



**HAL**  
open science

## How to Manage Incompleteness of Nutritional Food Sources?: A solution using FoodOn as pivot ontology

Patrice Buche, Julien Cufi, Stéphane Dervaux, Juliette Dibie, Liliana Ibanescu, Alrick Oudot, Magalie Weber

### ► To cite this version:

Patrice Buche, Julien Cufi, Stéphane Dervaux, Juliette Dibie, Liliana Ibanescu, et al.. How to Manage Incompleteness of Nutritional Food Sources?: A solution using FoodOn as pivot ontology. International Journal of Agricultural and Environmental Information Systems, 2021, 12 (4), pp.1-26. 10.4018/IJAEIS.20211001.0a4 . hal-03372310

**HAL Id: hal-03372310**

**<https://hal.science/hal-03372310>**

Submitted on 11 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.




Distributed under a Creative Commons Attribution 4.0 International License

# How to Manage Incompleteness of Nutritional Food Sources? A Solution Using FoodOn as Pivot Ontology

Patrice Buche, INRAE, Paris, France


Julien Cufi, INRAE, Paris, France

Stéphane Dervaux, INRAE, Paris, France

 <https://orcid.org/0000-0002-9825-3756>

Juliette Dibie, Université Paris-Saclay, France & AgroParisTech, France & INRAE, Paris, France

Liliana Ibanescu, AgroParisTech, France & INRAE, Paris, France

 <https://orcid.org/0000-0003-3373-437X>

Alrick Oudot, INRAE, Paris, France

Magalie Weber, INRAE, Paris, France

## ABSTRACT

In order to correctly assess the nutritional quality of a raw or manufactured food product, the first step is to obtain the associated nutritional values. Food composition databases (FCDBs) managed at national level provide values for nutrients of foods. 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, and 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 denote a given food in a given FCDB is usually different from the one used in others. In this paper, the authors address the problem of retrieving the nutritional value of foods by querying different FCDBs through FoodOn used as pivot ontology. The article presents a new food source alignment method between two FCDBs. The method has been evaluated on the French and United States food nutritional evaluation. The proposed solution for the incompleteness management task has been assessed with a real use case.

## KEYWORDS

ANSES, Background Knowledge, Food Composition Databases, FoodOn, Incompleteness Management, LanguaL, Ontology Alignment, Ontology Building, USDA

## 1. INTRODUCTION

For national and international food trading, it is a challenge to automatically generate the nutrition information panel required by regulation for raw or manufactured food products in many countries. The first challenge is to identify the nutritional values of the raw or manufactured food product either by its identification in an appropriate Food Composition DataBase (FCDB) (Pehrsson & Haytowitz, 2016), or by designing a specific experimental analysis procedure which may require high expertise,

DOI: 10.4018/IJAEIS.20211001.0a4

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

performant analysis tools and time. Unfortunately, values associated with nutrients of interest for a food 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 commonly used by nutritionists to deal with incompleteness.

This paper addresses the problem of semi-automatically identifying the nutritional value of raw or manufactured food products by querying different Food Composition Databases through a pivot vocabulary (named a master-code approach in Ireland & Moller, 2016) in order to deal with the lack of nutrient values. An additional issue arises because the vocabulary used to describe the ingredients of a food or a recipe in a given FCDB is usually different from the vocabulary used in others.

A lot of efforts have been done during the 30 last years in order to harmonize food nutritional data sources through world wide networks like INFOODS (INFOODS FAO, 2020) or EUROFIR (2020). A lot of standards exist concerning food classification and description systems, as reviewed and compared in Ireland and Moller (2016). LanguaL (2020), a multilingual thesaurus using faceted classification, is used in major food composition databases, *e.g.* in the United States (USDA, 2020), Europe (EuroFIR, 2020) and France (CiquaL, 2020) to define a food item by a set of standard controlled terms. Moreover, FoodOn (Dooley et al., 2018) is an ontology, initially based on a conversion of the LanguaL thesaurus, integrated with other resources and aiming to be the open standard controlled vocabulary for food science. In order to be able to integrate these major FCDBs and terminologies, called food sources, this paper proposes to align on FoodOn (that is therefore used as a pivot) a given food using both its LanguaL and English terminological descriptions commonly available in all FCDBs. In this proposed scenario, two foods from two different food sources being indexed with the same LanguaL description and same terminological English description are assumed to represent the same food.

This paper presents a new method to align a food source on a target one (i.e. FoodOn) using both food LanguaL description and the English terminological description. Our method has two main steps: (1) transformation of the food sources in food ontologies and (2) food product alignment computation based on semantic and syntactic information. In this approach, aligning a new FCDB on FoodOn will take benefit of FCDBs already aligned on FoodOn. Indeed, it allows by transitivity an automatic alignment of the new FCDB on FCDBs already aligned and avoids bilateral alignment efforts between FCDBs. During the French national Meatylab project gathering industrial and academic partners, this approach has been implemented in a new application called MultiDB explorer (MultiDB explorer, 2020) which currently integrates several national FCDBs including CiquaL and USDA. MultiDB explorer has been in particular used to deal with the lack of values in CiquaL for 3 nutrients of interest selected by industrial partners (Vitamin C, Vitamin B12, iron).

The paper is organized as follows. Section 2 presents the food sources. Section 3 gives details about the proposed alignment method and its assessment. Section 4 presents the main functionalities of MultiDB explorer and the assessment of CiquaL enrichment task for 3 nutrients of interest reusing USDA data. This new approach is compared to the state of the art in Section 5. Finally, the authors conclude and present their further work in Section 6.

