Understanding data quality

Fu, Qian; Easton, John

DOI:
10.1109/BigData.2017.8258380

Document Version
Peer reviewed version

Citation for published version (Harvard):
Fu, Q & Easton, J 2018, Understanding data quality: ensuring data quality by design in the rail industry. in *Proceedings of the 2017 IEEE International Conference on Big Data (BIGDATA).* IEEE Xplore, pp. 3792-3799, 2017 IEEE International Conference on Big Data, Boston, Massachusetts, United States, 11/12/17. https://doi.org/10.1109/BigData.2017.8258380

Understanding Data Quality

Fu, Qian; Easton, John

DOI:
10.1109/BigData.2017.8258380

Document Version
Peer reviewed version

Citation for published version (Harvard):
Fu, Q & Easton, J 2018, Understanding Data Quality: Ensuring Data Quality by Design in the Rail Industry. in Proceedings of the 2017 IEEE International Conference on Big Data (BIGDATA). IEEE Xplore, pp. 3792-3799, 2017 IEEE International Conference on Big Data, Boston, United States, 11/12/17. DOI: 10.1109/BigData.2017.8258380

Understanding Data Quality

Ensuring Data Quality by Design in the Rail Industry

Qian Fu* and John M. Easton
Birmingham Centre for Railway Research and Education
School of Engineering, University of Birmingham
Birmingham, B15 2TT, UK
*Email: q.fu@bham.ac.uk

***Abstract*—The railways worldwide are increasingly looking to the integration of their data resources coupled with advanced analytics to enhance traffic management, to provide new insights on the health of infrastructure assets, to provide soft linkages to other transport modes, and ultimately to enable them to better serve their customers. As in many industrial sectors, over the past decade the rail industry has been investing heavily in sensing technologies that record every aspect of the operation of the railway network. However, as any data scientist knows, it does not matter how good an algorithm is, if you put rubbish in, you get rubbish out; and as the traditional industry model of working with data only within the system that it was collected by becomes increasingly fragile, the industry is discovering that it knows less than it thought about the data it is gathering. When coupled with legacy data resources of unknown accuracy, such as design diagrams for assets that in many cases are decades old, the rail industry now faces a crisis in which its data may become essentially worthless due to a poor understanding of the quality of its data. This paper reports the findings of the first phase of a three-phase systematic review of literature about how data quality can be managed and evaluated in the rail domain. It begins by discussing why data quality matters in a rail context, before going on to define the quality, introduce and expand the concept of a data quality schema.**

Keywords-data quality; rail; quality by design; data quality schema.

I. INTRODUCTION

As the global railway industry moves towards a more data-driven decision making culture, an increasing proportion of operational and investment decisions are being based on facts and figures derived from existing data held within the industry, by supply chain, and in public repositories. Although the famous adage that "facts never lie" is largely true, the financial implications of incorrect decision making in an industry as infrastructure-heavy as the railways are substantial. There is thus a need for an effective integration of relevant data from heterogeneous (re-)sources to build sufficient understanding of all stakeholders on the value of the data available. On this basis, it will enhance the stakeholders' ability and confidence in answering specific questions before any decision is made. Ensuring data quality is a key element of ensuring the easy integration of data across system and organisation boundaries; the National Institute of Science and Technology (NIST) reported that in 2002 alone the US capital facilities industry could have saved around $15.8 billion (about 1% to 2% of the annual revenue) through improved information interoperability [1], so there is a substantial value proposition from building a better understanding of industry data quality. Indeed, if the NIST figure was translated into the context of the UK railways, this would have amounted to potential savings of between £82 million and £164 million for the financial year 2013/2014 (see [2]).

Looking forward, it will be important for the rail industry to be able to produce clear plans for its investments in data resources – a capability rooted in a clear understanding of the "fitness of purpose" of data resources for future tasks. Delivering this capability will require a shift from "find and fix" approaches to data management (e.g. [3]), to a "predict and prevent" based set of tools and processes, which allow data users to manage and utilise data resources in a more proactive way and with more confidence. This in turn relies on delivering enhanced data quality throughout a typical life cycle (including collection, processing, maintenance, and disposal) of data at reasonable cost and with a consistent quality level. Based on these facts, a three-phase systematic literature review is designed to investigate methodologies and toolkits for dealing with data quality in various fields, and to discuss them with an eye towards their feasibility of implementation within the railway industry. The techniques, which could be adapted from the other fields, will provide constructive suggestions enabling data quality management in the rail industry. The ultimate objective is to make recommendations for improvement and further development of the current information systems within the rail industry.

This paper presents the findings of the first phase of the complete systematic review; the second phase, which is nearing completion at the time of writing will report on a sociotechnical perspective in data quality design in the context of information and communications technology applications in the rail industry; and the third phase, which is expected to be completed by Christmas 2017, will wrap the topic up by discussing how data quality, and hence value in a given context, evolve over its lifecycle, turning data from a poorly-understood financial burden into an asset that can be commoditised by the industry.

