Big Data Quality - Towards an Explanation Model in a Smart City Context

MAKUS HELFERT, Dublin City University, Ireland
MOUZHI GE, Free University of Bozen-Bolzano, Italy

-Research in Progress-

In this paper we present initial research to develop a conceptual model for describing data quality effects in the context of Big Data. Despite the importance of data quality for modern businesses, current research on Big Data Quality is limited. It is particularly unknown how to apply previous data quality models to Big Data. Therefore in this paper we review data quality research from several perspectives and apply the data quality model developed by Helfert & Heinrich with its elements of quality of conformance and quality of design to the context of Big Data. We extend this model by analyzing the effect of three Big Data characteristics (Volume, Velocity and Variety) and discuss its application to the context of Smart Cities, as one interesting example in which Big Data is increasingly important. Although this paper provides only propositions and a first conceptual discussion, we believe that the paper can build a foundation for further empirical research to understand Big Data Quality and its implications in practice.

• **Information systems→Database management system engines, Information systems → Data warehouses**

Additional Key Words and Phrases: Data Quality, Data Quality Effects, Big Data, Data Analytics

1. INTRODUCTION

Big Data has received increasing attention in recent years, as organizations and cities are dealing with tremendous amounts of data. This data are fast moving, often changing in value, meaning and format, as well as can originate from various sources such as social networks, unstructured data from different websites or raw feeds from sensors. Thus Big Data presents us with a new challenge to ensure Data Quality in these environments. Big Data practitioners are experiencing a huge number of data quality problems, which can be time-consuming to solve or even lead to incorrect decisions. As [Warden 2011] stated "I probably spend more time turning messy source data into something usable than I do on the rest of the data analysis process combined". Therefore Big Data Quality (BDQ) should be one of the critical issues related to Big Data research and its applications. Since Big Data creates not only value in financial terms but also in terms of operational and strategic advantages [Haug and Arlbjørn 2010], exploring the value of Big Data and its quality management is crucial to the success of world-leading organizations and enterprises. This is particularly important within a Smart City context, as innovative cities "use ICTs and other means to improve quality of life, efficiency of urban operations and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social, environmental as well as cultural aspects" [ITU-T FG-SSC 2014].

Big Data is typically characterized by volume, velocity and variety [Laney 2001]. As a consequence research should investigate the influence of these characteristics to Data Quality, summarized as Big Data Quality (BDQ) research. Let us illustrate the challenge with an example from a Smart City context, in which many sensor data are used for planning and decision making. The data in Smart City applications are usually characterized by high volume, velocity and variety. In this environment, for example, higher data velocity can result in frequent changes in data specification. In a traffic surveillance information system, the traffic camera is taking a photo every 5 minutes. The data specification of photo quality is set to be 300 dpi. Any photo whose resolution is lower than 300 dpi (due to for example camera equipment, data transfer or weather condition) will be considered as low quality data. When time interval between two photos shots becomes 2 minutes or less, the data specification of photo

quality may be lowered, due to the fact that a constant flow of traffic photos can be analyzed in a different way than single photo shots. In this case, data specification can be affected by the data velocity. However, the relationship between the three typical Big Data characteristics and Data Quality is yet not well investigated.

Aim of this research-in-progress paper is to describe conceptual BDQ and the relationship between Big Data and Data Quality. This helps to derive indications for managing the value of Big Data. We believe that the relationship between Big Data characteristics and the value of the Big Data can be connected. This paper therefore investigates the relationship between the three Big Data characteristics from a quality lens. We have studied how to adapt traditional data quality research model in the context of Big Data. The Helfert & Heinrich model, with its elements conformance to specification and design, is the underpinning model in this paper. Each of the characteristics in Big Data may affect the quality model and accordingly cause different quality problems.

The rest of this paper is structured as follows. Section 2 presents a theoretical grounding for data quality research in the context of Big Data. We further model the BDQ by incorporating the quality concepts of conformance to specifications and conformance to design. Subsequently we examine the impact of data quality on Big Data and provide insights and discussions on how to manage the value of Big Data.

