AN ASSESSMENT IMPACT CLASSIFICATION OF DATA QUALITY REQUIREMENTS IN FOOD COMPOSITION DATABASE SYSTEMS

Karl Presser [1], *David Weber* [1], *Moira Norrie* [1]

[1] Department of Computer Science, ETH Zurich, Zurich, Switzerland, presser@inf.ethz.ch, weber@inf.eth.ch, norrie@inf.ethz.ch

Abstract – There exist established standards and quality controls for laboratory analysis on national, European and international levels. But there are only a few data quality recommendations for the electronic storage and exchange of analytical data. We therefore have collected explicit and implicit data quality requirements for food composition data and classified them according to their impact on implementation and quality assessment within a food composition database system.

Keywords: data quality requirements, data quality classification, data quality assessment, impact for implementation of database systems

1. INTRODUCTION

Data quality is an important aspect for laboratories so that an ISO standard ISO/IEC 17025 exists for which laboratories can be accredited in order to prove technical competence. There also exist European reference laboratories that coordinate a network of national reference laboratories to achieve high-quality results by provision of reference methods and materials, proficiency testing and training for laboratory staff. These data needs to be transferred from laboratories into specific databases such as a food composition database or a Total Diet Study (TDS) dataset. In the case of national food composition databases, values come mostly from different sources and data quality needs to be checked. A well-known issue was the high iron content of spinach, which was documented to be 10 times higher than measured. In 2012, we found in the Swiss food composition database that the sum of all nutrients per 100g edible portion summed up to around 118g.

Our goal was therefore to investigate a data quality framework in order to maintain and improve data quality in an electronic food composition database. The first step of this framework was to collect data quality requirements for food composition data and to analyse how they influence the implementation of a data quality assessment. A data quality requirement is a request to data to satisfy a criterion and can be as simple as 'food name must be provided' or more complex such as 'an adequate method should be used for analysis'.

2. BACKGROUND

There exist three different research areas that contribute to data quality of food composition data. These areas are data management, food composition itself and database systems.

In the area of data management, there exists the pioneer work from Wang and Strong [1]. They collected 179 data quality requirements from users and generated so-called dimensions which summarise the requirements into logical groups. The result is called a data quality framework. They came up with 15 dimensions which were divided into four groups; see Fig. 1.

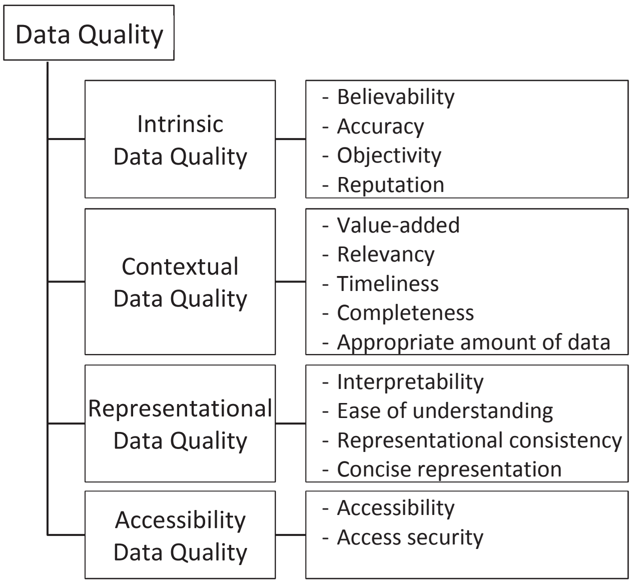

Fig. 1. A conceptual data quality framework with 15 dimensions identified by Wang and Strong in 1996.

Redman in [2] categorises data quality dimensions in three sets: those relating to the model or view, those relating to data values and those relating to the representation of records. Redman defines a view as 'part of the real world' to be captured in the data. The first category contains 15 dimensions, the second category contains four dimensions and the last contains eight dimensions. A comprehensive review with a focus on data quality dimension frameworks was done by Eppler in [3], where he identified 20 frameworks and investigated seven from different application contexts in more detail. His research showed that most of the frameworks are built for a certain domain and only a few general valid frameworks exist. The issue with these frameworks is that they are defined on an abstract level and do not give an idea of how data quality assessment must be considered. They do not consider concrete data quality requirements which are the criteria by which to assess data quality.

