The use of ETL and data profiling to integrate data and improve quality in food databases

Alexander Münzberg
Department of Computer Science and Microsystems Technology
University of Applied Sciences Kaiserslautern
Zweibrücken, Germany
alexander.muenzberg@hs-kl.de

Janina Sauer
Department of Computer Science and Microsystems Technology
University of Applied Sciences Kaiserslautern
Zweibrücken, Germany
janina.sauer@hs-kl.de

Andreas Hein
Department of Assistance Systems and Medical Technology
Carl von Ossietzky University Oldenburg, Germany
Andreas.Hein@uni-oldenburg.de

Norbert Rösch
Department of Computer Science and Microsystems Technology
University of Applied Sciences Kaiserslautern
Zweibrücken, Germany
norbert.roesch@hs-kl.de

Abstract— This paper focuses on integrating food data sources into a central database using extract, transform and load processing and the subsequent data quality enhancement. The obtained data will be transmitted by a food data web service to certain health apps for further use. Furthermore, it is planned to identify inconsistent, incorrect, duplicate and incomplete data using methods of data profiling so that they can be corrected. In order to quantify the data quality purposefully and appropriately, certain quality metrics were used. These metrics were calculated and evaluated using random test data selected from the food data.

Keywords—food data, food information service, ETL process, data warehousing, data profiling, data quality improvement, data quality metrics

I. INTRODUCTION

At the present time, comprehensive and standardized data sets on ingredients used and nutritional information of available food products are still missing. Although ingredient information needs to be printed on packaging of food products within European member states, there is no legal obligation to deliver digital food information such as information on ingredients, nutritional values and allergens to consumers in a standardized manner. For example, in the ingredient lists, different names may refer to the same ingredient. Updating food data whose information changes over time, such as the composition of ingredients, lead to further problems for the operators of food databases [1]. Due the missing of standardized data sources, providers of mobile health apps often have to revert to dubious external data sources or maintain their own databases, to fill these with information from product manufacturers or third parties. In particular, data sources compiled by Internet communities contain many inconsistencies and are often not adequately controlled and reviewed. The use of such data for an analysis in health applications, is therefore not recommended without extensive examination and data quality improvement.

Within the project named "Digitized Services in Dietry Counselling for People with Increased Health Risk Related to Malnutrition and Food Allergies" (DiDiER), funded by the Ministry of Education and Research (grant 02K14A150), high-quality food data are needed, which are used to evaluate digital nutrition diaries [2]. In the field of food intolerance and allergy, accurate ingredient lists and precise nutrition facts are required for the detection of elicitors. Within the project, a food data web service named "Food Information Service" (FIS) was developed. The attributes stored there, provide product related information, generic food information, European Article Numbers (EAN) [3], categories, brands, origins, contents, and information on ingredient lists, nutrition facts and allergens. The data comes from data sources of various external providers and are integrated during the data warehousing process "Extract, Transform and Load" (ETL) [4]. In order to increase the quality of the data, specific quality rules and ontologies were used and relations between the data elements of the different data sources were formed. The Data profiling process detect inconsistent, incorrect, missing and duplicate data in the data sets which must be corrected. Special metrics were used to quantify the data quality and evaluate its improvement. These metrics were computed with test data randomly selected from the food data and the results of these calculation were finally evaluated.

II. STATE OF THE ART

A. Data sources

Within the DiDiER project, partnerships have been established with various food data providers that provide the FIS with mostly trusted data. Furthermore, the open source data provider OpenFoodFacts [5] provides data that has been collected and compiled by an Internet community. These data have many inconsistencies and missing elements, so these data have to be specially examined for their quality. Overall, the FIS accesses the data of the following listed providers.

- Federal Food Key (German: Bundeslebensmittelschlüssel) [6]

- WikiFood (food product database) [7]
- Das-Ist-Drin (food product database) [8]
- OpenFoodFacts (food product database)
- Danone and Nutricia (diary-based and medical food companies) [9][10]

The providers of the above data sources have agreed to provide their data for the DiDiER project. For further use of this data, they were integrated into the central FIS database. For this, inconsistencies and missing data have to be corrected and contadictions between the data must be eliminated. The following sections provide more information about the data source providers shown above.

