JOURNAL OF SOFTWARE ENGINEERING & INTELLIGENT SYSTEMS
ISSN 2518-8739
30th April 2018, Volume 3, Issue 1, JSEIS, CAOMEI Copyright © 2016-2018
www.jseis.org

Big data and data quality dimensions: a survey

[1] Onyeabor Grace Amina, [2] Azman Ta'a
[1,2]School of Computing, Universiti Utara Malaysia, 06010, Sintok, Kedah, Malaysia
Email: [1]grace_amina@ahsgs.uum.edu.my, grameenah@gmail.com [2]azman@uum.edu.my

ABSTRACT

Data is a vital asset in virtually all types of organizations. These days data or information acquired from data analysis is the basis of decision making in various businesses or organizations in general and this offers numerous benefits by building accurate and dependable process. The degradation of its quality has erratic consequences resulting to wrong insights and decisions. Moreover, these are the days of Big Data (BD) which comes with varieties of vast amount of unprecedented data with unknown quality which makes its Data Quality (DQ) evaluation very challenging. DQ is therefore critical for the processes of data operations and management in order to detect associated performance problems. Besides, data of high quality has the ability to attain top services within an organization through enlarged prospects. Nonetheless, recognising different characteristics of DQ from its definition to the different Data Quality Dimensions (DQDs) are crucial for equipping methods and processes for the purpose of improving DQ. This paper focuses on the review of BD and the most commonly used DQDs for BD which are basis for the assessment and evaluation of the quality of BD.

Keywords: big data; data quality; data quality dimensions; big data quality;

1. INTRODUCTION

The rate of data explosion nowadays has never been apparent. Variety of data from diverse sources have been mounting immensely in large volume, with unprecedented velocity and the veracity of much of these data are uncertain. The Volume, Variety, Velocity and Veracity constitute the initial 4Vs definition of BD [1-5]. This new drift has given birth to the known phenomena called Big Data (BD). The trend has also prevailed on a change in organizational policy or strategy from the classical traditional management systems to Cloud enabled BD which brings flexible and scalable management of data and has proved to be cost effective and efficient [6-7]. Moreover, the growth of unstructured data especially indicates that data processing has gone beyond ordinary tables and rows [8-11]. This is noteworthy because data is considered as an asset in small and large business organizations especially in such an era where insights for business strategic decisions are drawn from BD [12-17]. According to [18], the insights offer new ways to the organizations by influencing fresh types of analytics on the new kinds of data. The challenge is now thrown to the organizations for the creation of fresh actions based on the profits offered by these sorts of analysis [19].

Bearing in mind that data from its sources and data analytics products are well-meaning for organizations and considering the great value of the organizations, practitioners and researchers view data as one of the significance benefit of business [20-21]. Due to the above fact, the requirement for more attention for Data Quality (DQ) in BD should not be overlooked [22-23]. One of the keys to achieving successful management data in an organization, is by attaining high DQ. Poor DQ has led organizations into several issues like wrong decisions, high cost and not being able to provide customer satisfaction [24]. As data is a vital resource in all areas of applications within business organizations and government agencies, DQ is vital for decision makers in the organizations to enable resolution performance connected concerns [25-27].

For the achievement of high quality data, there is a need to employ diverse techniques and strategies. According to [28-30] these strategies are divided into 1. Data-driven and 2. Process-driven. Data-driven strategies handles the data the way it is, by enhancing the DQ by altering directly the values of data applying techniques and activities such as integration or cleansing, while process-driven strategies make efforts to find poor DQ original sources and enhance the DQ by the redesign of the process of data creation or modification. Generally, process-driven DQ strategies has proven to perform better in comparison to data-driven strategies since its emphases is on removing the causes of DQ problems. Additionally, data-driven strategies appear to be costlier than the process-driven strategies either within the short long period [31]. There is a common phrase according to [32], by practitioners of quality control that one cannot improve what cannot be measured. Therefore, attempts should be made to operationally provide the definition and measurement of DQ. Sometimes, measurement of DQ is compared with measurement of physical product. So, [33-34] said comparing the measurement of physical product, DQ has a multidimensional problem.