## 2. AVAILABLE FOOD SOURCES

This section presents the food sources used in this paper to illustrate the proposed approach: LanguaL, a well-known multilingual thesaurus using faceted classification (see Section 2.1), FoodOn, a new international food ontology (see Section 2.2), CiquaL, the French food composition database (see Section 2.3) and finally USDA SR Legacy, the North American food composition database (see Section 2.4). The last two food sources are used as examples of national FCDBs to be linked, using FoodOn as a pivot.

Table 1. LanguaL characteristics and associated facets

CHARACTERISTIC	FACET
FOOD GROUP	A. Product Type including Codex Alimentarius classification for Food and Feeds and other international classifications
FOOD ORIGIN	<ul style="list-style-type: none"> <li>• B. Food Source species of plant or animal, or chemical food source</li> <li>• C. Part of Plant or Animal</li> </ul>
PHYSICAL ATTRIBUTES	E. Physical State, Shape or Form
PROCESSING	<ul style="list-style-type: none"> <li>• F. Extent of Heat Treatment.</li> <li>• G. Cooking method.</li> <li>• H. Treatment Applied.</li> </ul>
PACKAGING	<ul style="list-style-type: none"> <li>• J. Preservation Method</li> <li>• K. Packing Medium</li> <li>• M. Container or Wrapping</li> <li>• N. Food Contact</li> </ul>
DIETARY USES	P. Consumer Group/Dietary use
GEOGRAPHIC ORIGIN	R. Geographic Places and Regions
MISCELLANEOUS CHARACTERISTICS	Z. Adjunct Characteristics of Food

## 2.1 LanguaL

LanguaL (2020) is a well-known multilingual thesaurus using faceted classification (Ireland & Moller, 2000, 2010). LanguaL stands for “Langua aLimentaria” or “language of food”. More than 40,000 foods used in food composition databases are LanguaL described (LanguaL Indexed Datasets, 2020).

LanguaL is based on the following hypothesis: i) any food can be systematically described using a combination of characteristics; ii) these characteristics can be categorized into viewpoints, composed of one or several facets, and coded for computer processing; iii) the resulting food description codes can be used to retrieve food data from FCDBs. Table 1 shows the 14 facets of LanguaL.

Thanks to LanguaL, a food may be described by a combination of descriptors belonging to the standardized vocabulary associated with each facet presented in Table 1. A food can be described by several descriptors for one facet and it can also have no descriptor for a facet. Each descriptor is defined by a standard controlled term and a code. Each standard controlled term belongs to a specialization hierarchy, that is defined for each facet. An excerpt of the hierarchy of descriptors associated with facet A is presented in Figure 1.

In Table 2, the food *Cooked pork shoulder, choice* belonging to Ciqua database is described by 14 descriptors, *i.e.* a code associated with a controlled term, regrouped in 10 facets. For instance, this food has two descriptors for facet A, but is not described for facet K.

The descriptor of code A0797 associated with the controlled term PRESERVED MEAT (EUROFIR) belongs to the specialization hierarchy associated with facet A. **Product Type** of which an excerpt is presented in Figure 1.

## 2.2 FoodOn

FoodOn (Dooley et al., 2018) 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 translated 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 (called SIREN in the paper) food database that are organized in

Table 2. LanguAL descriptors for food Cooked pork shoulder, choice used in Ciqual database

LANGUAL CODE	LANGUAL TERM
A0279	CURED MEAT (US CFR)
A0797	PRESERVED MEAT (EUROFIR)
B1136	SWINE
C0270	SKELETAL MEAT PART, WITHOUT BONE AND SKIN, WITHOUT SEPARABLE FAT
E0137	SLICED
F0014	FULLY HEAT-TREATED
G0003	COOKING METHOD NOT APPLICABLE
H0253	CURED OR AGED
H0367	SALT ADDED
J0100	PRESERVED BY ADDING CHEMICALS
J0131	PRESERVED BY CHILLING
P0024	HUMAN CONSUMER, NO AGE SPECIFICATION
Z0010	CHOICE GRAD
Z0043	SHOULDER (MEAT CUT)

Figure 1. An excerpt of the specialization hierarchy of facet A in LanguAL with the controlled term PRESERVED MEAT (EUROFIR) presented in bold

- [-] A. PRODUCT TYPE
  - [+] DIETARY SUPPLEMENT
  - [+] FOOD ADDITIVES
  - [-] PRODUCT TYPE, EUROPEAN UNION
    - [+] CIAA FOOD CLASSIFICATION FOR FOOD ADDITIVES
    - [+] CLASSIFICATION OF PRODUCTS OF PLANT AND ANIMAL ORIGIN, EUROPEAN COMMUNITY
    - [+] EFSA FOOD CLASSIFICATION AND DESCRIPTION SYSTEM FOR EXPOSURE ASSESSMENT (EFSA FOODEX2)
    - [+] EUROCODE 2 FOOD CLASSIFICATION
    - [-] EUROFIR FOOD CLASSIFICATION
      - [+] BEVERAGE (NON-MILK) (EUROFIR)
      - [+] EGG OR EGG PRODUCT (EUROFIR)
      - [+] FAT OR OIL (EUROFIR)
      - [+] FRUIT OR FRUIT PRODUCT (EUROFIR)
      - [+] GRAIN OR GRAIN PRODUCT (EUROFIR)
      - [-] MEAT OR MEAT PRODUCT (EUROFIR)
        - MEAT ANALOGUE (EUROFIR)
        - MEAT DISH (EUROFIR)
        - OFFAL (EUROFIR)
        - POULTRY MEAT (EUROFIR)
        - **PRESERVED MEAT (EUROFIR)**
        - RED MEAT (EUROFIR)
        - SAUSAGE OR SIMILAR MEAT PRODUCT (EUROFIR)
      - [+] MILK, MILK PRODUCT OR MILK SUBSTITUTE (EUROFIR)
      - [+] MISCELLANEOUS FOOD PRODUCT (EUROFIR)
      - [+] NUT, SEED OR KERNEL (EUROFIR)
      - [+] PRODUCT FOR SPECIAL NUTRITIONAL USE OR DIETARY SUPPLEMENT (EUROFIR)
      - [+] SEAFOOD OR RELATED PRODUCT (EUROFIR)
      - [+] SUGAR OR SUGAR PRODUCT (EUROFIR)
      - [+] VEGETABLE OR VEGETABLE PRODUCT (EUROFIR)
    - [+] EUROPEAN FOOD GROUPS (EFG)
  - [+] PRODUCT TYPE, INTERNATIONAL