1) Federal Food Key

The Federal Food Key, called in German Bundeslebensmittelschlüssel (BLS), is a German nutrient database maintained by the Max Ruber Institute (MRI). The BLS contains nutrient information from approximately 15,000 foods collected by the MRI through literature research. However, the BLS provides information about standard foods and their compositions.

2) WikiFood

The European food product database WikiFood (WF) was developed by the Luxembourg Institute of Science and Technology (LIST) and supplies information about food products whose database has been filled by a European network of volunteers. The WF community writes ingredient lists of product packaging, reviews it and makes it public. Furthermore, many food manufacturers are involved in wording the databases. In addition to information about the manufacturer, content, ingredients and nutritional value, WF provides additional information about food allergens that are subject to labeling requirements.

3) Das-Ist-Drin

Similar to WF, the German product database Das-Ist-Drin (DID) provides information about food products. The database was developed by the organization snoopmedia and allows users to enter registered data on food products into the database. Similar to WF, DID provides information about allergens that require labeling.

4) OpenFoodFacts

OpenFoodFacts (OFF) is a open source database in which every user around the world can enter product data of food products. OFF delivers the most food data sets in the DiDiER Project and provides over 500 000 product data sets from many countries and regions. Like in most other crowed sourced food databases, the quality of many data sets remains often unclear and depends largely on the motivation of the individual data provider.

5) Danone and Nutricia

Danone as a german diary-based food company, provides all the data of all its food products to the DiDiER Project. Nutricia is the medical division of Danone and provides data of its medical drink and probes food.

Table I shows which food data elements is offered by the respective provider.

TABLE I. COMPARISON OF FOOD DATA PROVIDER

Data Elements	***BLS***	***WF***	***DID***	***OFF***	***Danone & Nutricia***
Product Related Information	no	yes	yes	yes	yes
Generic Food Information	yes	no	no	no	no
European Article Number (EAN)	no	yes	yes	yes	no
Category	yes	yes	yes	no	no
Brand name	no	yes	yes	yes	yes
Origin	no	yes	yes	yes	yes
Content	no	yes	yes	yes	yes
Ingredient List	no	yes	yes	yes	yes
Nutrition Facts	yes	yes	yes	yes	yes
Main Allergens	no	yes	yes	yes	yes

B. ETL Process

The ETL-Process, which is mainly used by data warehouses, has the task of converting data from various data sources into a standardized data structure and storing it in a central database. During the ETL process, mainly the functions listed below are performed [4].

- Identification and extraction of data from relevant (mostly heterogeneous) data sources.
- Clean up the data based on business rules.
- Integration of the data into a uniform format.
- Storing the data in a central database.

When integrating multiple data sources throughout the ETL process, it is important to solve the problem of heterogeneity. The data to be integrated are heterogeneous if the following are true [11].

- The interfaces of the different data sources for data access are different.
- The same facts are presented differently.
- The data models of the external data sources differ.
- The structures of the data models differ.
- The meaning, interpretation and usage of the data models differ.

In order to eliminate the above problems, certain ontologies between food database attributes and knowledge

databases are available for the integration of heterogeneous data sources into the central FIS database. To simplify the ETL process, extensive metadata models of the central database are important [12].

C. Knowledge data

For the creation of ontologies mentioned above, various knowledge databases are available, both for the ETL process and for data quality improvement after data integration. These knowledge databases represent collected information about parts of the data items stored in the food databases. They will be used, either to standardize the data items into a specific domain, for the transformation of data into a given range of values or for a consistency check of the data (e.g. in the context of food databases, if the knowledge data indicates that a certain food contains more than five gram sugar and this food is an ingredient of another food, a consistency check detects, when the other food contains less than five gram sugar). The following knowledge databases are used, either in the ETL process or in the data quality improvement after data integration.