Moreover, data has multidimensional concept that can be measured by various dimensions like consistency, accuracy and timeliness [35-36]. These dimensions are characteristics for the measurement and management of data and information quality across diverse domain and the metrics being used for measurement differ from context to context [37]. This paper review studies on DQ and the various dimensions from the time of the traditional data management system and their applicability in this era of BD. The rest of this paper is organized thus: Section 2 discusses BD and DQ, section 3 talked about DQDs, section 4 discussed DQDs for BD and the concluding remarks came in the last section.

2. BIG DATA AND DATA QUALITY

BD is a term used to describe huge data sets that are of diverse format created at a very high speed, the management of which is near impossible by using traditional database management systems. Organizations and businesses today are producing large datasets, the same way enormous number of data is being acquired and received from various sources and stored [38], [39]. This is the era of BD which started to be recognized a few years back. Its initial definition gives the term a poor definition of its representation; the only idea that it really conveys most frequently is of a huge volume of data too large to be managed by the current processors of computers [40], [7]. However, According to[41], [42], BD does not only concern the large volume of data but it also includes the ability to search, process, analyze and present meaningful information obtain from huge, varied and rapidly moving datasets. These three attributes lead to the foundational definition of BD regarding volume, variety, and velocity. Furthermore, [43], defined BD as high volume, high velocity and high variety assets of information demanding cost effective ground-breaking forms of information processing for improved insight and decision making. Data are created from an extensive range of sources such as social media, the internet, databases, websites, sensors, and so on. But before these data are stored, processing and cleansing with the help of numerous analytical algorithms are performed on them [44], [45]. However, because of the nature of BD, oftentimes, organizations encounter issues and challenges. BD acquired are in large volume, of different varieties and with unprecedented velocity which makes it challenging to manage the data. These concerns and challenges need to be looked into for the stored data to be simply retrieved for making proper business decisions prospectively [46]. [47] identified the challenges and the riskiest of them is DQ.

DQ as a concept is not easily defined. The studies related to DQ began as far back as in the 90s - the days of database management systems. Since then various researchers have proposed diverse definitions of DQ [48]. According to [35] the group of Total Data Quality Management led by Professor Richard Wang of the MIT University, with their in-depth research in the area of DQ defined it as fitness for use. And henceforth other researchers in the field came up with their own definition in the literature as meeting the users' expectations [49] (Sebastian-Coleman, 2012) or data suitable for use by data users [50-53], [54-55, 40]. DQ is defined by the International Organization for Standardization/International Electrotechnical Commission 25012 standard (ISO/IEC 2008) [56] as the extent to which a set of features of data meets requirements. All the above definitions of DQ clearly indicates that DQ is highly reliant on the context of the use of data and interactions to the customers' requirements, the ability to use and access data [18]. According to [57], it was pointed out, that to enhance DQ, two strategies are involved which are: data-driven and process-driven. The first strategy which is data-driven handles the data the way it is, making use of methods and actions like cleansing to enhance the quality of the data. And secondly, Process-driven strategy tries to detect originating poor DQ sources then redesigns the way the data is produced. DQ problems exist, right before the introduction of BD in the field. According to [13], the researchers categorized DQ issues and challenges according to (i) errors correction, (ii) unstructured data to structured conversion and (iii) integrating data from various sources of data. To add to the issues mentioned above, there exist quite a number of particular BD challenges, which include the large volume of data generated by web 2.0 moving at an unusual speed, contained within schema-less structures. Other BD quality issues are also identified related with BD features [35, 58-60]. Because of these joint issues, the processes of BD cleaning and sifting are phases to be implemented before the analyses of data with quality that is unknown. In [61] it is pointed out that DQ problems are more pronounced when dealing with data from multiple data sources. This problem obviously multiplies the data cleansing needs. Also, the huge amount of data sets that comes in at an unprecedented speed creates an overhead on the cleansing processes [13]. With the magnitude of data generated, the velocity at which the data arrives, and the huge variety of data, the quality of these data has left so much to be desired.