II. CONTEXT

Recent years have seen a prolific growth of specialist software in virtually every field of human activity, including of course, railway industry. This fact has not only presented unprecedented opportunities for development of data science, but also considerable challenges in terms of how data and its quality are viewed (i.e. interpreted, designed, managed, evaluated, or utilised) by its users (cf. [4]). In a safety-critical industry such as rail, problems caused by poor quality of data in data applications could potentially lead to irreversible loss or even disasters (see, for example, Fisher and Kingma [5], who reviewed, particularly from the standpoint of data quality, the causes of two tragic cases that had occurred in aerospace and aviation industry). It is therefore crucially important that the relevant decision makers have clear methods and an integrated process to:

- guarantee that gathered data meets well-defined quality levels, ideally, in the very first instance;
- enable and support staff in the selection of the most appropriate tools in their interactions with the held data resources; and
- ensure the inherent "value" of the data, which will inevitably change over time as the industry evolves (cf. [6]), is correctly understood.

The demanding requirements set out above would, in turn, call for greater attention to the needs of enhanced quality of the data itself. In this context, the paper aims to understand both theoretical and pragmatic perspectives on how issues regarding data quality may be approached by dint of an original design of data itself.

The remainder of this paper consists of three sections. Section III reviews and discusses definitions and schematic views of data from a quality perspective. Section IV further investigates what data quality is and how we may attain a predefined structure (also known as a schema) of data quality for its assessments and enhancement. Lastly, some concluding remarks are presented in Section V.

III. DATA AND DATA QUALITY

A. Definitions of Data

The definition of data may vary depending on the context in which it is discussed. Given different standpoints adopted by relevant literature published around 1980s, Fox, et al. [7] reviewed several ways of describing what data was. It was noted from the earlier studies that the definition of data was not often provided in an explicit fashion. In general, data was defined, given the underlying nature, as a set of collected facts, which could be stored, and may convey or develop useful information to its consumers (cf. e.g. [8, 9]). In that way, it raised two issues with respect to the literal meaning. Firstly, it implied that "information" was a consequence of processing "data" and so differentiated from "data" in their roles and status. But it is also arguable that the two terms "information" and "data" could be interchangeable – partly because "data" inherently carries information of its own, and partly because "information" is by definition facts as well. In this regard, information can also be deemed a collection of facts, notwithstanding being converted or transformed or its form of existence; and it, in essence, has the same characteristics as data. In this regard, the terms "data quality" and "information quality" are also interchangeable (see e.g. [10]), but can differ as needed in different stages of data life cycle. Another issue, as pointed out by Fox, et al. [7], was that any "fact" must be a thing that actually exists or is known to have occurred or proved to be true. However, in practice, data would not always fulfil the exact definition of what a fact is. In other words, while data may not always be "facts", defects in data would mostly be unavoidable. But in practical terms, we may simply circumvent the tangled definition since, semantically, a "fact" can sometimes also be hypothetical and hence deemed "a thing that is believed or claimed to be true". Thus, we shall complement the original definition of "facts" with this argument.

Where recorded data does not accurately capture the state of a system, it is believed that the defect in data (or the distorted fact) is principally due to human reasons, including the ways of how data is collected, or stored, or any further processing by its users. In this respect, the definition of data could be expressed in a more empirically-sound manner with involvement of data users (cf. [11]), namely, data is a result of recording (i.e. collecting by measuring and/or observing) the facts of the real world, and in turn represents or reflects the facts and is meaningful to its users. Nevertheless, both of the above approaches to the definition of data are lacking a specification of its storing and physical appearance (cf. [7]). There must be a detailed description of e.g. data types, notational representation and stored formats, which are prerequisites for the recording of data and establishment of data models. This aspect of data had been extensively discussed by e.g. Burch, et al. [12] as well as Stamper [11].

Further to the above, data has also been defined in terms of both its logical and conceptual representation from the perspective of database design (see e.g. [13-15]). A classic example is that data could be represented as a collection of data items (also known as "data points" or "datum"), each individual item being defined by an entity-attribute-value (EAV) triple that describes facts (cf. [16]). This approach is fundamentally reliant on the entity-relationship model that was firstly proposed and elaborated by Chen [17] (see also [18]). In a narrow sense, each of the EAV triples, as shown in Figure 1, is made up of a single "entity" that represents a thing (e.g. an object or an event), a single "attribute" that reflects a facet of the entity, and a single "value" that is recorded as a specific instance of the attribute and hence the entity. Figure 1 illustrates a schematic view of data defined with the basic EAV triple, where different data items are linked by solid lines over their entities, indicating certain relationships between them. The remaining links, each with a circled-end inside a basic EAV triple, indicate affiliation relationships between the elements of the EAV triples, which are essential and mandatory. That is to say, to define an entity would entail specification of its affiliated attribute-value tuple. For example, in Figure 1, the "Entity 1" is uniquely characterised by the "Attribute 1" that takes on the "Value 1".