Figure 2. SIREN term *pork shoulder (cooked, cured)* in FoodOn

ID	<a href="http://purl.obolibrary.org/obo/FOODON_03310969">http://purl.obolibrary.org/obo/FOODON_03310969</a>
comment	SIREN DB annotation: * has quality 'solid' ( <a href="http://purl.obolibrary.org/obo/FOODON_03430151">http://purl.obolibrary.org/obo/FOODON_03430151</a> ) * has quality 'fully heat-treated' ( <a href="http://purl.obolibrary.org/obo/FOODON_03440014">http://purl.obolibrary.org/obo/FOODON_03440014</a> ) * has quality 'shoulder (meat cut)' ( <a href="http://purl.obolibrary.org/obo/FOODON_03530043">http://purl.obolibrary.org/obo/FOODON_03530043</a> ) * derives from 'skeletal meat part, without bone or shell' ( <a href="http://purl.obolibrary.org/obo/FOODON_03420125">http://purl.obolibrary.org/obo/FOODON_03420125</a> ) * formed as a result of 'curing or aging process' ( <a href="http://purl.obolibrary.org/obo/FOODON_03460253">http://purl.obolibrary.org/obo/FOODON_03460253</a> )
database cross reference	SUBSET_SIREN:F10969
has curation status	<a href="http://purl.obolibrary.org/obo/IAO_0000428">http://purl.obolibrary.org/obo/IAO_0000428</a>
imported from	<a href="http://langual.org">http://langual.org</a>
inSubset	subset_siren
label	pork shoulder (cooked, cured)

families and described in LanguaL. Let us notice that the LanguaL descriptions of SIREN food terms are only available in a comment property in the current version of FoodOn. The LanguaL description for SIREN term *pork shoulder (cooked, cured)* in FoodOn is given in the comment property as presented in Figure 2.

### 2.3 Ciqual

Ciqual (Ciqual (2020)) is the French food nutritional composition database. It is managed by ANSES, the French Agency for Food, Environmental and Occupational Health and Safety. Ciqual 2017 version used in this paper includes 2,807 foods, of which 1,797 are fully described using LanguaL (see an example in Table 1). Up to 61 nutritional values, corresponding to different constituents, may be associated with a food (see Figure 3). Values are given for 100g of comestible food (e.g. without bone for meat).

### 2.4 USDA SR Legacy

USDA SR Legacy (2020) is the major source of food nutritional composition data in the United States. It is managed by USDA, the US department of agriculture. USDA SR Legacy 2018 version used in this paper includes 8,618 foods of which 5,137 have been fully described using LanguaL (see an example in Table 3). Up to 150 nutritional values, corresponding to different constituents may be associated with a food.

Table 3 gives the descriptors for a food, *Pork, cured, shoulder, arm picnic, separable lean and fat, roasted*. Let us notice that six of its descriptors—A0279, B1136, F0014, H0253, P0024 and Z0043—are the same as those used in Table 2, for *Cooked pork shoulder, choice*, but both foods are different.

## 3. A METHOD TO ALIGN FOOD SOURCES

This paper proposes a food sources alignment method using both LanguaL food description and food terminological description in English. The method is composed of two main steps. In the first one, the food sources are transformed into food ontologies using both LanguaL description and food terminological description in English. The second step consists in computing a new similarity score between two food concepts combining a syntactic similarity between food labels associated with food terminological descriptions in English and a semantic similarity between food LanguaL descriptions. Results of the alignment method are equivalence or subsumption relations between food concepts called respectively food matches and family matches in the following.

Figure 3. An excerpt of the nutritional values for Cooked pork shoulder, choice in Ciquil database

Nutritional information							
Component name	Value	Unit	Matrix unit	Value type	Method type	Method indicator	Reference type
zinc	2.94	milligram	per 100g edible portion	best estimate	Other method type	Analytical or calculation method not known	Webpage
water	71.7	gram	per 100g edible portion	best estimate	Other method type	Analytical or calculation method not known	File or Database
vitamin K-1	0	microgram	per 100g edible portion	logical zero	Other method type	Analytical or calculation method not known	Webpage
vitamin E: alpha-tocopherol equiv from E vitamer activities	0.26	alpha-tocopherol equivalent	per 100g edible portion	best estimate	Other method type	Analytical or calculation method not known	File or Database
vitamin D	0.9	microgram	per 100g edible portion	best estimate	Other method type	Analytical or calculation method not known	Webpage