2. THEORETICAL GROUNDING

In the following we ground our work by relating the concepts of Data Quality, Big Data and Value Chains. Several previous contributions have highlighted the positive correlations between high data quality and the value of information [Chen, Chiang, & Storey, 2012]; however again the value contribution of Data Quality in a Big Data environment is little understood. In this section we first discuss the usage and value of Big Data from a value chain perspective, and then relate Big Data to the topic of Data Quality. For an overview on common Data Quality concepts and approaches we refer to [Sadiq 2013]

2.1 Big Data and Value

Many authors refer to Big Data with the characteristics of volume, variety and velocity [Laney 2001]. In this paper, we define the three characteristics of Big Data as follows. Data volume refers to the considerable amount of data supply. Data variety means that the data are structured by different types from various sources. Data velocity indicates the speed of the new data is captured. Big Data analytics is using data for decision purposes. For example, an enterprise can use some data mining tools or analytical methods to find possible opportunities to increase firm performance. Big Data Analytics is particular important in a Smart City context, to analyze for example traffic patterns or environmental data.

Following the concept of [Porters' 1998] value chain, [Miller and Mock 2013] propose a value chain for Big Data. The chain includes three main steps of data discovery, data integration and data exploitation. This corresponds well to the traditional view in Data Quality of an information manufacturing system, transforming raw data into useful information [Wang 1998]. Also [Chaffey and Wood 2005] propose a similar model that focuses on the transformation from data to information to knowledge to action and then to results (DIKAR Model). Approaches to assess Data and Information Quality in information manufacturing systems have been proposed by e.g. [Ge & Helfert 2008]. Following form the Big Data Value Chain concept introduced above we view the concept of Big Data from an Information Manufacturing point of view, from data gathering to its final data usage and value creation. This relates to our earlier work, in which we developed an integrated

framework for IS/IT business value [Borek et al. 2011]. The framework relates resources and capabilities to IS/IT utilization in form of decisions and business value.

In contrast to traditional information systems environments, the value creation often comes from utilizing external information in the context of Big Data. These external data may be highly valuable for corporate decision-making or accumulating business knowledge [Chen et al., 2012]. It is particularly important in a Smart City context, in which the entire city could be a source of valuable data. In this regard Smart Cities are complex information manufacturing and value systems.

2.2 Contextual View of Data Quality

In order to adapt data quality concepts in the context of Big Data, we have reviewed data quality from a number of views. A classic definition of data quality is "fitness for use", i.e. the extent to which data successfully serves the purposes of users [Wang and Strong 1996]. Such a definition implies that the concept of Data Quality is contextual and subjective. It is highlighted for instance by dimensions relevance, believability, or usefulness that are highly contextual and subjective.

[Wang 1998] argues that data producing processes can be viewed as producing data products for data consumers, a view shared by many others. However, according to [Watts et al. 2009], many data quality assessment models have tended to ignore this, and the impact of contextual quality on information use and decision outcomes. Following this contextual view of Data Quality, researchers posit that it is the customer who is the ultimate judge of its quality. However, at the same time many researchers have acknowledged the difficulties to measure the extent to which a product/service meets and/or exceeds the customer's expectation. Since different customers may value different aspects of product or service attributes, considering the various quality expectations is challenging. Researchers have therefore argued, to distinguish between quality of design and quality of conformance in the Quality Management literature [Gilmore 1974], but also for Data Quality [Helfert & Heinrich 2003]. Considering both aspects (see section 3), we consider high BDQ as "data that conforms to data specifications and meets the user's expectations". More database and technical perspectives on data quality can be found for example in [Hoxmeier, 1998; Kim et al., 2003].

3. IMPACT OF BIG DATA ON DQ - FIRST CONCEPTUAL MODEL

In order to develop a conceptual model, we build on our earlier work [Helfert & Heinrich 2003], in which we propose a model to describe the impact of DQ on customer relationships. The model argues for distinguishing between quality of design and quality of conformance in Data Quality research. The view of quality of conformance has been widely accepted in general quality literature for some time. It allows that organizations can determine the quality of products by measuring how well the product conforms to an established specification. On the other hand, conformance to design considers the customer's expectation as the center of quality and meeting and/or exceeding the customer's expectation is critical to conformance to design [Gronroos 1983]. By adopting both quality views, we propose a first conceptual model to describe impacts of Big Data characteristics on BDQ and business value.