There has been an estimation of inaccurate data costing US businesses 600 billion dollars yearly [62]. The error rate in data as recorded by enterprises is typically estimated to between 1% and 5%, while for some organizations; it is well above 30% [63-64]. In the majority of data warehouse projects, data cleaning amounts to 30% to 80% of the developmental time plus the budget for enhancing the DQ against building the system. Regarding the web data, about 58% of the available files are XML, out of this volume, only 1/3 (one-third) of the XML documents with associated XSD/DTD are valid [65]. Also, about 14% of the documents are not well-formed, which is a simple mistake of tags that are mismatched and omitted tags that render the whole XML-technology unusable over these documents. All these pinpoint the pressing requirement for DQ management to make sure data in the databases exemplify the real world objects to that are refer in a reliable, consistent, precise, comprehensive, well-timed and exceptional way. There has been increase in demands by business organisations to develop DQ management systems, with the sole aim of detecting and efficiently correcting data errors. Thus, this move adds accurateness and value to the underlying business processes. Indeed, it is estimated that the rate of growth of DQ tools in the market is growing at 16% annually. This value is far above the average estimate of 7% for other IT sectors [66].

Data is exposed to auditing, profiling and the application of quality rules in a DQ system, with the aim of keeping and/or improving the quality. DQ concept has been known in the database community and it has not been a passive area of database management research for many years [67], [68-69]. Nevertheless, to apply directly these quality concepts to BD encounters serious problems as regards to the costing as well as the timing for data processing. This issue is made worst knowing the fact that these techniques were designed in the context of structured data [70]. Within the context of BD, any DQ application must be designated base on the origin, domain, format, and the data type it is being applied. It is essential to properly manage these DQ systems in solving the many problems rising in dealing with such vast data sets. In addition, for DQ to be managed, it must be measurable using the DQDs which is reviewed in the following section.

3. DATA QUALITY DIMENSIONS

DQ can be analyzed from multiple dimensions. A Data Quality Dimension (DQD) is a feature or information part use data requirements. DQD provides the way to measure and manage DQ [27, 57, 71-72], [27], [73]. It is a quantifiable property of DQ which is a representative of some feature of the data such as accuracy, consistency and completeness used in the guidance of the process of giving quality understanding [74]. Consequently, the description of some specific data could be said to be high in quality, depending on one or multiple dimensions. It is a usual phenomenon to find different terms denoting the same dimensions in the literature. For example, currency is sometimes referred to as timeliness due to the fact that use of data is universal [75]. Also, DQDs are on many occasions denoted as characteristics, or attributes [76]. Usually, data is altered owing to some factors such as the reading of sensor's devices, human data entry error, missing values in data, social media data and all sorts of unstructured data. These factors should be identified and categorised under the DQDs especially when the quality requires improvement and evaluation [13]. This is because DQ problems usually referred to as dirty or poor data are typically the particular problem existent and manifests within a DQD, for example, format glitches suffaces under the accuracy DQDs, and when data lacks in the appropriate format, it cannot be stared as quality data [77]. According to [78], various terms are used in the description of the data DQ related issues. as well as the mapping between the various problems to each of the relevant dimensions. The researchers in [56, 79] listed some details of dirty data affecting its quality component and the dimensions is associated with. Below is the table by Taleb, 2016 oa short list of the familiar DQ issues associted with DQDs.

Table. 1 Data quality issues vs data quality dimensions

		Data Quality Dimensions Related		
	Data Quality Issues	**Accuracy**	**Completeness**	**Consistency**
Instance Level	Missing data	X	X	
	Incorrect data, Data entry errors	X		
	Irrelevant data			X
	Outdated data	X		
	Misfielded and Contradictory values	X	X	X
Schema Level	Uniqueness constrains, Functional dependency violation	X		
	Wrong data type, poor schema design			X
	Lack of integrity constraints	X	X	X

3.1 Types of Data Quality Dimensions

There are several types of DQDs available in the literature and each of them is linked to specific metric [80-85]. The researchers in [86] identified forty DQDS that existed from 1985 to 2009. In addition, [73] revealed one hundred and twenty-seven DQDs from the analysis of sixteen sources selected for the study. Although, the DQDs that are commonly seen in the literature are categorized into intrinsic and contextual according to [32, 31, 80-81], both [35, 86] initially grouped DQs into four categories that are Intrinsic, Accessibility, Contextual, Representational in DQ field which are based on their dimensions.

