# Chapter 89
# From Data Quality to Big Data Quality

**Carlo Batini**
*University of Milano-Bicocca, Italy*

**Anisa Rula**
*University of Milano-Bicocca, Italy*

**Monica Scannapieco**
*Italian National Institute of Statistics (Istat), Italy*

**Gianluigi Viscusi**
*École Polytechnique Fédérale de Lausanne, Switzerland*

## ABSTRACT

*This chapter investigates the evolution of data quality issues from traditional structured data managed in relational databases to Big Data. In particular, the paper examines the nature of the relationship between Data Quality and several research coordinates that are relevant in Big Data, such as the variety of data types, data sources and application domains, focusing on maps, semi-structured texts, linked open data, sensor & sensor networks and official statistics. Consequently a set of structural characteristics is identified and a systematization of the a posteriori correlation between them and quality dimensions is provided. Finally, Big Data quality issues are considered in a conceptual framework suitable to map the evolution of the quality paradigm according to three core coordinates that are significant in the context of the Big Data phenomenon: the data type considered, the source of data, and the application domain. Thus, the framework allows ascertaining the relevant changes in data quality emerging with the Big Data phenomenon, through an integrative and theoretical literature review.*

## INTRODUCTION

The area of Big Data (BD) is currently subject of intense investigation in academic literature, pushed by the growth of data made available in the Web and collected by fixed and mobile sensors. According to (Dumbill, 2013) "Big data is data that exceeds the processing capacity of conventional database systems.

DOI: 10.4018/978-1-4666-9840-6.ch089

The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it".

Another issue that in recent years raised the attention of scholars and practitioners is Data Quality (DQ), a multifaceted concept, to the definition of which different dimensions concur. Data quality has been investigated focusing especially on data as represented in the relational model, traditionally adopted in Data Base Management Systems (for an extensive survey of DQ in the relational model see Batini & Scannapieco, 2006), notwithstanding the growing relevance and concerns of non-standard data such as text, music, design information and pictures (Rose, 1991). More recently, a variety of data types rising from linguistic and visual information, used and diffused through social networks, enterprise and public sector information systems as well as the Web, resulted in a deep investigation on how data quality concepts can be extended to such vast set of data types, encompassing, e.g., semi-structured texts, maps, images, linked open data. Thus, the information growth consequent to the BD phenomenon has deeply impacted on the diversity of available types of data, the proliferation of sources of data, and the consequent great expansion of application domains.

Taking the above issues into account, in this paper we investigate how the multifaceted issues making up DQ have evolved from the traditional domain of databases to the domain of BD. The first coordinate we chose to analyze the evolution of the DQ concept are data types adopted in BD. In particular, we will analyze semi-structured texts, maps, and linked open data. Then, we will consider two other coordinates: (ii) the sources that originate BD, and (iii) application domains in which Big Data are used/investigated. As to sources, we will focus on sensors & sensor networks and as to application domains, we will focus on official statistics.

The chapter is organized as follows. First, we describe the methodology followed in the chapter, that adopts an integrative review perspective for a theoretical purpose. Then we present the conceptual framework for analyzing the evolution of the DQ issues from relational databases to the diverse data types, application domains and sources considered in the following. As for DQ issues, we consider dimensions classified in terms of dimensions clusters, adopting the clusters proposed in Batini, Palmonari, and Viscusi (2012). The three BD coordinates, namely data types, sources and application domains are analyzed in terms of their structural characteristics. Subsequently, the evolution paths dealt with in the paper are introduced. Every path considers the evolution of a dimensions cluster from the relational domain to the issues target of the BD coordinates above introduced (i.e., data types, sources and application domains), further showing how the evolution of a given dimension can be interpreted a posteriori according to the structural characteristics considered. A final general discussion on DQ dimension clusters and BD coordinates concludes the chapter.

## METHODOLOGY

The chapter adopts an integrative review perspective (Beyea & Nicoll, 2015; Torraco, 2005; Whittemore & Knafl, 2005), aiming to summarize what is actually known on DQ that can provide insights on how to face the challenges of BD quality. In particular, the focus is on the evolution of data quality dimensions. The need for this review is motivated by the emergent nature of BD quality, that is more than the sum of its parts, exemplified by the data types, data sources and application domains analyzed in the subsequent sections. Consequently, these parts make up the conceptual framework guiding the analysis of the evolution of quality in BD, together with the key constructs