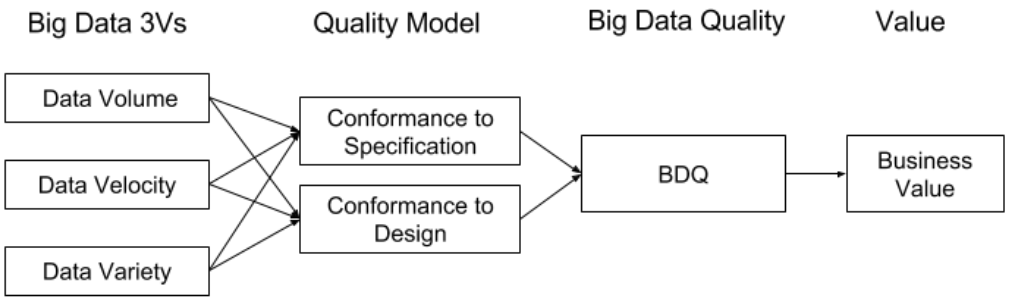

Figure 1, conceptual model for impact of Big Data Quality

In order to describe this relation, we use our earlier model with two aspects:

(1) Quality of design: a (standardized) quality function of a data user u at time t as $Q_{t,u}^{design}$ (I_t^{spec}, $I_{t,u}^{demand}$) $\in$ [0;1] describing the relation between demand and specificaiton, whereby the value 0 represents no quality and the value 1 represents maximum quality.

(2) Quality of conformance: $Q_t^{conform}$ (I_t^{spec}, I_t^{supply}) $\in$ [0;1] that describes the relation between specification and data provided. This function is independent from the data user, whereby the value 0 represents no quality and the value 1 represents maximum quality.

I_t^{spec}, $I_{t,u}^{demand}$ represent architectural descriptions of the demand and specification, whereas I_t^{supply} provides a measure for the information systems operations. $Q_{t,u}^{design}$ and $Q_t^{conform}$ measures the correspondence between its attributes. In other words, $Q_{t,u}^{design}$ describes the gathering of user requirements thus *user dependent*, and $Q_t^{conform}$ the implementation *and* operations of the information system according to some specification in form of models, architectures and processes; thus *user independent*. Data quality management aims to maximize both: consolidate requirements from various users and incorporate these into specifications; as well as ensuring that the information system fulfills the specification. In our model, the objective of data quality management is to maximize the total quality Q_t^{total} over all application areas, which can be described with the optimization variables I_t^{spec}, I_t^{supply} and $I_{t,u}^{demand}$ [Helfert & Heinrich 2003].

A good example of how this concept is used in IT Service management is the approach of ITIL as a set of practices for IT service management that focuses on aligning IT services with the needs of business [Cabinet Office, 2011]. Through a lifecycle concept, that is fundamental in ITIL, it translates strategy and demand into Service Designs. The phases of service transition and service operations translate the design in information systems implementations and operations.

As our previous works shows, the model can be useful to examine relation between the various elements of Data Quality. Generally, in Information Systems environments, it can be assumed that increasing I_t^{spec} results in higher Q_t^{design} and increasing $I_{t,u}^{demand}$ results in lower Q_t^{design} (exceptions have to be considered at a later stage). Similarly this applies to quality of conformance $Q_t^{conform}$, whereby increasing I_t^{spec} results in lower $Q_t^{conform}$ and increasing I_t^{supply} results in higher $Q_t^{conform}$.

However, how do Big data characteristics of volume, velocity and variety affect the underpinning relationships? Therefore, we have adopted the model and conceptually examined the effects of the 3 characteristics of Big Data (Volume, Velocity and Variety) on the quality function. In this article we do not consider any variations of

data demand I_t^{demond}, as we assume that demand is given at a certain point in time. In a first step, hypothetically we have postulated the relationships in Table 1.

In this situation, BDQ improvements can be done:

(a) by improving the specification I_t^{spec} or
(b) by increasing the data provided I_t^{supply}.

Referring back to our example of traffic management within a Smart City, we can identify following two elements to consider. On the one hand, we have standards, interfaces, and architectures and data schemas together with data gathering policies and routines defined. On the other hand, we have the operations together with infrastructure such as sensors, cameras, networks and servers as well as software components for the data extraction, integration and analyses. First specifies I_t^{spec}, whereas later refers to I_t^{supply}.

