Chapter XLII
Measuring Data Quality in Context

G. Shankaranarayanan
Boston University School of Management, USA

Adir Even
Ben Gurion University of the Negev, Israel

INTRODUCTION

Maintaining data at a high quality is critical to organizational success. Firms, aware of the consequences of poor data quality, have adopted methodologies and policies for measuring, monitoring, and improving it (Redman, 1996; Eckerson, 2002). Today's quality measurements are typically driven by physical characteristics of the data (e.g., item counts, time tags, or failure rates) and assume an objective quality standard, disregarding the context in which the data is used. The alternative is to derive quality metrics from data content and evaluate them within specific usage contexts. The former approach is termed as *structure-based (or structural)*, and the latter, *content-based* (Ballou and Pazer, 2003). In this chapter we propose a novel framework to assess data quality within specific usage contexts and link it to data utility (or utility of data) - a measure of the value contribution associated with data within specific usage contexts. Our utility-driven framework addresses the limitations of structural measurements and offers alternative measurements for evaluating completeness, validity, accuracy, and currency, as well as a single measure that aggregates these data quality dimensions.

BACKGROUND

Data quality is defined as fitness-for-use – the extent to which the data matches the data consumer's needs (Redman, 1996). However, in real-life settings, a single definition of the data quality may fail to support data management needs (Strong et al, 1997, Lee and Strong, 2003). Kulikowski (1971) suggests that data quality should be measured as a multi-dimensional vector that reflects different aspects of quality. Wang and Strong (1996) show that data customers perceive quality as having multiple dimensions such as accuracy, completeness, and currency. Quality, along each dimension, is often measured as a number between *0* (poor) and *1* (perfect). Pipino et al. (2002) identify three archetypes for quality metrics that adhere to this scale: (a) ratio between the actually obtained and the expected values, (b) min/max value among aggregations and (c) weighted average between multiple factors. Different measurement methods have been proposed along these archetypes (e.g., Redman, 1996; Pipino et al., 2002). Such measurements can be stored as quality metadata (Shankaranarayanan and Even, 2004), presented by software tools (Wang, 1998; Shankaranarayanan and Cai, 2006), tied to visual representations of data processes (Shankaranarayanan et al., 2003), and used for process optimization (Ballou et al., 1998).

Some quality dimensions (e.g., accuracy) are viewed as impartial (Wang and Strong, 1996) - i.e., the perception of quality along these dimensions is based on the data itself, regardless of usage. Others are viewed as contextual quality dimensions and perception of quality depends on the usage context (e.g., relevance). Pipino et al. (2002), however, argue that the same dimension can be measured impartially and/or contextually, depending on the purpose the measurement serves. As both impartial assessment and contextual assessment contribute to the overall perception of data quality, it is important to address both. We posit that within a usage context, the business value of data resources is reflected more by the data content and less by physical characteristics. Hence, we suggest that content-based measurement of quality is more appropriate for contextual assessment. We use utility functions (Ahituv, 1980) to link impartial information characteristics (here, data contents and presence of defects) onto tangible values within specific usages. Utility mapping has been used to examine tradeoffs between quality dimensions and optimize their configuration (Ballou et al., 1998; Ballou and Pazer, 1995, 2003).

The quality measurements proposed here are based on the traditional data hierarchy (adapted from Redman, 1996). The foundation of this hierarchy (figure 1) is the data item. The data item is defined as a triplet <*a,e,v*> of a data value *'v'* selected from the value domain attached to attribute *'a'* of entity *'e'* that represents a physical or logical real-world object. The data record is a collection of data items that represent the attributes of an entity instance. A dataset is a collection of records that belong to the same entity class (e.g., a subset of records in a table), and a database is a collection of datasets with meaningful interrelationships. Organizations typically have a collection of databases. Certain measurements evaluated here are defined at different hierarchical levels. The annotation used to differentiate the same measurement between levels is described in the glossary at the end of the chapter.

The framework examines the tabular dataset, assuming *N* identically structured records (rows, indexed by *[n]*), and *M* attributes per record (columns, indexed by *[m]*). Data contents, the actual attribute values of record *[n]*, are denoted $f^{E}_{n,1}$ through $f^{E}_{n,M}$. Each attribute has a valid set of values (e.g., integer, real, alphanumeric, or a finite set) defined by its value domain. We next develop the concept of utility specifically for datasets. We then use it to develop content-based and contextual data quality measurements at different data-hierarchy levels.

UTILITY OF DATA

The utility of a data resource is a non negative measurement of its value contribution. In commercial