- GeoLite2 city database [13] (data sets with country names and linked city names)
- Country codes database [14] – ISO 3166 (database with country names, linked ISO 3166 codes and linked top level domains)
- Language codes database [15] – ISO 639-1 (databse with language names and linked ISO 639-1 codes)
- Units database [16] - database with metric system units of mass
- Nutrition facts database – database of nutrition fact names collected from the food data sources that are used
- Allergen database [17] – database of 14 ingredients that must be declared as allergens in the EU
- Food category database – database of food categories and linked standard foods (such as pork, milk from the cow, tomatoes, etc.)
- BLS database

Because the BLS database acts as a food composition database, its data can be used as knowledge data to improve the quality of the records of other food databases, as they provide valuable information about the composition of different foods.

D. Data profiling

Data profiling is a process to get information about data. Data is analyzed for missing records, erroneous records or duplicates [18]. The data profiling process must be carried out incrementally after every ETL process in order to always be able to record and correct the inconsistencies of the latest data. It detects incorrect data records, that need to be corrected for a better quality. And it detects missing data or duplicates, that needs to be completed or deleted, respectively [18]. In general, data profiling includes the following tasks [12] [19].

- Recognition of patterns and data types.
- Outlier detection.
- Characterizing missing and default values.
- Data rule analysis (for example, detecting values that match certain regular expressions [20]).
- Analysis of column properties (an analysis performed by checking all values in a column and determining whether the values ar valid or not).
- Analysis of value dependencies accross columns.
- Detection of functional dependencies or foreign key dependencies [19].

III. Food Information Service

As already described, the FIS was developed to transfer food data to the e-health apps (with the scopes fraility and food intolerance) in the JavaScript Object Notation (JSON), a machine-readable data format [21]. Hence, the data is provided in a straightforward and both, human and machine-readable format, for further processing by the food diary apps. Only the data were collected, that are really needed by the respective use cases food intolerance, allergy and malnutrition. Modular and normalized relational data storaging makes it easy, to manipulate individual data items. To keep the clarity, a special database schema has been developed using predefined metadata. The integration of data from the various data sources is realized through the ETL process. After integration, the data is checked for quality using the methods mentioned above and corrected as needed. In Fig. 1, this process is graphically illustrated by the FIS architecture. The central FIS database stores the attributes shown in Table I. Fig. 2 shows the model of the FIS database. For this database, all functional and transitive dependencies were resolved.

A. ETL processing

During the ETL process, the data from external sources are loaded into the central FIS database. All data elements of the various data sources are stored unified. Knowledge databases are used to standardize text elements with different identifiers of the same meaning. Different units of measure are also unified. The figures below are illustrating this process using the data from the OFF database. This is the database with the greatest need for action, to extracting and transforming the data. First, a model was developed that illustrates, which OFF data is stored into which columns of the FIS database tables (Fig. 3).

The values of the OFF columns code, product_name, brands, origins and labels are stored without transforming in the FIS database. (Fig. 4).