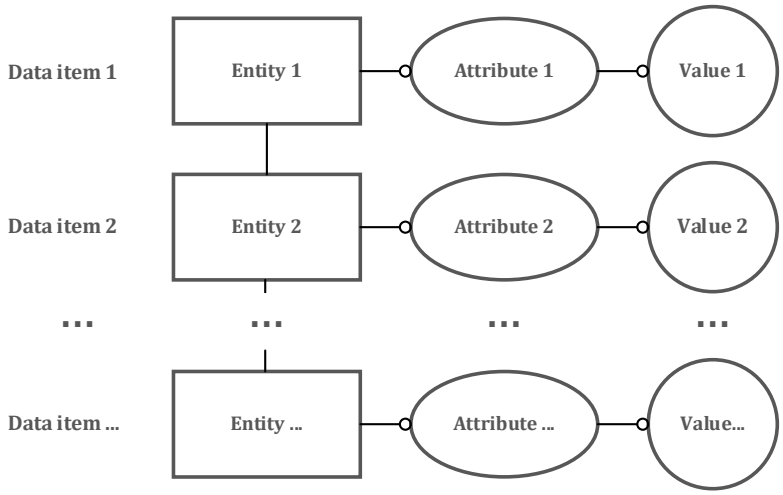

Figure 1. A simplified view of data modelled upon basic EAV triples

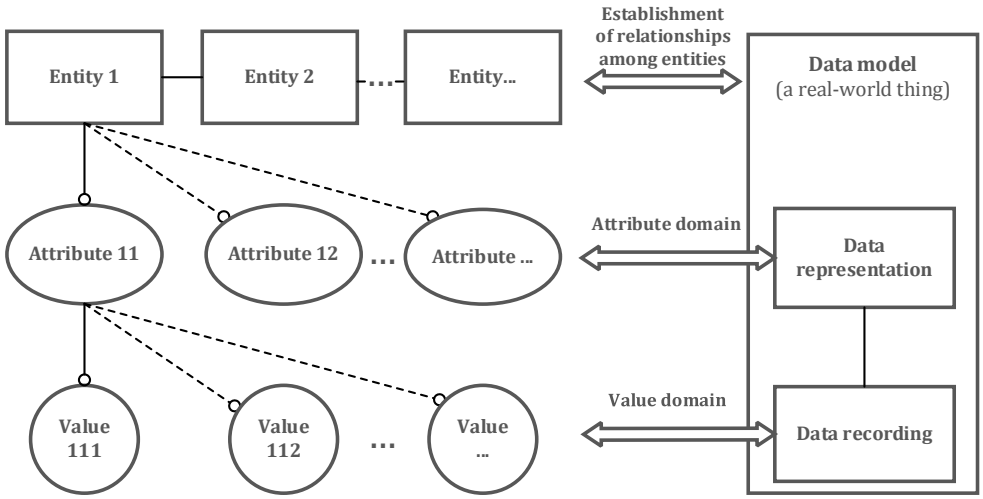

Figure 2. A simplified illustration of data modelled upon generalized-EAV triples

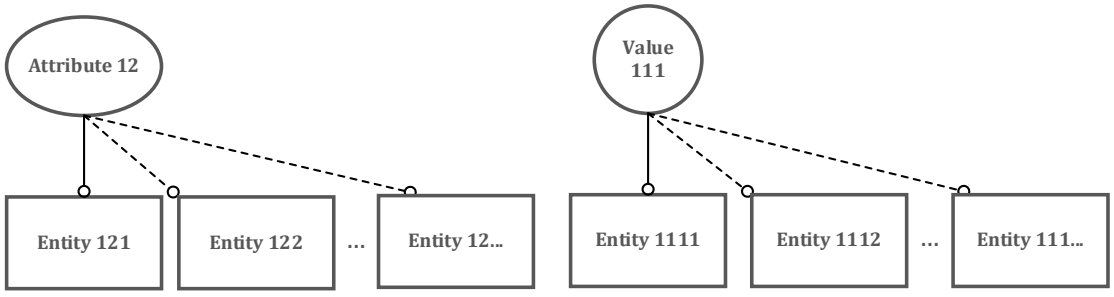

Figure 3. Further simplified illustrations of (left) an attribute domain, and (right) a value domain