Table 3. LanguaL descriptors for food Pork, cured, shoulder, arm picnic, separable lean and fat, roasted used in USDA database

LANGUAL CODE	LANGUAL TERM
A0279	CURED MEAT (US CFR)
A1280	PORK PRODUCTS (USDA SR)
B1136	SWINE
C0269	SKELETAL MEAT PART, WITHOUT BONE AND SKIN, WITH SEPARABLE FAT
E0150	WHOLE, NATURAL SHAPE
F0014	FULLY HEAT-TREATED
G0005	BAKED OR ROASTED
H0253	CURED OR AGED
J0108	PRESERVED BY TREATMENTS WITH CHEMICALS
P0024	HUMAN CONSUMER, NO AGE SPECIFICATION
Z0043	SHOULDER (MEAT CUT)

In the first step, a nested structure like JSON has not been chosen as it does not allow to represent the specialisation relation between food concepts. The csv format is not a nested structure and does not allow to represent the specialisation relation between food concepts. OWL has been chosen for three reasons. First, OWL allows to represent the three kinds of information used by the alignment method: the specialisation relation between food concepts, the labels associated with foods concepts and the LanguaL description associated with food concepts. Second, the SIREN part of FoodOn

which is studied in this paper is already available as an OWL ontology. Third, OWL is an international standard which favors open science.

### 3.1 Transforming a Food Source into an Ontology Using LanguaL

In this section, we present how to transform a food sources  $f_s$  into a food ontologie  $O_{fs}$  using both LanguaL description and food terminological description in English. First, a semantic representation of LanguaL thesaurus, called  $O_{Lang}$  is defined. Second, the definition of the ontology  $O_{fs}$  is provided.

In this paper, an ontology is defined as follows.

**Definition 1.** An ontology is a tuple  $O = \langle C, R, L, \leq_o, \theta \rangle$  such that:

1.  $C$  is a set of concepts;
2.  $L$  is a set of labels associated with concepts;
3.  $R$  is a set of relations in  $C \times C$  ;
4.  $\leq_o$  is a specialisation relation in  $C \times C$  ;
5.  $\theta$  is a denotation function from  $2^C$  to  $L$ .

**Definition 2 (Ontology  $O_{Lang}$ ).** Given the set of LanguaL facets  $h_i, i \in [1, 14]$ , the set of controlled terms  $t_j^i$  belonging to the specialization hierarchy associated with  $h_i$ , and  $\leq$  the specialisation relation between terms  $t_j^i$ , we denote  $t_{h_i}$  the father term of the specialization hierarchy associated with  $h_i$  such that  $\forall j, t_j^i \leq t_{h_i}$ .  $O_{Lang}$  is defined by the tuple  $\langle C_{Lang}, R_{Lang}, L_{Lang}, \leq_o, \theta \rangle$  with  $L_{Lang}$  the set of labels  $t_j^i$  such that  $\theta(c_{t_j^i}) = t_j^i$ ,  $C_{Lang}$  the set of concepts  $c_{t_j^i}$  corresponding to  $t_j^i$  (i.e.  $\theta(c_{t_j^i}) = t_j^i$ ) such that  $c_{t_j^i} \leq_o c_{t_j^i}$  if  $t_j^i \leq t_j^i$  and  $\forall j, c_{t_j^i} \leq_o c_{t_{h_i}}$  with  $\theta(c_{t_{h_i}}) = t_{h_i}$ ,  $R_{Lang}$  the empty set.

It may be noticed that  $O_{Lang}$  is an ontological representation of the LanguaL thesaurus using faceted classifications. It corresponds to a subpart of FoodOn ontology.

**Example 1.** Based on the excerpt of the specialization hierarchy associated with facet **A** presented

in Figure 1, we denote  $c_{MEAT\_OR\_MEAT\_PRODUCT} \in C_{Lang}$  the concept such that  $\theta(c_{MEAT\_OR\_MEAT\_PRODUCT}) = MEAT\_OR\_MEAT\_PRODUCT(EUROFIR)$ ,  
 $c_{PRESERVED\_MEAT} \in C_{Lang}$  the concept such that  $\theta(c_{PRESERVED\_MEAT}) = PRESERVED\_MEAT(EUROFIR)$  and  
 $c_{PRESERVED\_MEAT} \leq_o c_{MEAT\_OR\_MEAT\_PRODUCT}$ . URI associated with  $c_{MEAT\_OR\_MEAT\_PRODUCT}$  (resp.  $c_{PRESERVED\_MEAT}$ ) in FoodOn are [http://purl.obolibrary.org/obo/FOODON\\_03400793](http://purl.obolibrary.org/obo/FOODON_03400793) (resp. [http://purl.obolibrary.org/obo/FOODON\\_03400797](http://purl.obolibrary.org/obo/FOODON_03400797)).

The ontology  $O_{fs}$  is built from a given food source  $f_s$  using food LanguaL description and the expressiveness of OWL semantic language (OWL Language). As shown in Table 2, a food may be described by several facets  $h_i, i \in [1, 14]$ , and the LanguaL description of a food for a given facet  $h_i$  may be multi-valued with several descriptors denoted by  $c_{descr_1}^{h_i}, \dots, c_{descr_m}^{h_i}$  with  $c_{descr_j}^{h_i} \in C_{Lang}$ . We propose to use the someValueFrom OWL cardinality restriction to associate a food  $c_f$  with its



LanguaL description. Therefore, given a facet  $h_i$ ,  $i \in [1, 14]$ ,  $j \in [1, m_i]$ ,  
 $\leq_o \left( c_f, \text{someValuesFrom} \left( \text{has\_Facet}, c_{descr_j}^{h_i} \right) \right)$  means that each instance of  $c_f \in O_{f_s}$  must be  
 linked to at least one instance of  $c_{descr_j}^{h_i}$ .