In case (a), if we assume to improve the specification, we should improve quality of design; however quality of conformance maybe impacted as it requires more sophisticated data specification design and more frequent updates. This can be done by precisely capturing the data demand in a structured way and regularly updating the specification based on the data velocity.

For case (b), we consider increasing the quality of conformance by improving the quality of data supply. However, this is not necessarily the case in Big Data environments. Increasing Big Data does not necessarily increases the overall BDQ. For example, variety means the different types of data from various sources. Once the data variety increases, the data consistency becomes more dynamic and determining the consistency can be also complicated. As a consequence this means that Big Data not necessarily just means we have more data I_t^{supply} – thus better, but actually it depends on the data specification as well as data demand. Increasing the data volume can fulfill the data shortage or overfill the data according to the I_t^{spec} . Compared to traditional data quality that may only design the I_t^{spec} once, BDQ may require more updates on data specification. This can effect of the life cycle of data timeliness. Furthermore,

Table 1: Impact of Data Quality in Big Data

Big Data	I_t^{supply}	I_t^{spec}	Impact on BDQ
Volume Increase	Increase	-	$Q_t^{conform}$ more challenging
Velocity Increase	-	Need for more I_t^{spec} updates	$Q_t^{conform}$ and Q_t^{design} more challenging
Variety Increase	-	Complexity of I_t^{spec} increases	Q_t^{design} more challenging

In contrast to traditional Information environments, Table 1 show the propositions when we move towards a Big data environment, with increases in Volume, Velocity and Variety.

In this simple case of an increase in Data Volume, once would assume that the data supply I_t^{supply} should increase. However, this does not necessarily mean an increase in overall quality. As a consequence of the increased volume, assuring $Q_t^{conform}$ becomes more challenging (resulting of increased data transfers, data records

and potential inconsistencies, inaccuracies and incompletes). In our example related to Smart Cities, as the volume of traffic senor data and video streams increase, misreading and inconsistencies need to be detected, and thus quality assurance mechanism and data cleansing have to be increased accordingly.

In the second case, with an increase in Velocity of Big Data, we argue that the data specification requires more frequent updates as well. This is due to the fact that velocity includes data velocity as well as velocity of the data meaning and format. Changes in the data usually results in some delays of updating the system for data value as well as format and meaning. Indeed $Q_t^{conform}$ and Q_t^{design} are more challenging to address. In our Smart City example assume that velocity of traffic sensor data increases. As a result the quality requirement of the data should be changed such as less resolution for pictures and less strict rules between data capturing intervals. These changes should be captured in the data specification. However, when the velocity of traffic sensor data changes, there might be a delay for updating the data specification, and the system may filter out the useful traffic data with the current data velocity.

In the third case, if we assume that Variety of Big Data is increased, it will result in a more complicated I_t^{spec} and the designing Q_t^{design} gets more challenging. In our example related to Smart Cites, the traffic data can be obtained from sensors, traffic cameras, driver's report over telephone or Internet, or certain notification from a construction site. The data from different sources with different formats can be either structured or unstructured. Thus when the variety increases, I_t^{spec} becomes more complicated and Q_t^{design} gets more challenging due to potential data inconsistency.

From this table, we can see that when the three characteristics of Big Data change, how they impact the quality of conformance and design. It provides a theoretical guideline for Big Data practitioners to assure the DBQ. From our discussion, it can be seen that when measuring BDQ, the data quality criteria that are used to measure the traditional data quality will vary. Therefore it is critical to concern the feasibility of BDQ model when using big data analytics to create business value. In future research our theoretical concept can help to illustrate this effect related to data quality criteria.

4. CONCLUSIONS

In this research-in-progress paper we have presented data quality model in the context of Big Data. We have described the concept of BDQ and the Big Data value chain. The issues have been related to theoretical perspectives of data quality and the resource-based view on organisations. Following the work from [Helfert & Heinrich 2003] we have introduced a conceptual model that differentiates between quality of design and quality of conformance. This model has been applied and described within the context of BDQ. We argue that the three Vs of Big Data (Volume, Velocity and Variety) impact Data Quality and in turn the Value of Big Data. By applying the theoretical model, we have discussed how Data Quality changes in the context of Big Data. Furthermore, given the nature of Big Data, some data quality dimensions become more important in Big Data, like data consistency and concise presentation. Although many of these dimensions have been discussed in literature [Wang & Strong 1998], little research or insights into dimensions and DQ in the Big Data has been conducted yet. Thus some dimensions in DQ like completeness, timeliness, need to be re-considered or re-defined.