The introduction of the notion of "data item" provides a subtle view of "atomic structure" of data, which greatly simplifies the modelling of any piece of a real-world object with the establishment of relationships between defined entities. In the more generalised context, an entity could represent any predefined class, such as a composite of two or more entities, and any individual entity is exclusively identified by a set of attributes or rather, by an attribute domain (cf. [19, 20]). Different entities and their respective attribute domains may share some of the same attributes, but differ in attribute-value tuples, or rather, in attribute-value systems (cf. [13, 21]). Similarly, any individual attribute may have a value domain, which is bound up with one entity described by the attribute. Furthermore, any element in an attribute domain could be a predefined entity (and hence an EAV triple) that is not directly related to the entity over the attribute domain. This also applies to any value domain. So values are collected and stored via practicable methods of observation and/or measurement, on the understanding that those domains and necessary metrics have all been set. We may refer to such an extended EAV triple as a generalised data item (or a generalised-EAV triple). A simplified view of data modelled upon the generalised data items is illustrated in Figure 2; and further examples of an attribute domain and a value domain are illustrated in Figure 3. In these two illustrations, the dashed affiliation links indicate optional attributes and values, or optional relationships (cf. [20]). The generalised data item fully complies with the fundamentals of constructing any entity-relationship data model. In this regard, it must be pointed out that data quality would be pertinent to how different entities are interrelated; but for the current phase of study, the focus is on the EAV triple as a whole. Obviously, the establishment of a data model would entails that any element of a data item, especially the attribute and value domains are predefined by data users in accordance with their requirements for a specific application. Therefore, what is critical is to find out what attributes actually identify an entity and also their respective domains of permissible values. This has also been indicated by double arrows in Figure 2, meaning that it involves interactive processes between data and its users, where issues of data quality arise.

The definitions of data discussed above are summarised in Table 1 below.

B. Definitions of Data Quality

In general, the term, "data quality", is widely used to embody a set of "characteristics" or "facets" of data, such as its accuracy, completeness, consistency, and timeliness (cf. [22, 23]). Any of the many facets can be referred to as a variable, or more commonly, a "dimension" of data quality. In a broader sense, data quality is also supposed to be capable of indicating the degrees to which each of the dimensions conforms to data users' specific requirements in a given context (cf. [24, 25]). A poor (or low) level of data quality can have a severe impact on the overall effectiveness of the corresponding data applications (cf. [26]).

Empirically, given specific requirements of data users, data of various types is collected, stored, and utilised as inputs into an information system designed and implemented for practical applications. After data processing (i.e. a series of operations on the data inputs), the information system should produce desired outputs (or rather, information) for a certain application, and intermediate outputs that serve as "refined inputs" for other intended data applications (cf. [6]). The role that data plays along the path of the process flow may easily be regarded as similar to that raw material does in a production line in manufacturing engineering. In this regard, we may also refer to the whole process of the data flow as a data manufacturing process (cf. [27]).

Despite the comparability to a large extent between data and raw material, there, however, exist intrinsic dissimilarities in terms of their qualities. Firstly, the quality of data can be characterised in a lot more dimensions, for which counterparts may not be found in any kind of raw material; examples of this include accuracy and completeness (cf. [28]), to name a few. Secondly, the totality of data quality dimensions is common and shared among all types of data regardless of how data evolves in its manufacturing process, whereas this may not be universally true with respect to raw material. Further to this, as highlighted by Liu and Chi [6], existence, interpretation and any application of data are completely reliant on "theories", i.e. methods and models that enable data to evolve in its life cycle. Consequently, theories and control techniques that have been fairly well developed for raw material (or product) quality management, though, may not be applied mechanically to manage or analyse data quality without regard to specific life-cycle stages and requirements of data applications (cf. [29]).

Table 1. A summary of definitions of data in terms of its three facets

Facets	**Definitions**	**References**
Intrinsic nature	A set of facts, either realistic or hypothetical, that could be collected, stored and convey/develop useful information to data users	See e.g. [8, 9]
Physical representation	A retrievable form of the real-world facts as a result of measurement and/or observation, which is meaningful to data users	See e.g. [11, 12]
Logical/conceptual representation	A collection of (generalised) data items (i.e. entity-attribute-value triples), which describes/models the real-world facts	See e.g. [14-17]

IV. Dimensions of Data Quality and Associated Design Schema

All the sources reviewed for this paper acknowledge that data quality (hereafter referred to as DQ) is a multidimensional and hierarchical construct [23]. Yet, despite the existence of wide-ranging studies looking into this subject, it seems that there has been an inconclusive discussion about "what constitutes a good set of DQ dimensions" [26]. Uncertainties still remain as to what generic and/or rigid definitions of each DQ dimension are, and how the dimensions are correlated.