**Definition 3 (Ontology  $O_{f_s}$ ).** Given a food source  $f_s$ , each food  $f_k \in f_s$  is defined by its label  $t_{f_k}$   
 and its LanguaL description on several facets  $h_i$ ,  $i \in [1, 14]$ . The LanguaL description of  $f_k$  for  
 a given facet  $h_i$  is denoted by  $c_{descr_1}^{h_i}, \dots, c_{descr_{m_i}}^{h_i}$  with  $c_{descr_j}^{h_i} \in C_{Lang}$ .

$O_{f_s}$  is defined by the tuple  $\langle C_{O_{f_s}}, R_{O_{f_s}}, L_{O_{f_s}}, \leq_o, \theta \rangle$  with

- $L_{O_{f_s}} = L_{f_s} \cup L_{Lang}$  where  $L_{f_s}$  is the set of labels  $t_{f_k}$  such that  $\theta \left( c_{f_k} \right) = t_{f_k}$ ,
- $R_{O_{f_s}}$  the set composed of the single relation *has\_Facet*,
- $C_{O_{f_s}} = C_{f_s} \cup C_{Lang}$  where  $C_{f_s}$  is the set of concepts  $c_{f_k}$  corresponding to  $f_k$  such that  
 $\forall i \in [1, 14], \forall j \in [1, m_i], \leq_o \left( c_{f_k}, \text{someValuesFrom} \left( \text{has\_Facet}, c_{descr_j}^{h_i} \right) \right)$

In the OWL version of  $O_{f_s}$ , the set of concepts  $C_{O_{f_s}}$  corresponds to a set of OWL classes and  
 the set of binary relations  $R_{O_{f_s}}$  corresponds to a set of OWL object properties. The denotation function  
 $\theta$  corresponds to the property *skos:prefLabel*. Indeed, *skos:prefLabel* has been preferred to *rdfs:label*  
 because the property *skos:hiddenLabel* is also used to represent the lemmatized words associated  
 with a given label.

It must be noticed that the set of controlled terms associated with LanguaL facets are modeled  
 as concepts in Definition 2 (which correspond to OWL classes) as they are used to define in a general  
 way a given food concept. Moreover, in  $O_{f_s}$ , LanguaL descriptors already represented as OWL  
 classes in FoodOn are reused. For example, the descriptor CURED\_MEAT\_US\_CFR of Example 2  
 is represented by the OWL class associated with the URI [http://purl.obolibrary.org/obo/FOODON\\_03400279](http://purl.obolibrary.org/obo/FOODON_03400279).

In the Appendix, an excerpt of the OWL definition of the concept which corresponds to the food  
*Cooked pork shoulder, choice* presented in Table 2 is given.

### 3.2 Alignment Method

In this section, a new alignment method is presented for two food sources  $f_{s_1}$  and  $f_{s_2}$  relying on  
 their corresponding ontology  $O_{f_{s_1}}$  and  $O_{f_{s_2}}$  as given in Definition 3.  $O_{f_{s_1}}$  is considered as the source  
 ontology to be aligned and  $O_{f_{s_2}}$  as the target one. The proposed alignment method consists in finding  
 the best match in  $O_{f_{s_2}}$  for each food concept of  $O_{f_{s_1}}$ . For that, each food concept  $f \in O_{f_{s_1}}$  is  
 associated with a ranked list of food concepts  $g_i \in O_{f_{s_2}}$ . The final similarity score between  $f \in O_{f_{s_1}}$   
 and each  $g_i \in O_{f_{s_2}}$  is computed from a combination of a syntactic similarity score relying on the  
 concepts' labels and a semantic similarity score using the LanguaL food descriptions.

### 3.2.1 Syntactic Similarity

The definition of the syntactic similarity between a food concept  $f$  of the source ontology  $O_{fs_1}$  and each food concept of the target ontology  $O_{fs_2}$ , is based on their labels.

First, with each food concept  $f$  is associated a set of lemmatized words denoted  $W_f$  computed using the Python NLTK (Natural Language Toolkit from Bird et al. 2009) and the English Pen Treebank tagset (English Pen Treebank tagset (2020)) as follows:

$$\begin{aligned} w1 &= \text{PartOfSpeech\_Tagging}(\theta(f)) \\ w2 &= \text{WorldNetLemmatizer}(w1) \text{ using the WordNet Database} \\ W(f) &= \text{Delete\_Stop\_Words}(w2) \end{aligned}$$

**Example 2.** Let us consider the food concept having for label **Cooked pork shoulder, choice**. This label is transformed by the pre-processing step into the set of lemmatized words  $\{cook, pork, shoulder, choice\}$ .

The syntactic similarity between two food concepts  $f$  and  $g$  is computed using the cosine similarity:  $SyntSim(f, g) = \cos(W_f, W_g)$ .

**Example 3.** Let us consider the food concept  $f$  having for label **Cooked pork shoulder** and the food concept  $g$  having for label **Pork: Cooked pork shoulder**. Then  $W_f = \{cook, pork, shoulder\}$ ,  $W_g = \{cook, pork, pork, shoulder\}$  and  $SyntSim(f, g) = \cos(W_f, W_g) = 0,94$ .

