Food Chemistry 193 (2016) 166–172

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

A scope classification of data quality requirements for food composition data

Karl Presser [*], Hans Hinterberger, David Weber, Moira Norrie

Department of Computer Science, ETH Zurich, Zurich, Switzerland

ARTICLE INFO

Article history:
Received 10 February 2014
Received in revised form 17 December 2014
Accepted 15 March 2015
Available online 6 May 2015

Keywords:
Data quality
Classification of requirements
Food composition data

ABSTRACT

Data quality is an important issue when managing food composition data since the usage of the data can have a significant influence on policy making and further research. Although several frameworks for data quality have been proposed, general tools and measures are still lacking. As a first step in this direction, we investigated data quality requirements for an information system to manage food composition data, called FoodCASE. The objective of our investigation was to find out if different requirements have different impacts on the intrinsic data quality that must be regarded during data quality assessment and how these impacts can be described. We refer to the resulting classification with its categories as the scope classification of data quality requirements. As proof of feasibility, the scope classification has been implemented in the FoodCASE system.

1. Introduction

The European Food Information Resource (EuroFIR) project was a Network of Excellence (NoE) funded through the EU FP6 (2005–2010, 513944). One of its main objectives was to develop, for the first time in Europe, a single online platform with up-to-date food composition data across Europe (EuroFIR. EuroFIR history., 2013). In cooperation with the Swiss food compilers and EuroFIR, we have developed a food composition data management system called FoodCASE (Food Composition And System Environment). The implementation of FoodCASE was started in 2007 when we first implemented a database according to the EuroFIR proposals (Becker, Unwin, Ireland, & Møller, 2007; Becker et al., 2008) and the COST Action 99 project proposal, which together form the basis for the CEN standard on food composition data (Becker, 2010; CEN, 2012). Although software systems for the management of food composition data already existed in Switzerland, it was decided that the new FoodCASE system should be developed to follow the standards defined by EuroFIR as well as being flexible enough to support our research on data quality.

The long-term goal of our investigations on data quality is to develop a framework to support the implementation of information systems such as FoodCASE. The framework should provide easy and visual access to data quality information and metrics together with indicators of actions to be taken to address specific problems such as data that is incomplete, outdated or below a quality threshold.

As a first step in this direction, we carried out a detailed study of the data quality requirements for food composition data in FoodCASE. The objective of the study was to investigate whether all data quality requirements have the same impact on the intrinsic data quality or if there are differences which should be taken into account during data quality assessment. In addition, we wanted to find out how these differences in impact can be described and how the quality assessment is affected.

We start in Section 2 with a review of existing data quality frameworks. Section 3 then describes the three-phase study in corresponding subsections. Section 3.1 describes the different approaches that we had to take to collect all data quality requirements and the resulting requirement set. The criteria to distinguish the categories and the resulting classification are presented in Section 3.2. The implementation of our framework in FoodCASE is described in Section 3.3. Concluding remarks are given in Section 4.

2. Background

Comprehensive reviews of data quality frameworks have been done by Batini and Scannapieco (2006) and Eppler (2006). The latter identified 20 frameworks where most have a specific domain focus and only a few are general.

The most often cited general data quality framework is the one of Wang and Strong (1996). They interviewed data consumers

[*] Corresponding author at: ETH Zurich, Department of Computer Science, CNB E 102.2, Universitätsstr. 6, 8092 Zurich, Switzerland.
E-mail address: karl.presser@inf.ethz.ch (K. Presser).

http://dx.doi.org/10.1016/j.foodchem.2015.03.084
0308-8146/© 2015 Elsevier Ltd. All rights reserved.

about data quality and generated 179 dimensions. They condensed and summarised these dimensions to produce a final set of 16 dimensions and four categories that are presented in Fig. 1. For example, under the category intrinsic data quality, they grouped the data quality dimensions believability, accuracy, objectivity and reputation.

The term dimension comes from mathematics where a space can consist of *n* dimensions. This concept was mapped into the data quality framework and the notion of data quality having multiple dimensions evolved. A data quality dimension is, therefore, a specific data quality descriptor or property.

Redman (1996) grouped data quality dimensions into three sets: Those relating to the model or view, those relating to data values, and those relating to the representation of records. Redman defined the view as the "part of the real world" to be captured in the data. The first category contained 15 dimensions (relevance, obtainability, clarity of definition, comprehensiveness, essentialness, attribute granularity, domain precision, naturalness, occurrence identifiability, homogeneity, minimum redundancy, semantic consistency, structural consistency, robustness and flexibility). The second category contained four dimensions (accuracy, completeness, currency and value consistency), while the third category contained eight dimensions (appropriateness, interpretability, portability, format precision, format flexibility, ability to represent null values, efficient usage of recording media and representation consistency).