A. Dimensions of Data Quality

The four DQ dimensions that are mostly commonly referred to and discussed in the existing studies are accuracy, completeness, consistency, and time-related properties (e.g. timeliness or currency). For each of the four dimensions, Batini, et al. [30] made a comparison between the definitions that were proposed (up to around 2000s) under different contexts by other researchers. Although these dimensions could all be interpreted and understood based on the literal sense of the words, the key differences stemmed from the methods of assessment and measurement, with pertinent regards to the emphasis placed on the dimensions from the viewpoint of the users. For instance, the accuracy dimension of data may be quantified in terms of the extent to which data is precise [31]; while in some other cases, data users' interests are on whether or not the data is correct [22], which instead reduces the quantification problem to a binary question. Another important issue is, as also noted by Batini, et al. [30], that the DQ dimensions can be defined at different levels given the scope of how the relevant term "data" is defined. ranging from a basic data item (e.g. [10]), to data modelled on generalised data items, databases, information systems (e.g. [26]), and even more broadly, data warehouses (e.g. [32]).

For the sake of reducing ambiguity, Data Management Association (DAMA) UK Working Group on "Data Quality Dimensions" [33] put forward a set of core DQ dimensions in view of the group members' best practice, which, in addition to the aforementioned four, further include uniqueness and validity. How the six DQ dimensions are related is demonstrated in Figure 4. As shown in the figure, a dashed link indicate that the two data-quality dimensions connected by it are directly related to each other when being assessed. An arrowed link shows that there can be an implicational relationship between the two relevant dimensions. For example, a data value is accurate only if it is valid in terms of e.g. formats and permissible within the defined value domain, whereas a valid data value may not necessarily be accurate. This conceptual view considers only the multidimensionality of data quality.

It must be pointed out that DQ dimensions should not be only limited to the dimensions mentioned above; however, there can be much more characteristics that affect DQ (see e.g. [34, 35]), such as accessibility, interpretability, relevancy, and redundancy, to name but a few. All these dimensions could be further grouped into several categories, given their

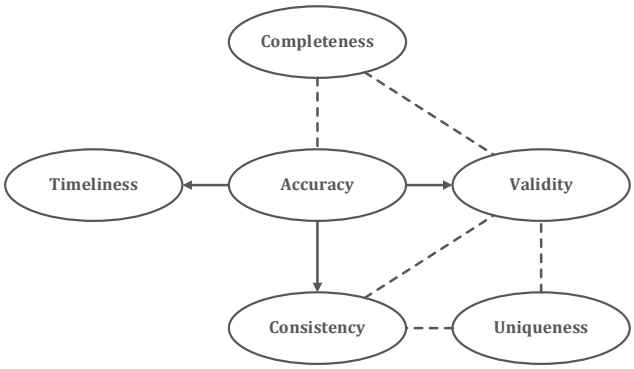

Figure 4. A multidimensional view of DQ, including six primary DQ dimensions as proposed in [33]

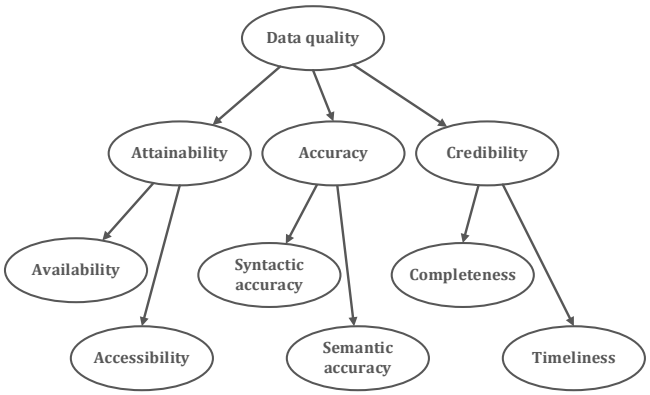

Figure 5. An example of a hierarchical structure of DQ dimensions (adapted from the example in [36])

own intrinsic attributes, requirements and contexts of data applications (cf. [6, 31]). The view of DQ structure depends largely on the understanding of which dimensions or which categories of dimensions are meaningful for data users (cf. [34]), which in turn requires understanding of relationships among DQ dimensions. In this regard, each individual dimension may be further resolved into sub-dimensions, meanwhile, with itself being a sub-dimension affiliated to another of a higher level. For example, the accuracy of DQ may be separated into semantic accuracy and syntactic aspects; and also it is a requisite for data credibility that describes to what extent the data could be believed or trusted. The credibility is treated as a higher-level DQ dimension (or a dimension category) that may contain some more sub-dimensions such as consistency and timeliness. A hierarchical structure of DQ dimensions is illustrated in Figure 5, which is adapted from the example originally provided by Wang, et al. [36].

From the above, it is obvious that the definition of DQ would entail a clear DQ hierarchy, which ought largely to be determined by data users' specific requirements for DQ.