Other publications propose how these dimensions can be assessed. The proposals range from very simple ones as in [4], where a simple ratio, a min/max operation and a weighted average can be used to more tailored ones that include weightings for dimensions based on surveys and using a scorecard index to compare the data quality level of a data record to others [5]. Metadata is additional information about an analytical result such as a sampling plan or analytical methods used. A claim of some authors [6][7] is the use of metadata to help to assess data quality. Rothenberg argued that information producers should perform verification, validation and certification of their data and that they should provide data quality metadata along with the datasets [7]. Naumann in [8] found a methodology to select data from different data sources according to their quality that is based on the metadata claim. Naumann's work presented a framework where data quality is involved in the query processing in a multi-database environment. This means that a mediator system, which is queried for molecular biology information and further queries other database systems to collect data, performs first a quality-driven source selection and then returns only most qualitative results to the user. Naumann took the data quality framework of Wang and Strong [1] and defined different, simplified scores for the data quality dimensions.

These proposals are based on the former data quality frameworks and are therefore also lacking in how concrete data quality requirements must be considered in a quality assessment.

In the area of food composition, the COST Action 99 project and the EuroFIR project established a harmonised data format for food composition data and defined guidelines for the process of food compilation [9][10][11]. Also in the EuroFIR project, a set of questions was proposed to determine a so-called quality index for values from literature [12] [13]. The set contains 34 questions grouped in seven categories. In contrast to quality dimensions, these definitions go into detail and cover a part of what needs to be considered when implementing a database system.

In the area of database management, there exists a work that classifies data quality requirements into hard, soft and indicator constraints [14]. Hard constraints are requirements that define data to be deficient, while soft constraints and indicators are saying that data quality is decreased but data is not deficient. The indicator even expresses that a decrease in data quality is assumed but not evidence-based. A typical example is that if a value is more than 10 years old, it is assumed that quality is decreased because of its age, but it is not proven. This so-called scope classification is another part of comprehensive data quality classification framework.

What is missing is a data quality framework that serves as an overview map for requirements and contains all the existing parts from literature. With the help of the framework, it should be possible to quickly identify where and how a requirement should be implemented and how it should be integrated into the quality assessment.

3. APPROACH

Data quality requirements from different sources were extracted to get a comprehensive set of requirements for food composition data.

The EuroFIR technical annex [11] contains attributes and data type definitions for food composition entities such as food, component or value. A data type is a classification of data such as integer, floating point number or Boolean. Each attribute and type definition is a data quality requirement that should be implemented in a database system. The 34 questions of the EuroFIR quality index [12] can mostly be answered with 'yes', 'no' or 'not applicable' and should be answered for every nutrient value. These are also data quality requirements which are expressed in question form. During the implementation of FoodCASE, which is an information system to manage food composition data, we deducted some requirements that are not explicitly stated in any document and we also obtained various requirements in communications with food scientists.

The final list was then verified by food scientists, and we investigated their impacts for the implementation of a database system as well as their impacts on the assessment of data quality.

4. RESULTS

In total, 451 data quality requirements were identified for the entities food, value, component, sample, method and reference. The EuroFIR technical annex contains 295 requirements [11], while 122 additional data quality requirements were deducted or obtained by communications with experts.

The analysis of the requirements showed that it is not possible to have one classification that covers all impact characteristics. Instead, we found six different classifications that target specific issues. In consequence, each requirement must be classified in all six classifications and belongs to a category in each classification. The six classifications that we identified to have an impact on implementation and assessment are: Scope, type, data dependency, user dependency, time dependency and answering.

4.1. Scope Classification

The scope classification was already described in [14]. In particular, hard constraints have the most common impact on software in that deficient data is mostly not allowed to be stored. In [14], there is also a proposal regarding the way soft constraints and indicators can be handled in software.

4.2. Type Classification

A data quality requirement can be of type atomic, composite or quality key. An atomic requirement is directly applied on attributes of an entity or on the entity itself. For instance, an atomic requirement is that 'a food name must have at least two letters' or 'a food item must have at least four macro nutrient values'.

In contrast, there are requirements that summarise other requirements, meaning that they are not applied on data directly but on other requirements. A composite requirement could be, for instance, a requirement that postulates that the two atomic requirements mentioned previously are satisfied. Such a composite requirement is used to calculate a summarised data quality assessment value; see Fig. 2. Additionally, a weighting factor is used to reflect that different requirements have different degrees of importance.