- Intrinsic DQDs refers to data features that are native to the data and objective
- Accessibility DQDs are categorized by fundamental issues relating to technical data access
- Contextual DQDs refers to data features that are reliant on the context in which the data are perceived or used
- Representational DQDs refers to how data is presented

The table below shows the categorization of DQDs:

Table. 2 Data quality categories and dimensions

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

Furthermore, other researchers have recognised different framework and methodology for the assessment and improvement of DQ using various approaches and methods on DQDs [27]. These scholars demonstrated descriptions for DQDs and brought to recognition more significant DQDs [27, 82 84, 87], [2, 11, 12, 22].

4. DATA QUALITY DIMENSIONS FOR BIG DATA

Some studies have been conducted in organizations and in the academics regarding DQDs in BD. It has been observed from the various research works that in most cases the DQDs used in measuring that are used on the traditional data management systems are applicable in the measurement of DQ in BD. However, the DQDs for BD are categorized into intrinsic and contextual and intrinsic [32]. Contextual DQDs are connected with the values of data and intrinsic DQDs are related to the data intention, that is, the schema of the data [31, 13]. The intrinsic is commonly used and frequently found in the literature [18, 13]. The intrinsic DQ consist of the following dimensions: Accuracy, Completeness, Consistency and Timeliness. The above DQDs are associated with the ability of data to map the interest of the data user [88].

Intrinsic DQ dimensions comprises of i. Accuracy: which measures whether logging of data was done correctly and shows precise values. ii. Timeliness: This measures that if data is up to date or not which is occasionally signified as data volatility and currency [89]. iii. Consistency: This measures agreement of data with its format and structure. Studies on BDQ refers to conditional functional dependencies as DQ rules to identify semantic faults [90-91]. iv. Completeness: This measures that if all data that are relevant are correctly recorded without missing values or entries [13]. The features of BD, that is, volume, velocity, variety and veracity have a more or less result on the area of DQ. A concern is that the DQ cannot just be described by the traditional DQDs, but also requires care to be taken considering BD characteristics. This is called BD quality dimensions. The authors in [19] even merge BD characteristics – 3 Vs with DQDs based on International Standard Organization/ International Electrotechnical Commission ISO/IEC. Thus, for example, additional dimensions such as performance, relevancy, popularity and credibility are measured for quality of social media data [92-93]**.** Accuracy, Completeness, consistency and timeliness were also used to evaluate BD in health sector [8].

5. CONCLUSION

DQ issue is a serious issue for organizational operation processes to be able to identify associated performance issues since data is a vital resource in organizations, businesses and governmental agencies [28, 31, 62]. Organizational data are no longer limited to just databases as new technologies emerge. BD sources have turned out to be significant in organizations. This paper reviewed the literature on BD and DQDs from both the traditional data and BD. From the viewpoint BD quality research as compared with traditional data still has much to cover in using DQDs for the assessment and evaluation of BD. The literatures right from the inception of DQ have defined DQ in different ways and have identified various vital DQDs. Reaching up to one hundred and twenty-seven DQDs. It is also reviewed that the most common DQDs used for BD are Accuracy, Consistency, Completeness and Timeliness. Therefore, there's still much to be covered in BD for DQDs for effective and efficient measurement of BD quality.