B. Design of Schema for Modelling Data Quality

Based on the entity-relationship model, Wang, et al. [34] proposed a methodology of requirement analysis for DQ. They introduced a notion of DQ-attribute that consists of a

set of DQ indicators and DQ parameters, in order to interrogate DQ from an objective and a subjective perspective, respectively. According to their definitions, a DQ parameter is qualitative, and in essence refers to a DQ dimension as discussed above, e.g. accuracy and credibility, which should be meaningful for a data user; and a DQ indicator is a specific "data value" for the DQ parameter. Given this, we may construct a conceptual view of DQ, as shown in Figure 6 (cf. [36]), in a similar fashion to a data item (cf. Figure 2). Thus, DQ is treated as, and it actually is, data, in that the DQ-attribute, as defined by Wang, et al. [34], could be deemed an "attribute-value" tuple being characterised instead by a parameter-indicator tuple of DQ. The only difference is that each attribute-value tuple acts as a "DQ-entity" of its quality parameter-indicator tuple.

Conventionally, the design of a conceptual schema of a data model for a data application starts with users eliciting application attributes (i.e. data entities and attributes for application) according to the application requirements. For the purpose of incorporating data quality as part of the modelling process, as illustrated in Figure 7, a set of affiliated DQ-attributes need also be identified or determined for each application attribute (cf. [34, 36-38]).

Note, again, that the identification and determination of DQ-attribute are largely subject to users' own consideration. In practice, therefore, there can be a set of alternative views for both the application data and the corresponding DQ. It would then require an approach that could effectively combine all the alternative DQ views into a single DQ schema for the targeted data application. (This issues will be investigated in the third phase of the planned systematic review in the near future.)

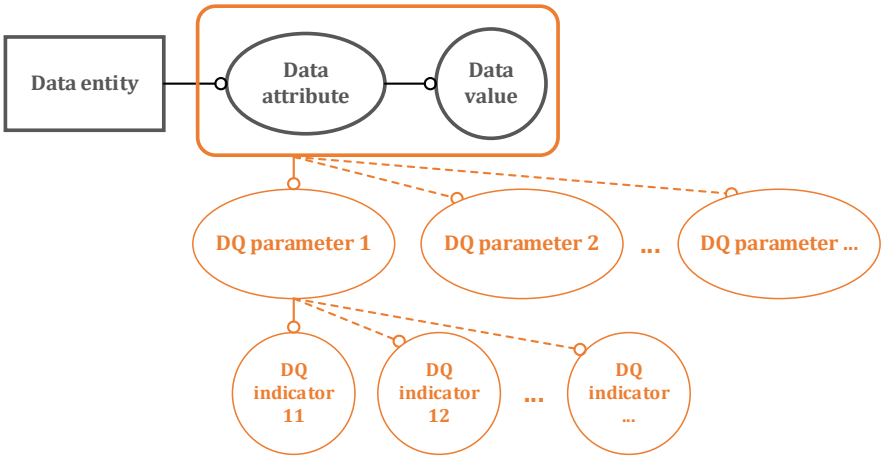

Figure 6. A simple example of data quality view

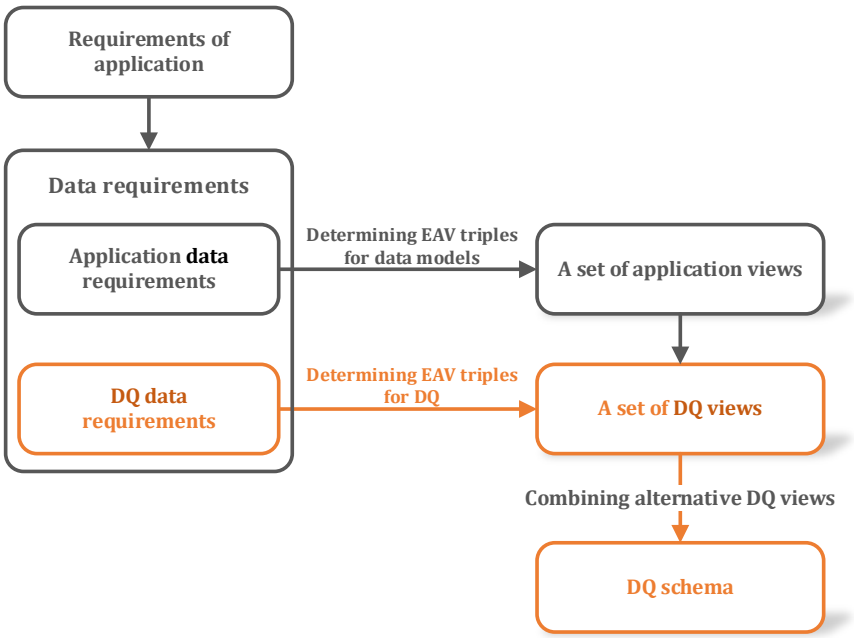

Figure 7. An illustration of the process for deriving a data quality schema embedded in a data model