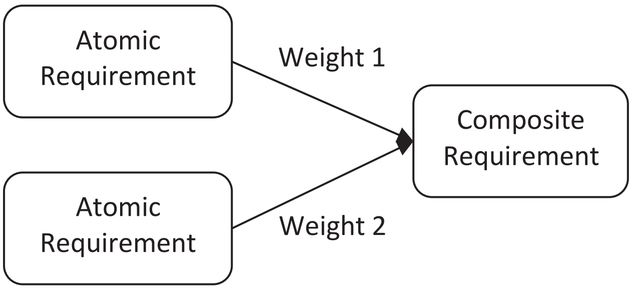

Fig. 2. Weights of contributing requirements to composite or quality index requirements vary for different users.

A quality key requirement is a special composite requirement. It is the top requirement that normally includes all other requirements. It is normally used to calculate some sort of summary or aggregate data quality assessment value that is disseminated together with the data. The assessment can be calculated on each data record or over all data records. In an assessment approach, the types also define the order of the assessment. First, all atomic requirements must be assessed before the composite requirements can be evaluated and all composite requirements must be evaluated before the quality key.

The EuroFIR data quality index forms a good example for the type classification. The 34 questions, which are atomic data quality requirements, are summarised in seven categories that are of type composite. The seven categories are also summarised to a single quality index in the range between 7 and 35, which forms the quality key.

4.3. Data Dependency Classification

The classification data dependency has two categories: data general and data individual.

The data individual category represents requirements that can only be applied to certain data records. As an example, consider the requirement 'is the fatty acid conversion factor provided'. This requirement is only applicable for a nutrient value of total fat or total fatty acid because one can be calculated using the other and the conversion factor. But it is not applicable for carbohydrate or alcohol. Consequently, the assessment approach has to take into account that some data records of the same entity can have a different number of requirements. This impact on the data quality assessment directly affects the implementation of a database system. Normally a single user input mask is implemented and re-used to enter data records of the same entity, such as a nutrient value. This user input mask needs additional functionality to recognise what data is entered to add or remove certain requirement checks.

In contrast, requirements from the data general classification can be applied to all data records of the same entity. For example, the requirement that a food name must be unique can be applied to all foods.

4.4. User Dependency Classification

In an importance rating study for data quality dimensions [15], we showed that different user groups rate the importance of some requirements significantly different. We even showed that the importance rating was different within the same user group. This finding was interesting, as the food compiler user group was expected to rate requirements intention-independent as they collect food composition data independently of later use. The user dependency classification is directly derived from these findings. The classification user dependency contains the categories user general and user individual requirements.

For user individual requirements, the community of users does not find a consensus and hence some of the users support the existence of a certain requirement while others deny the existence. But even if there is an agreement on the existence of a data quality requirement, there can be a disagreement on the importance of the requirement. If users agree on the existence and importance of a data quality requirement, it belongs to the user general category; otherwise, it belongs to the user individual category.

The user dependent requirements have a direct impact on assessment and implementation in that they need to be added or removed based on the current user. On the other hand, the differences in importance rating influence the assessment for requirements of type composite and quality key. Depending on the importance, the weighting factors of composite or quality keys are different as shown in Fig. 2.

4.5. Time Dependency Classification

The classification time dependency has the two categories during input and after input. The category during input contains data quality requirements that can be validated, while a user is entering data while the after input category contains requirements that first can be checked after some data is entered. The first category has the advantage that a system can prevent deficient data from being entered. The category to which a requirement belongs can depend on the application design. Consider the requirement that ‘the sum of all nutrient values per 100 g must be close to 100 g’. If an application provides a user input mask where all nutrient values must be entered before the save operation can be performed, the requirement can be checked at input time. As the amount of nutrient values per food can be bigger than 100, the user should be able to store at any time. In this case the requirement belongs to the category after input because the requirements can first be checked after the last nutrient value was entered.

The impact for the implementation of a database system is that an approach is needed where users can inform the system that data entry has finished and that time-dependent requirements can be validated. A challenging situation is when a requirement is a hard constraint and belongs to the after input category. In such a case the implemented data quality framework must notify users to solve the situation.

4.6. Answering Classification

In the classification answering there are the two categories automatic answer and user answer. In the automatic answer category are requirements for which the answer as to whether a requirement is satisfied can be automatically evaluated by the system. In contrast, the user answer category contains requirements for which the answer cannot be automatically generated but needs user input. For instance, a data quality requirement that an ‘English name must have at least three characters’ can be evaluated automatically. In contrast, a requirement such as ‘is it clear what part of an animal was taken for sampling’ is hardly determinable by the system and needs user input.