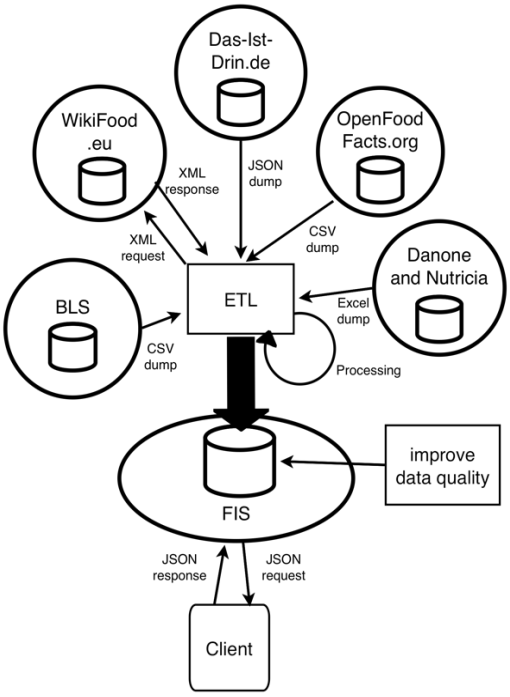

Fig. 1. The FIS architecture

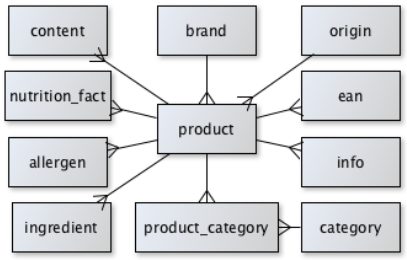

Fig. 2. Model of the FIS database

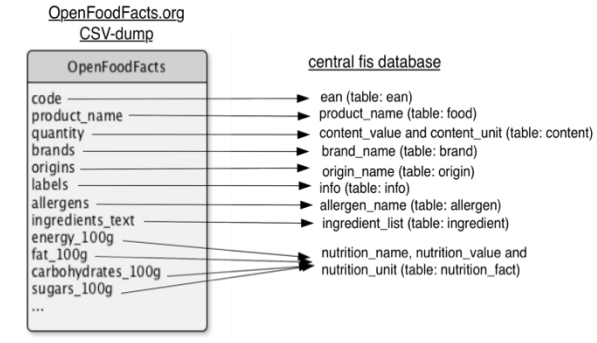

Fig. 3. Assignment of OFF columns to the columns of the FIS database tables

The value of the OFF attribute quantity is a string in form of the following entire regular expression format (ERE) [20].

$$[:digit:]\{1,\}\.[:Digit:]\{1,\}[:space:][:alpha:]\{1,2\} \quad (1)$$

It must be split in the following three formats.

$$[:digit:]\{1,\}\.[:digit:]\{1,\} \quad (2)$$

$$[:space:] \quad (3)$$

$$[:alpha:]\{1,2\} \quad (4)$$

Part (2) stands for any decimal number, part (3) stands for the space character and part (4) stands for one or two characters of the alphabet. The strings matching parts (2) and (4) are stored in the columns content_value and content_unit of the FIS database table content (Fig. 5).

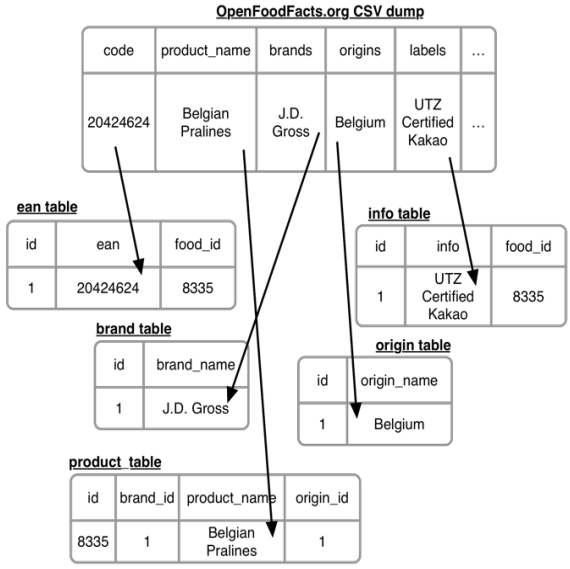

Fig. 4. Storing of the OFF column values code, product_name, brands, origins and labels in the FIS database

Several allergen identifiers for a product are stored as a single string under the OFF attribute named allergens. However, the allergen names must be stored individually in the allergen table of the fis database. For this storage process, the string must be split up in the individual allergen names (Fig. 6).

In addition to the ingredient lists, the language of the ingredient lists, in accordance with the ISO 639-1 standard, is also stored in the corresponding FIS database table. In order to recognize the respective languages of the ingredient lists in the OFF data, the python Application Programming Interface (API) py-googletrans is used (an open source translation API for the programming language Python [22][23]). Based on the language codes knowledge database, the respective language designation is selected according to the ISO 639-1 standard. This name is stored together with the ingredient list string in the FIS database table ingredient (Fig. 7).