REFERENCES

1. Chen, M., S. Mao, and Y. Liu, *Big data: A survey*. Mob. Netw. Appl., 2014. 19(2): p. 171–209.
2. Philip Chen C. L. and. C.-Y. Zhang, *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.* Inf. Sci., 2014. 275: p. 314–347.
3. Wielki, J., *The opportunities and challenges connected with implementation of the big data concept,* in *Advances in ICT for Business, Industry and Public Sector,*
4. C. M. Olszak, and T. Pe_ech-Pilichowski, Editors. 2015, Springer. p. 171–189.
5. Hashem, I. A. T., I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, *The rise of 'big data' on cloud computing: Review and open research issues.* Inf. Syst., 2015. 47: p. 98–115, 2015.
6. Hu, H.,Y. Wen, T.-S. Chua, and X. Li, *Toward Scalable Systems for Big Data Analytics: A Technology Tutorial,* IEEE Access 2014. 2, p. 652–687.
7. N. I. of Standards, *Draft NIST big data interoperability framework: security and privacy,* Report, 2015.US. Department of Commerce.
8. Serhani, M. A., H. T. El Kassabi, I. Taleb &A. Nujum, *An hybrid approach to quality evaluation across big data value chain.* In *Big Data (BigData Congress.* 2016. IEEE International Congress. p. 418-425.
9. N. I. of Standards, *Draft NIST big data interoperability framework: reference architecture,* Report, 6. 2015. US. Department of Commerce.
10. Gubbi, J. R. Buyya, S. Marusic, M. Palaniswami, *Internet of things (IoT): A vision, architectural elements, and future directions,* Future Generation. Computer. Syst. 2013. 29(7), p. 1645–1660. Including Special sections: Cyber-enabled Distributed Computing for Ubiquitous Cloud and Network Services and Cloud Computing and Scientific Applications Big Data, Scalable Analytics, and Beyond. http://dx.doi.org/10.1016/j.future.2013.01.010.URL: http://www.sciencedirect.com/science/article/pii/S0167739X13000241
11. Pääkkönen, P. and D. Pakkala, *Reference architecture and classification of technologies, products and services for big data systems,* Big Data Res., 2015. 10.1016/j.bdr.2015.01.001.
12. Madnick, S. E., R. Y. Wang, Y. W. Lee, and H. Zhu, *Overview and framework for data and information quality research,* Journal of Data Inf. Quality, 2009. 2.
13. Taleb, I., H. T. El Kassabi, M. A. Serhani, Dssouli, R., & Bouhaddioui, C. *Big data quality: A quality dimensions evaluation* in *Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016.* Intl IEEE Conferences, p.759-765).
14. Bhatia, S., J. Li,W. Peng, and T. Sun, *Monitoring and analyzing customer feedback through social media platforms for identifying and remedying customer problems,* in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining (ASONAM).* 2013, p. 1147-1154.
15. Antunes, F. and J. P. Costa, *Integrating decision support and social networks,* Adv. Human-Comput. Interact., 2012, 9.
16. Fabijan, A., H. H. Olsson, and J. Bosch, *Customer feedback and data collection techniques in software R&D: A literature review* in *Software Business* (Lecture Notes in Business Information Processing), 2015. 210. Springer. p. 139-153.
17. Ferrando-Llopis, R., D. Lopez-Berzosa, and C. Mulligan, *Advancing value creation and value capture in data-intensive contexts* in *Proc. IEEE Int. Conf. Big Data*. 2013, p. 5-9.
18. Izham Jaya, M., F. Sidi, I. Ishak, L.I. L. L. Y. Suriani Affendey, & M. A. Jabar, *A Review Of Data Quality Research In Achieving High Data Quality Within Organization*. Journal of Theoretical & Applied Information Technology, 2017.*95*(12).
19. Merino, J., I. Caballero, B. Rivas, M. Serrano, & M. Piattini, *A data quality in use model for big data.* Future Generation Computer Systems,2016. *63*, p.123-130.

20. Gandomi, A., and M. Haider, *Beyond the hype: Big data concepts, methods, and analytics*, Int. J. Inf. Manage., 2015. 35, p. 137–144.

21. Lesser, E., and R. Shockley, *Analytics: The new path to value.* 2014. URL: http://www-935.ibm.com/services/us/gbs/thoughtleadership/ ibv-embedding-analytics.html.