In our further research we aim to develop an experiment to refine and test this theoretical model. We plan to use virtual machines and build a Big Data infrastructure to test the proposition in Table 1. Some virtual machines will act as

data sources generating large amount of "real-time" sensor data. We are able to set parameters such as volume, velocity and variety for the data generation. Other virtual machines will be used to integrate data and analyze the data by, for example, a Hadoop cluster and analytics tools. Using a common approach to assess data quality, see for example [Helfert, et al. 2009] we are able to assessing the impact of data quality on the generated Big Data.

ACKNOWLEDGMENTS

This work was partly supported with the financial support of the Science Foundation Ireland grant 13/RC/2094 to Lero - the Irish Software Research Centre (www.lero.ie).

The authors would like to thank the anonymous reviewers for their insightful reviews.

REFERENCES

Borek, A., Helfert, M., Ge, M. and Parlikad, A.K.N. 2011. An information oriented framework for relating IS/IT resources and business value. In Proceedings of the International Conference on Enterprise Information Systems (ICEIS). Beijing, China

Cabinet Office (2011), ITIL Service Strategy. Service Design, Service Transition, Service Operation,Continual Service Improvement, TSO, London.

Cappiello, C., Francalanci, C. and Pernici, B. (2004), Time-related factors of data quality in multi-channel information systems, Journal of Management Information Systems, 20(3), pp. 71-91.

Chaffey and Wood (2005) Chaffey D., Wood S. Business Information Management: Improving Performance Using Information Systems, Pearson Education, 2005

Chen, Chiang, & Storey, 2012 Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Quarterly 36(4):1165–1188.

Gilmore, H.L. (1974), Product conformance cost, Quality Progress, 7, pp. 16-19.

Gronroos, C. (1983) Strategic management and marketing in the service sector. Marketing Science Institute Massachusetts, USA.

Ge M, Helfert M: Data and Information Quality Assessment in Information Manufacturing Systems Business Information Systems, Volume 7, Lecture Notes in Business Information Processing (2008), pp 380-389.

Helfert, M.; Heinrich, B.: Analyzing Data Quality Investments in CRM – A model-based approach, in Eppler, M.J., Helfert, M (eds): Eighth International Conference on Information Quality (IQ 2003), 7th to 9th November, MIT Sloan School of Management, Cambridge, pp. 80-95.

Helfert, M., Foley, O. Ge, M., Cappiello, C.: Analysing the effect of security on information quality dimensions. In 17th European Conference on Information Systems (Newell S, Whitley EA, Pouloudi N, Wareham J, Mathiassen L eds.), 2785-2797, Verona, Italy. (ISBN 978-88-6129-391-5), 2009.

Hoxmeier, 1998; Hoxmeier, J.A. (1998), "Typology of database quality factors", Software Quality Journal, Vol. 7, Nos 3-4, pp. 179-93.

Haug, A,, Arlbjørn J.S. (2010) Barriers to master data quality, Journal of Enterprise Information Management. Vol. 24 No. 3, pp. 288-303

ITU-T FG-SSC 2014 Smart sustainable cities: An analysis of definitions, TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU (10/2014), available at https://www.itu.int/en/ITU-T/focusgroups/ssc/Documents/Approved_Deliverables/TR-Definitions.docx

Kim et al., 2003 Kim, W., Choi, B-J., Hong, E-K., Kim, S-K. and Lee, D. (2003), "A taxonomy of dirty data", Data Mining and Knowledge Discovery, Vol. 7 No. 1, pp. 81-99.

Kwon 2014 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.

Laney, D. (2001), '3D Data Management: Controlling Data Volume, Velocity, and Variety' , Technical report, META Group .

Miller and Mock (2013) H. Gilbert Miller, Peter Mork: From Data to Decisions: A Value Chain for Big Data. IT Professional 15(1): 57-59 (2013)

Sadiq, S. (2013), Handbook of Data Quality, Research and Practice, 2013, Publisher: Springer

Wang, R.Y. and Strong, D.M. (1996), Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, 12(4), pp. 5-34.

Warden P. (2011) Big Data Glossary, O'Reilly publishing.

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