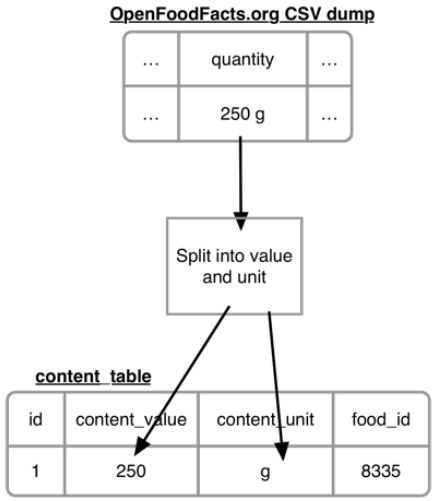

Fig. 5. Storing of the OFF column quantity in the FIS database

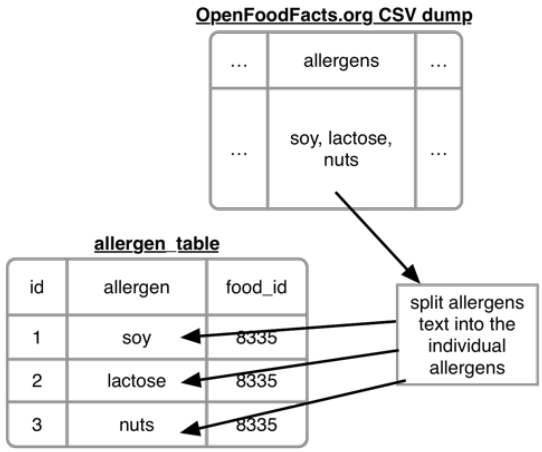

Fig. 6. Storing of the OFF column allergens in the FIS database

In the OFF data, otherwise as in the FIS database, each nutrition fact represents its own table column with the nutrition name as the column name and the nutrition value as the column value. Details about nutrition units of the OFF data are stored separately in a knowledge database table. In addition, there is a knowledge database table for the standardization of nutrition names, in which the nutrition names of all external data sources are assigned to the nutrition names chosen for the FIS. To integrate all OFF nutrition fact data, the assigned FIS nutrition name is first selected from the corresponding knowledge database table using the OFF nutrition name. Subsequently, the (to the nutrition name) related unit is selected from the corresponding knowledge database table. Then, the nutrition facts nutrition name, nutrition value and nutrition unit can be loaded into the FIS database (Fig. 8).

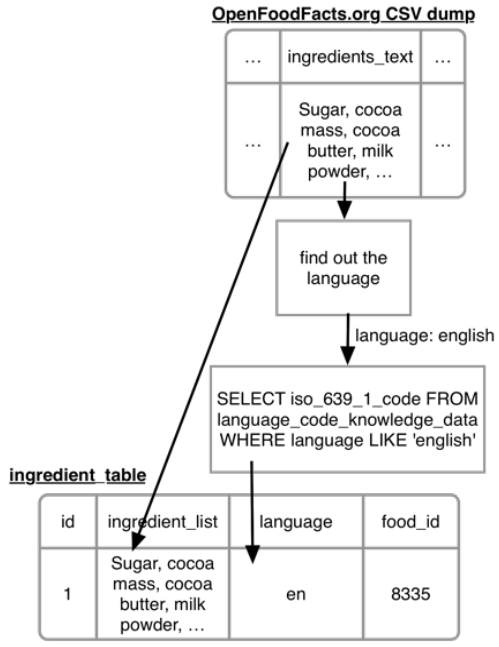

Fig. 7. Storing of the OFF column ingredient_text in the FIS database

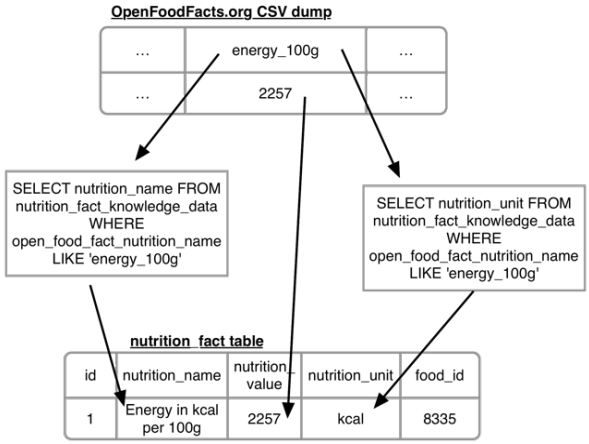

Fig. 8. Storing of nutrition name, nutrition value and nutrition unit of the OFF nutrition fact columns in the FIS database

B. *Data profiling*

In order to optimize the data quality of the FIS database, the data of the FIS database is checked for inconsistencies during the execution of data profiling using consistency rules. Furthermore, the outlier detection is used to examine the outliers of numerical values more precisely for erroneous values. Table II shows the consistency rules used in data profiling. The datatype column specifies the data types to which an attribute value must correspond. The regular expression column contains, as the name indicates, expressions in the regular expression form, to which a respective attribute value must correspond. Furthermore, there is a column named knowledge database. This column

specifies the knowledge databases that are used to examine if there is a specific identifier in it.

