Chapter XLI
Data Quality Assessment

Juliusz L. Kulikowski
Institute of Biocybernetics and Biomedical Engineering PAS, Warsaw, Poland

INTRODUCTION

For many years the fact that for a high information processing systems' effectiveness high quality of data is not less important than high systems' technological performance was not widely understood and accepted. The way to understanding the complexity of data quality notion was also long, as it will be shown below. However, a progress in modern information processing systems development is not possible without improvement of data quality assessment and control methods. Data quality is closely connected both with *data form* and *value of information* carried by the data. High-quality data can be understood as data having an appropriate form and containing valuable information. Therefore, at least two aspects of data are reflected in this notion: 1st - technical facility of data processing, and 2nd - usefulness of information supplied by the data in education, science, decision making, etc.

BACKGROUND

In the early years of information theory development a difference between the *quantity* and the *value* of information was noticed; however, originally little attention was paid to the information value problem. R. Hartley interpreting information value as its psychological aspect stated that it is desirable to eliminate any additional psychological factors and to establish an information measure based on purely physical terms only (Klir 2006, pp. 27-29). C.E. Shannon and W. Weaver created a mathematical communication theory based on statistical concepts, fully neglecting the information value aspects (Klir ,2006, p.68). In most of later works concerning information theory backgrounds attention was focused on extension of the *uncertainty* concept rather than on this of information value. Nevertheless, L. Brillouin tried to establish a relationship between the quantity and the value of information stating that for an

information user the relative information value is smaller than or equal to the absolute information, i.e. to its quantity (Brillouin,1956, Chapt. 20.6). M.M. Bongard (Bongard, 1960) and A.A. Kharkevitsch (Kharkevitsch, 1960) have proposed to combine the information value concept with the one of a statistical decision risk. This concept has also been developed by R.L. Stratonovitsch (Stratonovitsch, 1975, Chapts. 9, 10). This approach leads to an economic point of view on information value as profits earned due to information using (Beynon-Davies, 1998, Chapt. 34.5). Such approach to information value assessment is limited to the cases in which economic profits can be quantitatively evaluated. In physical and technical measurements data accuracy (described by a mean-square error or by a confidence interval length) is used as the main data quality descriptor. In medical diagnosis data actuality, relevance and credibility as well as their influence on diagnostic sensitivity and specificity play relatively higher role than data accuracy (Wulff, 1981). This indicates that, in general, no universal set of data quality descriptors exists; they rather should be chosen according to the application area specificity. In the last years data quality became one of the main problems posed by the world wide web (WWW) development (Baeza-Yates & B. Ribeiro-Neto, 1999, Chapt. 13.2). The focus in the domain of finding information in the WWW increasingly shifts from merely locating relevant information to differentiating high-quality from low-quality information (Oberweis & Perc, 2000, pp. 14-15). In the recommendations for databases of the Committee for Data in Science and Technology (CODATA) several different quality types of data are distinguished: 1st primary (rough) data whose quality is subjected to individually or locally accepted rules or constraints, 2nd qualified data, broadly accessible and satisfying national or international (ISO) standards in the given application domain, 3rd recommended data – the highest quality broadly accessible data (like physical fundamental constants) that have passed a set of special data quality tests. In the last decades several technological tools for formal data incorrectness detection and rectifying have been proposed (Shankaranarayan & Ziad & Wang, 2003). In some countries the interests of information users are legally protected from distribution of certain types of incredible or misguided data. On the other hand, a governmental intervention into the activity of open-access databases is also limited by international legal acts protecting human rights to free distribution of information.

BASIC PROBLEMS OF DATA QUALITY ASSESSMENT

The idea that information value can be better characterized by a multi-component vector than by a scalar value arose in late 60-ths of the 20th century. In such case the following problems arise: 1st, what information (or – data, as its particular form) aspects should be taken into account for its value characterization, 2nd, what does it mean that a multi-aspect information value is "better" or "higher" than another one, 3rd, how simple data multi-aspect values can be extended on higher-level data structures. J.L. Kulikowski (Kulikowski, 1971) proposed to describe information value by an element of semi-ordered linear vector space (Kantorovitsch space). For a multi-aspect data quality evaluation data validity (actuality), relevance, credibility, accuracy, and operability as quality factors were originally proposed. In the ensuing years the list of proposed data quality factors by other authors has been extended up to almost two hundreds (Wang & Strong, 1996; Shanks & Darke, 1998; Pipino & Lee & al., 2002). Between arbitrarily chosen data quality factors hidden functional dependence may exist, as illustrated in Fig. 1.

Fig. 1a illustrates a hypothetical dependence between data validity v_{vd} and data operability v_{op}: improving data operability is a time-consuming operation reducing data validity. Fig 1b shows that