V. Concluding Remarks

This paper lays a theoretical foundation for the whole systematic review of understanding data quality. Up to the current phase of the planned systematic review, we have revisited a few fundamental considerations and questions about data and DQ, from a wider perspective. Firstly, the definition of "data" has been critically dissected and discussed in terms of three aspects, including its intrinsic nature, physical appearance, and conceptual representation. It is reasonable to state that an integrated view that encompasses all of the three aspects is essential for a decent definition of data and hence any data model (cf. [7]). An in-depth discussion has been presented over the conceptual (basic and generalised) view of a data item; and that would inherently lead to a stratified system in the assessment of its quality. (Issues of measuring and assessing the DQ will be investigated in the third phase of the planned systematic review.) Secondly, definitions for DQ and its dimensions have been discussed. However, there is also a lack of consensus on what constitute a good set of data quality dimensions. It has been particularly noted that many issues of DQ are stem from the interactive processes between data and its users. (The involvement of data users' views deserves extra attention, which will be further investigated in the second phase of the planned systematic review.) In this regard, a DQ perspective is not only to monitor and control quality of data at different stages of a typical data life cycle, but also to prioritise data quality in the original design of data, databases, and even information systems. Primarily, there can be two aspects of the design for DQ. One is the conceptual view of DQ structure, which ought to facilitate assessment and measurement of DQ. Another is related to internal control and correction procedures within a database or an information system, which ought to help ensure quality data at an acceptable level.

A generalised data item would, in some cases, become or be treated as a database. As such, the ways of inspecting the quality of data item would also be applicable for quality assessments of database. It must be noted that many of the reviewed methodologies for assessing DQ are limited in the context of the narrow view of data, especially focused on the quality of data values. It will require quality assessments to be performed at a higher resolution (e.g. [39]), such as for data modelled on the generalised data item. This will therefore refer to the database management that further involves data definition facility.

The "attribute-based approach" proposed by Wang, et al. [36] is a cornerstone. The modelling process entails the specification of the DQ-data that are viewed by the data users as essential for estimating, determining, or enhancing data quality, thus laying a foundation for the development of a quality perspective in database design. Defects in data within an information system may evolve and be propagated; and good data may also be contaminated in various ways [22]. The further encapsulation of DQ-data may assist in, and hence increase information system's capability of, tracing and handling these problems. One issue about this approach needs to be noted. That is, a data application may possess not a single but rather a set of alternative application views or data models (see also Figure 7).

In reality, "error-free" data or information system are hardly likely to be always available to the data users. Nevertheless, any data user may also need to have a rethink of a question, as noted by Ballou and Pazer [40] nearly 30 years ago, that is, whether the gain in enhanced data quality (and even in information system as whole) from additional quality control procedures would be commensurate with the cost incurred to achieve it. For instance, as mentioned by Ballou and Pazer [22], approaches that used to manage the accuracy dimension were often through "extensive and often elaborate edit checks and controls" [41]. Obviously, such approaches would not be able to meet the requirements for managing the large volume of data. Issues as such will be addressed and further discussed in the remaining phases of the planned systematic review. In addition, the concepts and metrologies presented in this paper will also be elaborated in the future studies with specific examples of data applications within the rail industry.

Acknowledgment

This study is funded by Network Rail through their Strategic Partnership in Data Integration and Management with the University of Birmingham. The authors would like to express their gratitude to Network Rail for their continued support.

References

[1] M. P. Gallaher, A. C. O'Connor, J. L. Dettbarn, Jr., and L. T. Gilday, "Cost analysis of inadequate interoperability in the US capital facilities industry," *National Institute of Standards and Technology (NIST)*, 2004.

[2] J. Tutcher, M. Easton John, and C. Roberts, "Enabling data integration in the rail industry using RDF and OWL: The RaCoOn ontology," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, vol. 3, p. F4015001, 2017/06/01 2017.

[3] D. P. Ballou and G. K. Tayi, "Methodology for allocating resources for data quality enhancement," *Communications of the ACM*, vol. 32, pp. 320-329, 1989.

[4] C. R. Adams, "How management users view information systems," *Decision Sciences*, vol. 6, pp. 337-345, 1975.

[5] C. W. Fisher and B. R. Kingma, "Criticality of data quality as exemplified in two disasters," *Information & Management*, vol. 39, pp. 109-116, 12// 2001.

[6] L. Liu and L. Chi, "Evolutional data quality: a theory-specific view," presented at the Seventh International Conference on Information Quality, Boston, MA, 2002.

[7] C. Fox, A. Levitin, and T. Redman, "The notion of data and its quality dimensions," *Information Processing & Management*, vol. 30, pp. 9-19, January/February 1994 1994.

[8] J. P. Fry and E. H. Sibley, "Evolution of data-base management systems," *ACM Computing Surveys (CSUR)*, vol. 8, pp. 7-42, 1976.

[9] H. D. Clifton, *Business data systems: a practical guide to systems analysis and data processing*. Englewood Cliffs, London (etc.): Prentice-Hall, 1978.