TABLE II. CONSISTENCY RULES FOR DATA PROFILING

attributes	*datatype*	*regular expressiosn*	*knowledge database*
product_name	string	^[:alnum:]*$	-
brand_name	string	^[:alnum:]*$	-
origin_name	string	-	Geo2Lite and country codes
info	string	^*$	-
ean	string	(^[:digit:]{8}$)\| (^[:digit:]{13}$)	-
content_value	float	-	-
content_unit	string	-	units of mass
ingredient_list	string	^[:alnum:]*$	-
language	string	-	language codes
allergen	string	-	14 ingredients that must be declared as allergens in the EU
nutrition_name	string	-	nutrition fact names
nutrition_value	float	-	-
nutrition_unit	string	-	units of mass

The stored value in the content_value column must be a floating-point number (the data type is float). The regular expression on the ean column indicates, that the number must be either 8 or 13 characters long. And the value for the column nutrition_name must be found in the nutrition facts knowledge database. If a column value violates any of the consistency rules specified in Table II, the value must be corrected.

C. Data quality measurement

The following functions calculate a metric to assess the quality criterion completeness [24].

$$Q_{completeness}(v) := \begin{cases} 0 & if\ v = NULL\ or\ v \Leftrightarrow NULL \\ 1 & else \end{cases} \quad (5)$$

$$Q_{completeness}(T) := \frac{\sum_{i=1}^{|A|} Q_{completeness}(T.A_i)}{|A|} \quad (6)$$

v is an attribute value in the information system and formula (5) symbolized the quality metric of completeness for v. Formula (6) calculate the quality metric of completeness for a tuple T with the attribute values $T.A_i$. The quality metric of consistency is calculated as follows.

$$r_s(v) := \begin{cases} 1 & if\ v\ satisfied\ the\ consistency\ rule\ r_s \\ 0 & else \end{cases} \quad (7)$$

$$Q_{consistency}(T) := \frac{\sum_{i=1}^{|A|} r_s(T.A_i)}{|A|} \quad (8)$$

Formula (7) symbolizes a consistency rule for the value v and formula (8) calculate the quality metric of completeness for a tuple T with the attribute values $T.A_i$.

In the following, the values for the quality metrics completeness and consistency with respect to 1000 randomly selected FIS data sets, are calculated with the aid of the above-mentioned functions. When calculating the consistency measured value, null values are not taken into account, as this would distort the result. Table III and Table IV show the calculated measured values from completeness and consistency to the respective FIS database columns, before and after the recognition of missing and inconsistent data and their corrections. In brackets next to the calculated measured values, the ratio of missing or consistent data to the total data is shown.

TABLE III. RESULTS OF COMPLETENESS QUALITY MEASUREMENT

attribute name	*completeness quality measure values before data profiling and correction*	*completeness quality measure values after data profiling and correction*
product_name	1 (1000/1000)	1 (1000/1000)
brand_name	0,846 (846/1000)	0,942 (846/898)
origin_name	0,809 (809/1000)	0,896 (809/903)
info	0,622 (622/1000)	1 (622/622)
ean	0,912 (912/1000)	1 (912/912)
content_value	0,794 (794/1000)	0,875 (794/909)
content_unit	0,802 (802/1000)	0,882 (802/909)
ingredient_list	0,803 (803/1000)	0,85 (803/945)
language	0,791 (791/1000)	0,842 (796/945)
allergen	1 (1102/1102)	1 (1102/1102)
nutrition_name	1 (2728/2728)	1 (2728/2728)
nutrition_value	0,812 (2216/2728)	0,812 (2216/2728)
nutrition_unit	1 (2728/2728)	1 (2728/2728)