**Definition 4 (Syntactic similarity).** Given a food concept  $f$  of the source ontology  $O_{fs_1}$  and the target ontology  $O_{fs_2} = \langle C_{O_{fs_2}}, R_{O_{fs_2}}, L_{O_{fs_2}}, \leq_o, \theta \rangle$  with  $C_{O_{fs_2}} = \{g_1, \dots, g_n\}$ , the syntactic similarity of  $f$  with  $O_{fs_2}$  is the list of couples

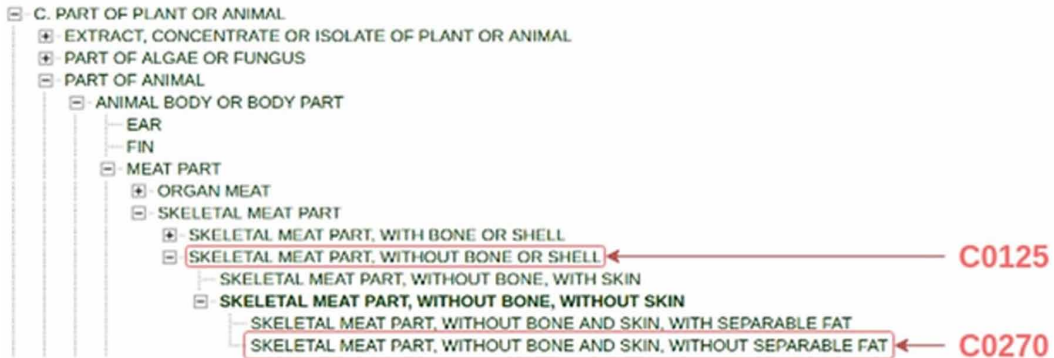
$$List\_SyntSim(f, O_{fs_2}) = \left\{ (g_1, SyntSim(f, g_1)), (g_2, SyntSim(f, g_2)), \dots, (g_n, SyntSim(f, g_n)) \right\}$$

### 3.2.2 Semantic Similarity

The semantic similarity between a food concept  $f$  of the source ontology  $O_{fs_1}$  and each food concept of the target ontology  $O_{fs_2}$  relies on their LanguaL descriptions. In order to compute the semantic similarity  $SemSim(c_{descr^f, h_i}, c_{descr^g, h_i})$  between two descriptors  $c_{descr^f, h_i}$  and  $c_{descr^g, h_i}$  of two food concepts  $f$  and  $g$  for the LanguaL facet  $h_i$ , the Wu-Palmer similarity (Wu and Palmer (1994)) is used; that allows one to compute how closely these descriptors are related in the specialization hierarchy of the facet. Notice that the Wu-Palmer similarity, recalled in Equation 1, assumes that the similarity between two concepts is the function of the depth of both concepts and least common ancestor in the hierarchy of concepts.

$$SemSim(c_1, c_2) = WuPalmer(c_1, c_2) = \frac{2 \times \text{depth}(lcs(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \quad (1)$$

Figure 4. An excerpt of the specialization hierarchy of facet C in  $O_{Lang}$



where  $depth(c_i)$  is the depth of the concept  $c_i$  in the considered specialization hierarchy belonging to  $O_{Lang}$ , and  $lcs(c_1, c_2)$  is the least common ancestor of  $c_1$  and  $c_2$  in this hierarchy.

**Example 4.** Let us consider the food concept  $f \in O_{fs_1}$  having for label *Cooked pork shoulder, choice* presented in Table 2. Let us denote by  $c_{C0270}$  its descriptor for facet C associated with the term **SKELETAL MEAT PART, WITHOUT BONE AND SKIN, WITHOUT SEPARABLE FAT**. Let us also consider the food concept  $g \in O_{fs_2}$  having for label *pork shoulder (cooked, cured)* presented in Figure 2 that has for descriptor the concept  $c_{C0125}$  for facet C associated with the term **SKELETAL MEAT PART, WITHOUT BONE OR SHELL**. As shown in Figure 4,

$$SemSim(c_{C0270}, c_{C0125}) = \frac{2 * 7}{7 + 9} = 0,875 \text{ with } Depth(c_{C0270}) = 9, \text{ } Depth(c_{C0125}) = 7 \text{ and}$$

$$lcs(c_{C0270}, c_{C0125}) = c_{C0125}.$$

Given a LanguaL facet, a food concept may be described by several descriptors as already mentioned (see the example in Figure 1). As different annotators belonging to different agencies have defined those descriptors, it is not obvious that similar concepts from two distinct food sources have been defined with the same descriptors for a given facet. Consequently, we propose to define the similarity on a facet for two food concepts belonging respectively to  $O_{fs_1}$  and  $O_{fs_2}$  as the maximum LanguaL descriptors similarities for all descriptor pairs.

**Definition 5 (Facet semantic similarity).** Let us consider a LanguaL facet  $h_i$ , a food concept  $f \in O_{fs_1}$  with its LanguaL description for  $h_i$  being denoted by  $c_{descr_1}^{h_i}, \dots, c_{descr_{m_k}}^{h_i}$  with  $c_{descr_j}^{h_i} \in C_{Lang}$  and a food concept  $g \in O_{fs_2}$  with its LanguaL description for  $h_i$  being denoted by  $c_{descr_1}^{h_i}, \dots, c_{descr_{m_l}}^{h_i}$  with  $c_{descr_j}^{h_i} \in C_{Lang}$  according to Definition 3. Their semantic similarity on facet  $h_i$  is defined by Equation 2:

$$FacetSemSim(f, g, h_i) = \max_{o=1, m_k} \max_{p=1, m_l} SemSim(c_{desc_o}^{h_i}, c_{desc_p}^{h_i}) \quad (2)$$

Let us notice that if the food concepts  $f$  or  $g$  are not defined for a given facet  $h_i$  then  $FacetSemSim(f, g, h_i) = 0$ . Since LanguaL facets do not have the same importance in the definition of a given food, the semantic similarity for two food concepts is defined as the weighted sum of the fourteen facets' semantic similarities, the higher the weight, more important is the facet.