22. Finch, G., S. Davidson, C. Kirschniak, M. Weikersheimer, C. Rodenbeck Reese, and R. Shockley, *Analytics: The speed advantage. why data-driven organizations are winning the race in today’s marketplace*. 2014. URL:http://www-935.ibm.com/services/us/gbs/thoughtleadership/2014analytics/?cm_mc_uid=48208801620614296137181&cm_mc_sid_50200000=1429613718.

23. Lundquist, E., *Data quality is first step toward reliable data analysis.* URL: http://search.ebscohost.com/login.aspx?direct=true& b=bth&AN=89867448&lang=es&site=ehost-live.

24. Kwon, O., N. Lee, B. Shin, *Data quality management, data usage experience and acquisition intention of big data analytics*, Int. J. Inf. Manage. 2014. 34, p.387–394. URL: http://www.sciencedirect.com/science/article/pii/ S0268401214000127

25. Strong, D. M., Y. W. Lee, and R. Y. Wang, *Data Quality in Context* , Communications of the ACM, 1997. 40(5), p. 103–110.

26. Eckerson W., *Data Warehousing Special Report: Data quality and the bottom line, Applications Development Trends*, 2002.

27. Batini, C., C. Cappiello, C. Francalanci, and A. Maurino, *Methodologies for data quality assessment and improvement.* ACM Computing Surveys, 2009. 41(3), p. 1–52.

28. Tee, S. W., P. L. Bowen, P. Doyle and F. H. Rohde, *Factors influencing organizations to improve data quality in their information systems,* Accounting and Finance, 2007. 47(2), p. 335–355.

29. Sidi, F., P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, *Data quality: A survey of data quality dimensions* in *International Conference on Information Retrieval Knowledge Management (CAMP)*, 2012. p. 300 –304.

30. Glowalla, P., P. Balazy, D. Basten, and A. Sunyaev,*Process-driven data quality management – An application of the combined conceptual life cycle model* in *47th Hawaii International Conference on System Sciences (HICSS)*, 2014. p. 4700–4709.

31. Batini C, Cappiello C et al. (2009). Methodologies for data quality assessment and improvement, ACM Computing Surveys, vol 41(3), 1–52, doi:[10.1145/1541880.1541883].

32. Hazen, B. T., C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, *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, 2014. 154, p. 72-80.

33. Garvin, D.A., *What does product quality really mean?.* Sloan Manage. Rev., 1984. 26 (1),p.25–43.

34. Garvin, D.A., *Competing on the eight dimensions of quality*. Harvard Bus. Rev., 1987. 65 (6), p.101–109.

35. Wang, R. Y, and D. M. Strong, (1996). *Beyond accuracy: What data quality means to data consumers,* Journal of Management Information Systems, 1996. 12(4), p. 5–33.

36. Heinrich, B., M. Kaiser, and M. Klier, *How to measure data quality? A metric-based approach,* Twenty Eighth International Conference on Information Systems, Montreal, 2007, p. 101–122.

37. Huang H, B. Stvilia, C. Jorgensen and H. W. Bass, (2012). Prioritization of data quality dimensions and skills requirements in Genome annotation work, Journal of the American Society for Information Science and Technology, 2012. 63(1), p. 195–207.

38. Maier, M., A. Serebrenik, and I. T. P. Vanderfeesten, 2013). *Towards a big data reference architecture,* 2013. University of Eindhoven

39. Immonen, A., P. Pääkkönen, & E. Ovaska, *Evaluating the quality of social media data in big data architecture*. *IEEE Access*, 2015. *3*, p. 2028-2043.

40. Loshin, D., *Big Data Analytics: From strategic planning to enterprise integration with tools, techniques.* No Sql, and Graph, 2013. Elsevier.

41. Malik, P., *Governing Big Data: Principles and Practices*. IBM Journal of Research and Development, 2013.

42. Mahanti, R. *Critical success factor for implementing data profiling: The first step toward data quality*. Softw. Qual. Prof., 2014. 16 (2), p. 13–26.

