/11311/1057520/295709/FutureGeneration.pdf

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From Data Quality to Big Data Quality. *Journal of Database Management, 26*(1), 60-82. https://www.igi-global.com/article/from-data-quality-to-big-data-quality/140546

Becker, D., King, T. D., & McMullen, B. (2015). Big data, big data quality problem. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 2644-2653). https://ieeexplore.ieee.org/abstract/document/7364064

Cai, L. & Zhu, Y., (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal, 14*(2), 1-10. https://datascience.codata.org/articles/10.5334/dsj-2015-002/

Catarci, T., Scannapieco, M., Console, M., & Demetrescu, C. (2017). My (fair) big data. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 2974-2979). https://ieeexplore.ieee.org/abstract/document/8258267

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications. http://us.sagepub.com/en-us/nam/research-design/book255675

DAMA, (2013). *Defining Data Quality Dimensions.* Data Management Association (DAMA)/ UK Working Group. https://is.gd/dama_def_data_quality_dim

El Alaoui, I., Gahi, Y., & Messoussi, R. (2019). Big Data Quality Metrics for Sentiment Analysis Approaches. In *Proceedings of the 2019 International Conference on Big Data Engineering* (pp. 36-43). https://dl.acm.org/citation.cfm?id=3341629

Fu, Q., & Easton, J. M. (2017). Understanding data quality: Ensuring data quality by design in the rail industry. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 3792-3799). https://ieeexplore.ieee.org/abstract/document/8258380

Hazen, B. T., Boone, C. A., Ezell, J. D. & Jones-Farmer, L. A., (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics, 154*, 72-80. https://www.sciencedirect.com/science/article/abs/pii/S0925527314001339

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews.* Keele University. http://www.it.hiof.no/~haraldh/misc/2016-08-22-smat/Kitchenham-Systematic-Review-2004.pdf

Kwon, O., Lee, N. & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management, 34*, 387-394. https://www.sciencedirect.com/science/article/pii/S0268401214000127

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A survey on data quality: classifying poor data. In *2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC)* (pp. 179-188). https://ieeexplore.ieee.org/abstract/document/7371861

Liu, J., Li, J., Li, W. & Wub, J. (2015). Rethinking big data: A review on the data quality and usage issues. *Journal of Photogrammetry and Remote Sensing, 115*, 134-142. https://www.sciencedirect.com/science/article/abs/pii/S0924271615002567

Rao, D., Gudivada, V. N., & Raghavan, V. V. (2015). Data quality issues in big data. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 2654-2660). https://ieeexplore.ieee.org/abstract/document/7364065

Serhani, M. A., El Kassabi, H. T., Taleb, I., & Nujum, A. (2016). A hybrid approach to quality evaluation across big data value chain. In *2016 IEEE International Congress on Big Data (BigData Congress)* (pp. 418-425). IEEE.https://ieeexplore.ieee.org/abstract/document/7584971

Taleb, I., El Kassabi, H. T., Serhani, M. A., Dssouli, R., & Bouhaddioui, C. (2016). Big data quality: A quality dimensions evaluation. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing*, (pp. 759-765).

Taleb, I., Serhani, M. A., & Dssouli, R. (2018). Big data quality: A survey. In *2018 IEEE International Congress on Big Data (BigData Congress)* (pp. 166-173). https://ieeexplore.ieee.org/abstract/document/8457745

Taleb, I., & Serhani, M. A. (2017). Big Data Pre-Processing: Closing the Data Quality Enforcement Loop. In *2017 IEEE International Congress on Big Data (BigData Congress)* (pp. 498-501). https://ieeexplore.ieee.org/abstract/document/8029366

Taleb, N., (2013). *Beware the big errors of 'Big Data'.* https://www.wired.com/2013/02/big-data-means-big-errors-people/

The world's most valuable resource is no longer oil, but data. (2017, May 6). *The Economist*. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

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

Woodall, P. et al., (2014). An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics. In *19th International Conference on Information Quality (ICIQ)*, Xi'an, China. https://is.gd/Woodall_et_al_big_data

Woodall, P., Borek, A. & Kumar Parlikad, A., (2013). Data quality assessment: The Hybrid Approach. *Information & Management, 50*(7), 396-382. https://www.sciencedirect.com/science/article/abs/pii/S0378720613000517

Xie, C., Gao, J., & Tao, C. (2017). Big data validation case study. In 2*017 IEEE third international conference on big data computing service and applications (BigDataService)* (pp. 281-286). https://ieeexplore.ieee.org/abstract/document/7944952

Zhang, P., Xiong, F., Gao, J., & Wang, J. (2017). Data quality in big data processing: Issues, solutions and open problems. In *2017 IEEE Smart World, Ubiquitous Intelligence & Computing*. (pp. 1-7). https://ieeexplore.ieee.org/abstract/document/8397554