The impact on implementation and quality assessment is obviously that the system needs input from users before being able to evaluate data quality. An additional issue is that a user answer is an opinion and can differ from one user to another.

5. IMPACT ON ASSESSMENT SCALE

We have seen that all six classifications have an impact on the data quality assessment in a food composition database system. Additionally, four of the classifications must be considered when determining an assessment scale: scope, data dependency, user dependency and time dependency. Fig. 3 is used to describe the ranges of an assessment scale.

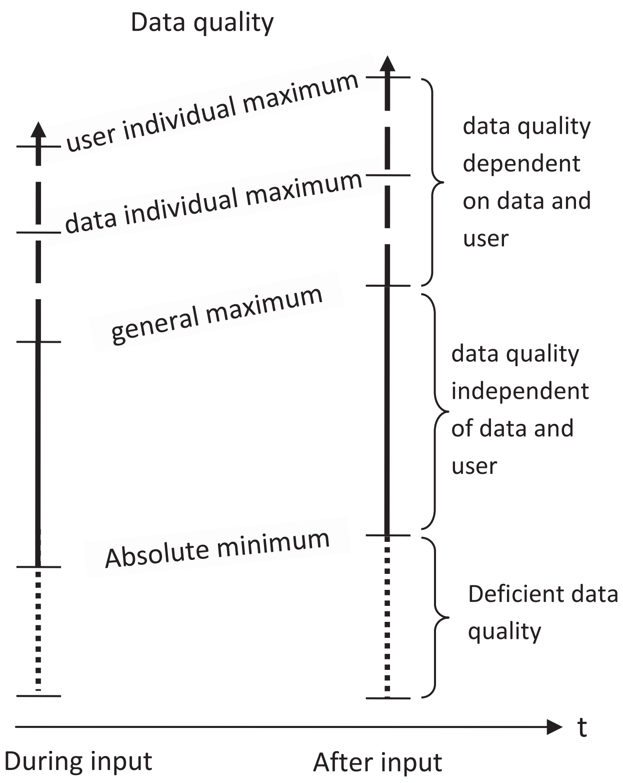

Fig. 3. Requirement classifications and their categories' impact on the data quality assessment scale.

The two data quality value axes, bold arrows in Fig. 3, go from bottom to top and represent the data quality rating of an entity such as food or nutrient value. The higher a data quality value is, the more requirements are satisfied with respect to their importance. To explain why there are two data quality axes having different sizes, we have to consider each of the four classifications. Each data quality requirement belongs to one category of each classification. The scope classification has the categories hard constraint and soft constraint/indicator. We have seen in the result section that hard constraints invalidate data if they are not satisfied. In Fig. 3, this fact is represented in that a scale has an absolute minimum, under which the quality of data is deficient. It can be argued that the area under the absolute minimum is not a range but a single point for deficient data. But as there is mostly more than one hard constraint, we see it as a small range, on which the amount of unsatisfied hard constraints influences how far the quality of a data record is from the absolute minimum. It is only a small range because the main message is that data is deficient and therefore the range under the absolute minimum does not need to be distinctive. If a data record satisfied all hard constraints, it has the absolute minimum data quality.

As seen in the description of the scope classification, soft constraints and indicators do not make data quality deficient. Their impact on the assessment scale is the same and therefore these two scopes are summarised. The three categories together determine the section on the data quality scale from the absolute minimum to the general maximum. General means that these requirements are applied against all data records and are not dependent on data or users.

In contrast, data-specific requirements can only be applied to certain data. Hence for individual entities, it is possible to reach a higher data quality than for others. This behaviour is represented as a dashed line in Fig. 3 from the general maximum to the data individual maximum. There is a similar situation with user individual requirements. Certain users or user groups can have more data quality requirements so that data records can reach higher data quality values than they can for other users or user groups. This behaviour is also represented as a dashed line from the data individual maximum to the user individual maximum.

The last categories that influence the data quality scale are the categories of the time dependency classification. The during input category determines the assessment range in which a data record can be when data is entered into the system. In Fig. 3 the arrow on the left side represents the whole during input scale. The requirements of the after input category can first be satisfied after certain data is entered. In Fig. 3 this behaviour is indicated by the second data quality scale on the right side. It can be seen that these requirements can have an impact on all classification sectors on the data quality scale.

Other reasons why a data quality scale can increase over time are either that new requirements are taken into account or the rating scales of requirements increase. As an example, consider the assessment of the analytical method used to determine a nutrient value. New analytical methods are invented over time that are more accurate and will result in higher quality scores for a nutrient value. So the quality score will increase with every new and more accurate method.