**Definition 6 (Food semantic similarity).** Given a food concept  $f \in O_{fs_1}$  and a food concept  $g \in O_{fs_2}$ , their food semantic similarity is defined by Equation 3:

$$FoodSemSim(f, g) = \frac{\sum_{i=1}^{14} w(h_i) * FacetSemSim(f, g, h_i)}{\sum_{i=1}^{14} w(h_i)} \quad (3)$$

where  $w(h_i)$  is the weight factor associated with the facet  $h_i$ .

Section 3.3.1 presents the different tests run to empirically determine the weight associated with facets.

**Example 5.** Let us consider the food concept  $f \in O_{fs_1}$  having for label **Cooked pork shoulder, choice** presented in Table 2 and the food concept  $g \in O_{fs_2}$  having for label **pork shoulder (cooked, cured)** presented in Figure 2. Thanks to facet descriptors associated with  $f$  and  $g$  and weighted similarities given in Figure 5, then

$$FoodSemSim(f, g) = \frac{4 + 4 + 3.5 + 0.6 + 1 + 1 + 1}{4 + 4 + 4 + 1 + 1 + 1 + 1} = \frac{15.1}{18} = 0.8388. \text{ In this example, the Food semantic similarity is high which indicates that both food products are similar.}$$

The semantic similarity of a food concept  $f \in O_{fs_1}$  with the food concepts of  $O_{fs_2}$  is defined as follows.

**Definition 7 (Semantic similarity).** Given a food concept  $f$  of the source ontology  $O_{fs_1}$  and the target ontology  $O_{fs_2} = \langle C_{O_{fs_2}}, R_{O_{fs_2}}, L_{O_{fs_2}}, \leq_o, \theta \rangle$  with  $C_{O_{fs_2}} = \{g_1, \dots, g_n\}$ , the semantic similarity of  $f$  with  $O_{fs_2}$  is the list of couples

$$List\_SemSim(f, O_{fs_2}) = \{(g_1, FoodSemSim(f, g_1)), (g_2, FoodSemSim(f, g_2)), \dots, (g_n, FoodSemSim(f, g_n))\}$$

Finally, the similarity between a food concept  $f \in O_{fs_1}$  with the food concepts of  $O_{fs_2}$  is presented in Algorithm 1 which computes the ranked list of the most similar food concepts  $g_i \in O_{fs_2}$  associated with  $f \in O_{fs_1}$  from the list  $ListSyntSim(f, O_{fs_2})$  as defined in Definition 4 and the  $ListSemSim(f, O_{fs_2})$  as defined in Definition 7.

Figure 5. Facet descriptors associated with  $f \in O_{fs_1}$  and  $g \in O_{fs_2}$  with their weight experimentally defined

f in $O_{fs_1}$	g in $O_{fs_2}$	FacetSemSim	Facet	Weighted similarity
A0279	A0279	1	A	4
B1136	B1136	1	B	4
C0270	C0125	0,875	C	3,5
E0137	E0151	0,6	E	0,6
F0014	F0014	1	F	1
H0253	H0253	1	H	1
Z0043	Z0043	1	Z	1

Facet	Weight
A	4
B	4
C	4
E	1
F	1
G	1
H	1
J	1
K	1
M	1
N	1
P	1
R	1
Z	1

**Algorithm 1** Similarity of  $f \in O_{fs_1}$  with  $O_{fs_2}$

$$C_{O_{fs_2}} = \{g_1, \dots, g_n\}$$

```

result = emptyList()
FORALL  $g_i$  in  $O_{fs_2}$  DO
    score = avg ( SyntSim( $f, g_i$ ), FoodSemSim( $f, g_i$ ) )
    result.addItem( $g_i$ , score)
ENDFOR
RETURN result.sort()
    
```

### 3.3 Alignment Method Assessment

The food sources alignment method presented in Section 3.2 has been tested on Ciqual food database and the SIREN subpart of FoodOn ontology, the foods of both sources being described with LanguaL descriptions.

Let us call  $O_{fs_1}$  the source ontology built from Ciqual food database using Definition 3,  $O_{fs_2}$  the extension of the SIREN subpart of FoodOn (that is already an ontology) also using Definition 3. It must be noticed that  $O_{fs_1}$  (resp.  $O_{fs_2}$ ) is a flat list of food concepts described by their LanguaL facets. The alignment method takes into account the LanguaL description to compute food matches.  $O_{fs_1}$  (resp.  $O_{fs_2}$ ) is available in the dataset <https://doi.org/10.15454/6CEYU3> (resp. dataset <https://doi.org/10.15454/5LLGVY>) stored in OWL format in the institutional INRAE dataverse.  $O_{fs_1}$  (resp.  $O_{fs_2}$ ) is composed of 3863 (resp. 7862) classes and 24218 (resp. 53832) logical axioms.