Both of the frameworks described above have a separate category for intrinsic data quality, where the inherent quality of data is regarded, and other aspects, such as the purpose of data usage or presentation of data, are excluded. In our investigations, we focused on this category since it deals with the quality of the data managed by the system rather than what might be considered as contextual properties of how data is represented and used.

What is necessary is all the additional information about a food composition value required to determine its quality. For instance, it is not enough to simply have a value of 5 mg vitamin C for the food item apple since it is also necessary to have information about how the sampling and analytical measurements were done in order to estimate the data quality.

The use of metadata to evaluate data quality has been proposed by other researchers, for example (Mihaila, Raschid, & Vidal, 2000) and (Rothenberg, 1996). Specifically, Rothenberg argued that information producers should perform verification, validation and certification of their data, and then provide data quality metadata along with the datasets. Also, Naumann, Leser, and Freytag (1999) presented a mediation framework for the querying of data in molecular biology where data are selected from different data sources based on data quality information stored as metadata. The authors based their approach on the data quality framework of Wang and Strong (1996), defining scores for each of the data quality dimensions and normalising them to build a weighted sum.

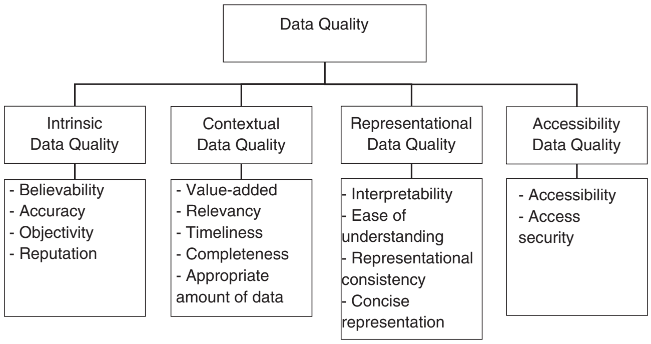

Fig. 1. A data quality framework with 15 dimensions identified by Wand and Strong in 1996.

Most of the frameworks that have been proposed are of a conceptual nature. Thus, while the frameworks proposed for example by Wang and Strong (1996) and Redman (1996) categorise dimensions in slightly different ways, they are similar in their approach in that their definitions of the dimensions tend to be descriptive and subjective (Naumann & Rolker, 2000), often using adjectives for which the semantics are overlapping or fuzzy.

Research on tools and metrics to support data quality management in practice tends to be limited to specific dimensions or domains. For example, specific metrics have been proposed for the data quality dimensions of timeliness (Ballou, Wang, Pazer, & Tayi, 1998; Hinrichs, 2002; Klier, 2007) and accuracy (Hinrichs, 2002; Klier, 2007). Pipino, Yang, and Wang (2002), on the other hand, presented three general ways of deriving measures of data quality based on simple ratio, the minimum or maximum operation and the weighted average. The simple measure is the ratio of current to total outcomes. For instance, if a column of a table should contain at least one occurrence of all 50 states, but only contains 43 states, the population is incomplete and a ratio of 43/50 generated. Minimum or maximum can be used to aggregate multiple data quality dimensions by simply selecting the minimum or maximum value, respectively, from the normalised data quality values of the individual data quality dimensions. Alternatively, one could use a weighted average so that certain dimensions are given more importance than others in determining data quality.

A challenge for the measurement of data quality dimensions is the high abstraction level of the dimensions. A dimension normally consists of several data quality requirements that are more concrete and demand a specific constraint be satisfied. Examples for data quality requirements are the uniqueness of food names or that component values should be floating point numbers.

In the area of food composition data, there are proposals with concrete metrics for data quality requirements that can be used to determine overall data quality. Holden, Bhagwat, and Patterson (2002) generalised and expanded their existing data quality evaluation system in the U.S. to be valid for all nutrients. The evaluation system consists of five categories: sampling plan, number of samples, sample handling, analytical method and analytical quality control. One modification was the extension of the rating scale for every category from 0–3 to 0–20 to have a more continuous scale. The sum of the five categories determines the so-called quality index of a nutrient value. Quality index is a data quality rating value that indicates the reliability of a food composition value. In a second step, nutrient data from several acceptable sources are aggregated to give an overall estimate of the nutrient content of that food. In this step, the quality indexes are also aggregated into the so-called confidence code.

To explain the confidence code, consider again a database that has three vitamin C values for an apple from three different sources. For the three values, their quality index would be calculated as described above. To have one representational vitamin C value that can be published, the three values are aggregated. For the aggregated vitamin C value, a confidence code indicates how reliable the value is and, therefore, is similar to the quality index. The scale of the confidence code is simplified and consists only of the letters A, B, C and D, where A is the highest rating.

Specific rules are applied in the aggregation of the confidence code. For example, if individual sources have not received high ratings because the samples were only regional, the confidence code might be higher than the simple sum because the regions were not intersecting and hence a bigger area of the country was covered. In addition, the sum of the confidence codes is fitted to a range, the maximum of which is 100, to avoid aggregation of