TABLE IV. RESULTS OF CONSISTENCY QUALITY MEASUREMENT

attribute name	*consitency quality meassure values before data profiling and correction*	*consitency quality meassure values after data profiling and correction*
product_name	1 (1000/1000)	1 (1000/1000)
brand_name	1 (846/846)	1 (846/846)

attribute name	*consitency quality meassure values before data profiling and correction*	*consitency quality meassure values after data profiling and correction*
origin_name	0,721 (583/809)	0,934 (756/809)
info	1 (622/622)	1 (622/622)
ean	0,779 (710/912)	1 (710/710)
content_value	1 (794/794)	1 (794/794)
content_unit	0,739 (582/802)	0,971 (765/788)
ingredient_list	0,991 (796/803)	1 (803/803)
language	1 (796/796)	1 (796/796)
allergen	0,829 (912/1102)	0,933 (1028/1102)
nutrition_name	1 (2728/2728)	1 (2728/2728)
nutrition_value	1 (2216/2216)	1 (2216/2216)
nutrition_unit	1 (2728/2728)	1 (2728/2728)

D. Results of data quality improvement

The result shows, that the quality of the data has improved. For data of unpackaged food, getting from the BLS data source, there is no information about brand names, origins, EAN, contents, ingredient lists and their languages, therefore, the rows with occuring null values was deleted. The rows of the table info with null values in the info column were all deleted, because the product related information for the use of the product data is not absolutely necessary. Completing missing nutrition values proves to be very difficult. For this, complex relationships between the nutritional values of the BLS (if available) and the ingredients or the nutrition facts of the different products must be established. The use of data mining methods using the BLS nutrition values and similar product data of other data sources could also provide positive results here [25]. EAN values that violating the consistency rule, during the measurement of the consistency, were deleted, as a correction of missing or incorrect characters of the EAN is not possible without further information. The allergens with names or acronyms where the correct allergen was not sufficiently recognizable (for example, a "ce" could stand for celery, crustaceans or cereals) were deleted for the same reason. The deletions also improved the quality measured value of the data.

IV. CONCLUSION AND FURTHER WORKS

Through the ETL process, a promising data basis for the data processing and forwarding to e-health apps by the FIS, was established. With data profiling, duplicated, erroneous, inconsistent and incomplete data can be detected and handled. Through quality metrics completeness and consistency could be quantified, and thus the need for action in quality improvement could be recognized. Overall, this improved the data quality of the FIS database, even though it is not yet optimal. In particular, the recognition and improvement of data quality using consistency rules, has produced good results. Although not all missing data sets could be completed and not all inconsistencies could be fixed, nevertheless, upon closer examination of the data, important findings were collected from the data metrics. These findings can in turn be incorporated into the ETL and quality improvement process. Often missing or incorrect data was deleted, but also the absence of such data increases the trust of the data user in the remaining data. To further optimize data quality, further work on the project involves detection and correction of erroneous data, as well as completion of missing data, by the formation of complex ontologies between similar product data and between product data and the nutrition facts of the BLS database. Furthermore, data mining is used to detect similarities or relations between the data. In addition, the data will be evaluated in the next project steps on the basis of further quality criteria (for example, believability, timeliness, accessability, etc.).

REFERENCES

[1] A. Arens-Volland, N. Rösch, F. Feidert, R. Herbst, R. Mösges, "Change frequency of ingredient descriptions and free-of labels of food items concern food allergy sufferers", Allergy: European Journal of Allergy and Clinical Immunology, 394, Volume 65, 2010

[2] P. Elfert, M. Eichelberg, J. Tröger, J. Britz, J. Alexandersson, et al., "DiDiER Digitized Services in Dietry Counselling for People with Increased Health Risk Related to Malnutrition and Food Allergies. In: Computers and Communications (ISCC), IEEE Symposium on. IEEE, pp. 100–104, 2017

[3] Arens A, Rösch N, Feidert F, Harpes P, Herbst R, Mösges R. Mobile electronic patient diaries with barcode based food identification for the treatment of food allergies. GMS Med Inform Biom Epidemiol. 2008;4(3):Doc14.