43. G. Inc., Gartner’s IT glossary. URL: http://www.gartner.com/it-glossary/bigdata 2015

44. Lock, M. (2012). *Data management for BI big data*. Aberdeen Group, 2012. p. 4-14. http://www.facebook.com/l.php?u=http%3A%2F%2Fvertica.com%2Fwpcontent%2Fuploads%2F2012%2F03%2FDataManagementforBI_Aberdeen.pdf&h=oAQE7lPo5

45. Soares, S., *Big Data Governance: An Emerging Imperative*, MC Press, 2012.

46. Feldman M., *The big data challenge: Intelligent tiered storage at scale.* White Paper, 2013, p. 7-8.

47. Parkinson, J., *Six big data challenges*. CIO Insight. 2012. URL: http://www.cioinsight.com/c/a/Expert-Voices/Managing-Big-Data-Six- Operational-Challenges- 484979.

48. Cai, L., and Zhu, Y., *The challenges of data quality and data quality assessment in the big data era*. Data Science Journal, 2015. 14.

49. Sebastian-Coleman, L., *Measuring data quality for ongoing improvement: a data quality assessment framework*, 2012. Newnes.

50. Strong, D. M., Y. W. Lee, and R. Y. Wang, *Data Quality in Context.* Communications of the ACM, 1997, 40(5), p. 103–110.

51. Lee Y. W. and D. M. Strong, *Knowing- Why About Data Processes and Data Quality. Journal of Management Information Systems*, 2003, 20(3) p. 13–39.

52. Levitin A. V. and T. C. Redman, *Data as a resource: properties, implications, and prescriptions.* Sloan Management Review, 1998, 40, p. 89–101.

53. Wang, R. Y., *A product perspective on total data quality management*. Communications of the ACM, 1998, 41(2), p. 58–65.

54. Sidi, F., P. H. S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim and A. Mustapha, *Data quality: A survey of data quality dimensions* in Information Retrieval & Knowledge Management (CAMP), International Conference, 2012 (pp. 300-304). IEEE.

55. Alizamini, F. G., M.M. Pedram, M. Alishahi, and K. Badie, *Data quality improvement using fuzzy association rules*, 2010, p.468-472.

56. Chen, M., M. Song, J. Han, and E. Haihong, *Survey on data quality* in World Congress on Information and Communication Technologies (WICT), 2012, p.1009–1013.

57. Glowalla, P., P. Balazy, D. Basten, and A. Sunyaev, *Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model* in 47th Hawaii International Conference on System Sciences (HICSS), 2014, p. 4700–4709.

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

59. L. Cai, L. and Y. Zhu, *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*. Data Sci. J., 2015. 14(0), p. 2.

60. Krogstie, J. and S. Gao, *A semiotic approach to investigate quality issues of open big data ecosystems* in Information and Knowledge Management in Complex Systems, K. Liu, K. Nakata, W. Li, and D. Galarreta, Editors. Springer. 2015, p. 41–50.

61. Rahm, E. and H. H. Do, *Data Cleaning: Problems and current approaches*, IEEE Data Eng Bull, 2000. 23(4) p. 3–13.

62. Eckerson W. W., *Data quality and the bottom line: achieving business success through a commitment to high-quality data.* Data Warehousing Institute, 2002.

63. Fan W. and F. Geerts, *Foundations of data quality management.* Morgan & Claypool, 2012.

64. Fan, W., F. Geerts, X. Jia, and A. Kementsietsidis, *Conditional functional dependencies for capturing data inconsistencies. ACM TODS, 2008,* 33(2).

65. Grijzenhout S., and M. Marx, *The quality of the XML web*. CIKM, 2011. p.1719-1724.

66. Gartner: *Forecast 2007 Data quality tools worldwide*. 2006-2011 Technical report, Gartner.

67. Yeh, P. Z. and C. A. Puri, An efficient and robust approach for discovering data quality rules in 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2010. 1, p. 248–255.

68. Floridi, L. *Big data and information quality* in The Philosophy of Information Quality, L. Floridi and P. Illari, Editors. Springer. 2014, p. 303–315.

69. Zhou, H., J. G. Lou, H. Zhang, H. Lin, H. Lin, and T. Qin, *An Empirical Study on Quality Issues of Production Big Data Platform* in IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE*)*, 2015, 2, p. 17–26.

