A new alignment method based on FoodOn as pivot ontology to manage incompleteness in nutritional legacy data sources

Patrice BUCHE[a, b,1], Julien CUFI[b] , Stéphane DERVAUX[c] , Juliette DIBIE[c] , Liliana IBANESCU[c] , Alrick OUDOT[b] and Magalie WEBER[d]

[a] *LIRMM, Univ Montpellier, CNRS, INRIA GraphIK, Montpellier, France*
[b] *IATE, Univ Montpellier, INRA, CIRAD, Montpellier SupAgro, Montpellier, France*
[c] *UMR MIA-Paris, AgroParisTech, INRA, University Paris-Saclay, Paris, France*
[d] *BIA INRA, Nantes, France*

Abstract. In order to correctly assess the nutritional quality of a meal or a manufactured food product in a given country, the first step is to assess the nutritional values for its ingredients. Food composition databases (FCDBs) available in a lot of countries and managed at national level provide values for energy and nutrients of food components. Unfortunately, values associated with some nutrients of interest may be lacking in the FCDB of the country in which the nutritional quality must be assessed. Finding values associated with nutrients for similar foods in other FCDBs is a way to deal with incompleteness. An additional issue arises because the vocabulary used to describe the ingredients of a meal or a recipe in a given FCDB is usually different from the one used in other ones. In this paper we address the problem of identifying the nutritional value of a recipe's ingredients by querying different FCDBs through FoodOn as pivot ontology. We present a new alignment method between two distinct FCDBs, based on syntactic and semantic approaches, whose vocabulary is previously transformed into an ontology. Our method has been evaluated on Ciqual, the French food nutritional database and USDA, the United States food nutritional. The incompleteness management task based on FoodOn as pivot ontology has been assessed with a real use-case concerning iron, Vitamin B12, Vitamin C nutrients.

Keywords. Ontology alignment, Food composition databases, FoodOn, LanguaL

In the framework of the French Meatylab project including industrial partners, we were asked to propose a solution to combine data retrieved from several Food composition databases (FCDB) [1] managed by different agencies and countries to assess the nutritional values of a recipe. Each agency has its own way of describing a food product, whether in terms of labeling or categorization in different facets. Unfortunately, values associated with some nutrients of interest may be lacking in the FCDB of the country in which the nutritional quality must be assessed. Finding values associated with nutrients for similar foods in other FCDBs is a way to deal with incompleteness. Consequently, our goal is to offer a multi-base query tool based on an ontology to establish links

[1] Corresponding Author

between similar food concepts. This functionality is not available in state of the art tools (eg. EuroFIR FoodExplorer).

As an example, this tool should be able to use a term (for example example ‘Courgette, puree’ from CIQUAL FCDB), to be able to recover all the products and associated nutritional values (e.g. ‘Squash, winter, acorn, cooked, boiled, mashed, without salt’ in USDA FCDB). To achieve this, we use as background knowledge, the LanguaL description [2] associated with the food term defined in each national agency. LanguaL stands for "Langua aLimentaria" or "language of food". These descriptions provide a multi-facets semantic definition of a given food expressed in a standardized vocabulary that we will use to find similarities between food products belonging to the vocabulary of different agencies. More than 40.000 foods used in food composition databases are LanguaL described [5].

Our method also takes into account English labels associated with food products in FCDBs. The pivot of all these vocabularies is FoodOn [3], an ontology dedicated to food description. FoodOn is a food ontology initially based on a conversion of the LanguaL thesaurus. For instance, each specialization terms' hierarchy associated with each LanguaL facet was tranlasted in FoodOn into a specialization concepts' hierarchy. Additionally, FoodOn includes 9.500 food terms imported from the Scientific Information and Retrieval Exchange Network of the US Food and Drug administration food database that are organized in families and described in LanguaL.

Our approach aligns the food products of the different FCDBs on FoodOn, based on LanguaL faceted descriptions (semantic approach) in addition to product labels (syntactic approach). This combination of both approaches permits to overcome both the lack of faceted description for some products and the gaps in a purely syntactic comparison (the same food may be denoted differently in different FCDBs).

The main originality of our alignment approach is to reuse Langual descriptions associated with FCDBs food terms available on Langual website combining relevant alignment methods already known in the state of the art [4]. We will present main principles of the approach and results obtained to deal with the lack of values associated with Vitamin C, Vitamin B12 and iron for a set of Ciqual food products reusing values associated with similar foods in USDA.

References

[1] Pehrsson, P. and Haytowitz, D. (2016). Food composition databases. In Caballero, B., Finglas, P. M., and Toldra_, F., editors, Encyclopedia of Food and Health, pages 16-21. Academic Press, Oxford.

[2] Ireland, J. and Moller, A. (2016). Food classification and description. In Caballero, B., Finglas, P. M., and Toldra_, F., editors, Encyclopedia of Food and Health, pages 1-6. Academic Press, Oxford.

[3] Dooley, D. M., Griffiths, E. J., Gosal, G., Buttigieg, P. L., Hoehndorf, R., Lange, M., Schriml, L. M., Brinkman, F. S. L., and Hsiao,W.W. L. (2018). Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. In npj Science of Food.

[4] Shvaiko, P. and Euzenat, J. (2013). Ontology matching: State of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering, 25:158-176.

[5] LanguaL indexed Datasets (2020) The LanguaL indexed Datasets. http://langual.org/langual_indexed_datasets.asp