[4] J. Lihong, C. Hongming, X. Boyi "A domain ontology approach in the ETL process of Data Warehousing", IEEE International Conference on E-Business Engineering, pp. 30-35, 2010.

[5] OpenFoodFacts, "OpenFoodFacts – World", [Online], Available: https://world.openfoodfacts.org, [Accessed 2018-07-31]

[6] Max Rubner Institue (MRI) Karlsruhe, "Bundeslebensmittelschlüssel", 2017, [Online], Available: https://www.blsdb.de/, [Accessed 2018-07-31]

[7] WikiFood, "WikiFood: Knowing what's inside, the Wiki for foodstuffs", 2013, [Online], Available: http://www.wikifood.eu/wikifood/en/struts/welcome.do, [Accessed 2018-07-31]

[8] snoopmedia GmbH, "das ist drin: gemeinsam besser leben", [Online], Available: http://das-ist-drin.de/, [Accessed 2018-07-31]

[9] Danone GmbH, "Danone", [Online], Available: http://www.danone.de/, [Accessed: 2018-07-31]

[10] Nutricia GmbH, "Nutricia: Advanced Medical Nutrition", 2018, [Online], Available: http://www.nutricia.de/, [Accessed: 2018-07-31]

[11] U. Leser, F. Naumann, "Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen", Dpunkt-Verlag, 2007, ISBN 3898644006

[12] J.E. Olson, "Data Quality: The Accuracy Dimension", Morgan Kaufmann Publishers, 2003, ISBN 9781558608917

[13] MaxMind, "MaxMind: GeoLite2 Free Downloadable Databases", 2018, [Online], Available: https://dev.maxmind.com/geoip/geoip2/geolite2/, [Accessed: 2018-07-31]

[14] International Organization of Standardization (ISO), „Country Codes – ISO 3166“, [Online], Available: https://www.iso.org/iso-3166-country-codes.html, [Accessed: 2018-07-31]

[15] International Organization of Standardization (ISO), „Language Codes – ISO 639“, [Online], Available: https://www.iso.org/iso-639-language-codes.html, [Accessed: 2018-07-31]

[16] National Institute of Standards and Technology (NIST), “The NIST Reference on Constants, Units and Uncertainty”, 2017, [Online], Available: https://physics.nist.gov/cuu/Constants/index.html, [Accessed: 2018-07-31]

[17] Food Safety, “List of 14 Allergens”, 2014, [Online], Available: https://www.fsai.ie/legislation/food_legislation/food_information/14_allergens.html, [Accessed: 2018-07-31]

[18] T.F. Kusumasari, Fitria, “Data Profiling for Data Quality Improvement with Openrefine”, IEEE International Conference on Information Technology Systems and Innovation (ICITSI), 2016

[19] Z. Abedjan, L. Golab, F. Naumann, “Data Profiling”, IEEE International Conference on Data Engineering (ICDE), pp. 1432-1435, 2016

[20] The IEEE and The Open Group, “The Open Group Base Specifications Issue 6”, “9. Refular Expressions”, 2004, [Online], Available: http://pubs.opengroup.org/onlinepubs/009695399/basedefs/xbd_chap09.html#tag_09_03_05, [Accessed: 2018-07-31]

[21] ECMA International, “The JSON Data Interchange Syntax”, Standard ECMA-404, 2nd Edition, 2017

[22] SuHun Han, “Googletrans: Free and unlimited Google translate API for Python”, Googletrans 2.3.0 documentation, 2018, [Online], Available: http://py-googletrans.readthedocs.io/en/latest/, [Accessed: 2018-07-31]

[23] Python Software Foundation, “Pythton”, 2018, [Online], Available: https://www.python.org/, [Accessed: 2018-07-31]

[24] B. Heinrich, M. Klier, “Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement”, in: K. Hildebrand, M. Gebauer, H. Hinrichs, M. Mielke (ed.), “Daten und Informationsqualität”, 3. Edition, Springer Verlag, 2015, pp. 49-67

[25] J. Cleve, U. Lämmel, “Data Mining”, De Gruyter Verlag, 2014, ISBN 9783486713916