6. CONCLUSION

We showed that there are six classifications (scope, type, data dependency, user dependency, time dependency and answering) in which data quality requirements can be classified. These six classifications are targeted on the implementation of data quality assessment in a food composition database system. We showed how four of the six classifications influence the data quality scale so that every data quality rating must be regarded in the context of the scale. We also showed that certain classification combinations can increase the challenge to handle requirements correctly. A database system needs an advanced data quality framework that can manage all these issues.

With the implementation of a data quality framework in FoodCASE, we showed that it is possible to consider all these requirement classifications. The framework needed two parts: input validation and a quality analysis toolkit which represents the two categories of the time dependency classification. The input validation is validating all requirements that can be checked during data entry regarding scope, data dependency and answering requirements. The quality analysis toolkit allows users to define ratings of each data quality requirement as well as the weights for composite and quality keys. It also allows the definition of requirements from the scope, type, data dependency, user dependency and time dependency classification.

ACKNOWLEDGEMENTS

The authors would like to thank all EuroFIR project colleagues who have contributed to this work.

This work was partially supported by the EuroFIR Nexus project (GA-265967) funded under the EU 7th Framework Programme.

REFERENCES

[1] R.Y. Wang and D.M. Strong, Beyond Accuracy: “What Data Quality Means to Data Consumers”, *Journal of Management Information Systems*, vol. 12, nº. 4, pp. 5-33, 1996.

[2] T.C. Redman, *Data Quality for the Information Age*, ARTECH HOUSE, 1996.

[3] J.M. Eppler, *Managing Information Quality*, Springer Verlag, 2006.

[4] L. Pipino, L. Yang and R.Y. Wang, Data Quality Assessment, *Communications of the ACM*, vol. 45, pp. 211-218, 2002.

[5] T.-H. Moges, K. Dejaeger, W. Lemahieu and B. Baesens, “A multidimensional analysis of data quality for credit risk management: New insights and challenges”, *Information & Management*, vol. 50, nº. 1, pp. 43-58, 2013.

[6] G. A. Mihaila, L. Raschid, and M.-E. Vidal, “Using Quality of Data Metadata for Source Selection and Ranking”, *3rd international WebDB Workshop*, Dallas, Texas, USA, 2000.

[7] J. Rothenberg, “Metadata to Support Data Quality and Longevity”, *First IEEE metadata conference*, Silver Spring, Maryland, USA 1996.

[8] F. Naumann, U. Leser and J. C. Freytag, “Quality-driven Integration of Heterogeneous Information Systems”, *25th International Conference on Very Large Data Bases*, pp. 447-458, Edinburgh, Scotland, 1999.

[9] F. Schlotke, W. Becker, J. Ireland, A. Møller, M.-L. Ovaskainen, J. Monspart, et al., “Eurofoods Recommendations for Food Composition Database Management and Data Interchange”, *Journal of Food Composition and Analysis*, vol. 13, nº. 4, pp. 709–744, 2000.

[10] W. Becker, “Towards a CEN Standard on Food Data”, *European Journal of Clinical Nutrition*, vol. 64, pp. 49–52, 2010.

[11] W. Becker, A. Møller, J. Ireland, M. Roe, I. D. Unwin and H. Pakkala, “Proposal for Structure and Detail of a EuroFIR Standard on Food Composition Data”, EuroFIR Technical Report, 2008.

[12] S. Salvini, M. Oseredczuk, M. Roe, and A. Møller, “Guidelines for Quality Index Attribution to Original Data from Scientific Literature or Reports for EuroFIR Data Interchange”, EuroFIR Report, 2009.

[13] S. Westenbrink, M. Oseredczuk, I. Castanheira, and M. Roe, “Food Composition databases: The EuroFIR Approach to develop Tools to Assure the Quality of the Data Compilation Process”, *Food Chemistry*, vol. 113, nº. 3, pp. 759–767, 2009.

[14] K. Presser, H. Hinterberger, D. Weber and M. Norrie, “A scope classification of data quality requirements for food composition data”, *Food Chemistry*, vol. 193 (2016), pp. 166-172, 2015.

[15] K. Presser, D. Weber and M. Norrie, “A Study of Data Quality Requirements for Empirical Data in the Food Sciences”, *Proceedings of the European Conference on Information Systems (ECIS)*, Tel Aviv, Israel, 2014.