[10] M. Bovee, R. P. Srivastava, and B. Mak, "A conceptual framework and belief - function approach to assessing

overall information quality," *International Journal of Intelligent Systems*, vol. 18, pp. 51-74, 2003.
[11] R. K. Stamper, "Towards a theory of information - Information: Mystical fluid or a subject for scientific enquiry?," *The Computer Journal*, vol. 28, pp. 195-199, 1985.
[12] J. G. Burch, G. Grudnitski, and F. R. Strater, *Information systems: theory and practice*, 3rd ed.: John Wiley & Sons, Inc., 1983.
[13] L. W. Barsalou and C. R. Hale, "Components of conceptual representation: from feature lists to recursive frames," in *Categories and Concepts: Theoretical Views and Inductive Data Analysis*, I. Van Mechelen, J. Hampton, R. Michalski, and P. Theuns, Eds., ed San Diego: Academic Press, 1993, pp. 97-144.
[14] J. A. Zachman, "A framework for information systems architecture," *IBM Systems Journal*, vol. 26, pp. 276-292, 1987.
[15] C. Batini, S. Ceri, and S. B. Navathe, *Conceptual database design: an entity-relationship approach*. Redwood City; Wokingham: Benjamin/Cummings Publishing Company, 1992.
[16] D. C. Tsichritzis and F. H. Lochovsky, *Data models*. Englewood Cliffs, New Jersey: Prentice-Hall, 1982.
[17] P. P.-S. Chen, "The entity-relationship model - Toward a unified view of data," *ACM Transactions on Database Systems (TODS) - Special issue: papers from the international conference on very large data bases: 22–24 September 1975, Framingham, MA*, vol. 1, pp. 9-36, 22-24 September 1975 1976.
[18] P. P.-S. Chen, "A preliminary framework for entity-relationship models," presented at the Second International Conference on the Entity-Relationship Approach to Information Modeling and Analysis, 1983.
[19] S. Y. Tu and R. Y. Wang, "Modeling data quality and context through extension of the ER model," presented at the Third Workshop on Information Technology and Systems, Orlando, Florida 1993.
[20] M. West, *Developing high quality data models*: Morgan Kaufmann Publishers Inc., 2011.
[21] T. Jones, "Attribute value systems: an overview," ed. Department of Cognitive Science, University of California at San Diego, 1998.
[22] D. P. Ballou and H. L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," *Management Science*, vol. 31, pp. 150-162, February 1985 1985.
[23] L. Jiang, "Data quality by design: a goal-oriented approach," Ph.D. Thesis, Computer Science, University of Toronto, 2010.
[24] ISO 9000: 2015, "Quality management systems - Fundamentals and vocabulary," in *Terms related to requirement*, ed: International Organization for Standardization, 2015.
[25] IAIDQ. (2017, 02 March 2017). *International Association for Information and Data Quality - IQ/DQ glossary*. Available: http://iaidq.org/main/glossary.shtml
[26] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, vol. 39, pp. 86-95, 1996.
[27] R. Y. Wang, V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, pp. 623-640, August 1995.
[28] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, pp. 58-65, February 1998.
[29] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," *Communications of the ACM*, vol. 40, pp. 103-110, May 1997 1997.
[30] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, pp. 1-52, 2009.
[31] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, pp. 5-33, March 1996.
[32] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, *Fundamentals of data warehouses*, 2nd ed.: Springer Berlin Heidelberg, 2003.
[33] Data Management Association (DAMA) UK Working Group on "Data Quality Dimensions", "The six primary dimensions for data quality assessment - Defining data quality dimensions (Final Version)," ed. UK, 2013.
[34] R. Y. Wang, H. B. Kon, and S. E. Madnick, "Data quality requirements analysis and modeling," presented at the The Ninth International Conference on Data Engineering, Vienna, Austria, 1993.
[35] A. Levitin and T. Redman, "Quality dimensions of a conceptual view," *Information Processing & Management*, vol. 31, pp. 81-88, 1995/01/01 1995.
[36] R. Y. Wang, M. P. Reddy, and H. B. Kon, "Toward quality data: an attribute-based approach," *Decision Support Systems - Special issue on information technologies and systems*, vol. 13, pp. 349-372, March 1995.
[37] R. Y. Wang and S. E. Madnick, "A polygon model for heterogeneous database systems: the source tagging perspective," presented at the 16th International Conference on Very Large Databases, Brisbane, Australia, 1990.
[38] R. Y. Wang and S. E. Madnick, "A source tagging theory for heterogeneous database systems," presented at the International Conference on Information Systems, Copenhagen, Denmark, 1990.
[39] T. C. Redman, *Data quality: management and technology*: Bantam Books, Inc., 1992.
[40] D. P. Ballou and H. L. Pazer, "Cost/quality tradeoffs for control procedures in information systems," *Omega*, vol. 15, pp. 509-521, 1987/01/01 1987.
[41] J. Martin, *Design and strategy for distributed data processing*. Englewood Cliffs, London: Prentice-Hall, 1981.