70. Becla, J., D.L. Wang, and K.T. Lim, *Report from the 5th workshop on extremely large databases.* Data Sci. J. 11, 2012, p. 37–45. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84859721986&partnerID=40&md5=fc683361d4e5427bd6fe1780713b0c51

71. McGilvray, D., *Executing data quality projects: Ten steps to quality data and trusted information*: Morgan Kaufmann, 2008.

72. Sidi, F., 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 (CAMP)*, 2012, pp. 300 –304.

73. Jayawardene, V., S., Sadiq, and M. Indulska, *An analysis of data quality dimensions*, 2015, (1-32).

74. Scannapieco M., P. Data quality at a glance, Datenbank-Spektrum, 2005, 14, p. 6–14.

75. Batini C., M. Palmonari, G. Viscusi, *Opening the closed world: A survey of information quality research in the wild* in The Philosophy of Information Quality. Springer. 2014, p. 43–73.

76. Loshin D., *The practitioner's guide to data quality improvement*. Elsevier, 2011.

77. Rahm, E. and H. H. Do, *Data Cleaning: Problems and current approaches*, IEEE Data Eng Bull, 2000. 23(4), p. 3–13.

78. Oliveira P., *A formal definition of data quality problems in* IQ, 2005.

79. Laranjeiro, N., S. N. Soydemir, and J. Bernardino, *A survey on data quality: Classifying poor data* in IEEE21st Pacific Rim International Symposium on Dependable Computing *(PRDC)*, 2015, p. 179– 188.

80. Ballou, D.P., H. L. Pazer, *Modelingdataandprocessqualityinmulti-input, multi-output informationsystems.* Manage.Sci., 1985, 31(2),150–162.

81. Wang, D. P., H. Pazer, G. K. Tayi, Modeling information manufacturing systems to determine information product quality. Manage.Sci., 1998, 44(4), 462–484.

82. Pipino, L.L., Y. W. Lee, and R. W.Wang, *Data quality assessment*. Commun.ACM, 2002, 45 (4), p. 211–218.

83. Redman, T.C., *Data Quality for the Information Age*. Artech House Publishers, Norwood, MA, 1996.

84. Wand,Y., and R. Y. Wang, *Anchoring data quality dimensions in ontological foundations.* Commun. ACM, 1996, 39(11), p. 86–95.

85. Wang, R.Y., and D. M. Strong, *Beyond accuracy: What data quality means to data consumers*. Journal of Manage.Inf.Syst., 1996, 12(4), p. 5–33.

86. Lee, Y.W., D. M. Strong, B. K., Kahn, R. Y. Wang, *AIMQ: A methodology for information quality assessment.* Inf.Manage. 2002, 40 (2), p. 133–146.

87. Wang, K. Q., S. R. Tong, L. Roucoules, and B. Eynard, *Analysis of data quality and information quality problems in digital manufacturing*,2008, pp. 439-443.

88. Bovee, M., R. P. Srivastava, and B. Mak, *A conceptual framework and belief function approach to assessing overall information quality*, International Journal of Intelligent Systems, 2003, 18(1), p. 51–74.

89. Fan, W. F. Geerts, and J. Wijsen, *Determining the currency of data*," ACM Trans. Database Syst. TODS, 2012, 37(4), p. 25.

90. Saha, B. and D. Srivastava, *Data quality: The other face of Big Data* in *IEEE 30th International Conference on Data Engineering (ICDE)*, 2014, p. 1294–1297.

91. Tang, N. *Big Data Cleaning* in Web Technologies and Applications, L. Chen, Y. Jia, T. Sellis, and G. Liu, Editors. Springer. 2014, p. 13–24.

92. Bhatia, S., J. Li, W. Peng, and T. Sun, *Monitoring and Analyzing Customer Feedback Through Social Media Platforms for Identifying and Remedying Customer Problems,* 2013, p. 1147–1154.

93. Immonen, A., M. Palviainen, and E. Ovaska, *Requirements of an Open Data Based Business Ecosystem*, 2, 2014.

AUTHORS PROFILE