A set of alignments, called GS (for gold standard) especially designed for this work have been used to test the method. GS has been designed using a subset of 181 food products from Ciqual fully described in LanguaL and including different kinds of foods (fruits and vegetables, meats and fishes,

milk and dairy products). Each food has been manually aligned with the closest food concept in  $O_{fs_2}$ , called *food match*. Let us notice that it was not always possible to find a similar food concept (only possible for 73 food products from 181) in  $O_{fs_2}$ . Manual food alignments have been updated using the results obtained by the alignment method presented in Section 3 when more relevant. It happened for 14 foods upon 73 (20% of food alignments) which shows that this method can also be used in an iterative way to enhance GS quality.

$O_{fs_2}$  is a flat list of food products belonging to the SIREN subpart of FoodOn. The organization in families already defined in the SIREN subpart of FoodOn ontology has been used to determine food family matches. Indeed, it was always possible to find a similar family in  $O_{fs_2}$  for the 181 food products from Ciqual. It is the reason why each food has been manually aligned with its closest family in  $O_{fs_2}$ , called *family match*. In the following, the set of 73 food products from Ciqual associated both with a food match and a family match in  $O_{fs_2}$  is called  $GS_{73}$ . The set of 181 food products from Ciqual only associated with a family match in  $O_{fs_2}$  is called  $GS_{181}$ . It must be noticed that  $GS_{73}$  is included in  $GS_{181}$ .  $GS_{73}$  and  $GS_{181}$  are available in the dataset <https://doi.org/10.15454/BVXD7I> stored in Excel format in the institutional INRAE dataverse. Those datasets include also the LanguaL description of Ciqual foods. The alignment results are presented and discussed in the following.

**Example 6.** The food *beef chuck (raw)* (resp. the food family *beef food product*) from  $O_{fs_2}$  is an example of food match (resp. family match) of the food *Beef, chuck, raw* from  $O_{fs_1}$ .

Table 4 provides comparison results in terms of precision, recall and F-measure between 5 methods: the method presented in Algorithm 1, the syntactic score alone (Def.4), the semantic score alone (Def. 7) and two state of the art alignment tools OnAGUI (OnAGUI (2020)) and AML (Faria et al. (2013)). Results have been obtained using  $GS_{73}$  to assess food matches. AML has been used in automatic mode and OnAGUI using the Levenstein distance. For each alignment tool and for each food concept  $f \in O_{fs_1}$  the decreasing ordered list of the 5 best matches  $g_i \in O_{fs_2}, i \in [1, 5]$  has been considered. If  $g_1$  corresponds to the food concept of  $O_{fs_2}$  aligned with  $f$  in  $GS_{73}$ , it is considered as a food first match in Table 4. If at least one of the 5 food concepts  $g_i \in O_{fs_2}, i \in [1, 5]$ , corresponds to the food concept of  $O_{fs_2}$  aligned with  $f$  in  $GS_{73}$ , it is considered as a food best five match in Table 4. Thresholds used to define the set of considered alignments for the five methods have been chosen independently for each method to optimize the F-measure in order to compare the best results which can be obtained for each method. The syntactic score (Def. 4) alone provides better results than the method presented in Algorithm 1 for food first match. Considering the method presented in Algorithm 1 and the syntactic score (Def. 4) alone, the best results in terms of F-measure are obtained for food first match with a threshold of 0.6 on the similarity score. It can be noticed that for both kinds of considered alignments (first match and five first matches), F-measure is always better for the method presented in Algorithm 1 and the syntactic score (Def.4) alone compared to OnAGUI and AML.

Additional assessments are presented in Table 5 concerning family matches with the method presented in Algorithm 1, the syntactic score alone (Def. 4) and the semantic score alone (Def. 7) for  $GS_{181}$ . For a given food concept  $f \in O_{fs_1}$  the decreasing ordered list of the 5 best matches  $g_i \in O_{fs_2}, i \in [1, 5]$  has been considered. If  $g_1$  belongs to the closest  $O_{fs_2}$  family associated with  $f$  in  $GS_{181}$ , it is considered as a family first match in Table 5. If at least one of the 5 food concepts  $g_i \in O_{fs_2}, i \in [1, 5]$ , belongs to the closest  $O_{fs_2}$  family associated with  $f$  in  $GS_{181}$ , it is considered as

Table 4. Food matches results with  $GS_{73}$

Alignment method	first				best five			
	Threshold	P	R	F-M	Threshold	P	R	F-M
Combination Syntactic score (Def. 4)- Semantic score (Def. 7) (Algo. 1)	0.6	0.58	0.52	0.55	0.9	0.77	0.27	<b>0.40</b>
Syntactic score (Def. 4)- Semantic score (Def. 7)	0.6	0.66	0.53	<b>0.59</b>	0.9	1.0	0.23	0.38
OnAGUI (Levenstein)	0.6	0.33	0.33	0.33	0.9	0.12	0.56	0.20
OnAGUI (Levenstein)	0.8	0.29	0.19	0.23	0.7	0.15	0.31	0.21
AML	0.6	0.89	0.23	0.37	0.9	0.94	0.22	0.35

a family best five match in Table 5. The best result in terms of F-measure is obtained for the family first match with the semantic score alone. The best result obtained for the family 5 best matches is obtained for the method presented in Algorithm 1 which shows that in this case the combination of the syntactic and the semantic scores allows to obtain a better result.

In order to evaluate the weight associated with each facet in Equation 3 from Definition 6, the assumption that facets A, B and C are the more important to characterize a given food has been made. The similarity score was computed using different combinations of weighting values for facets A, B and C on  $GS_{181}$ . Best results in terms of first match were obtained for  $w_{facet_A} = 7$ ,  $w_{facet_B} = 5$  and  $w_C = 1$ , weights associated with other facets were set to 1.

Results presented in Table 5 show that the semantic score proposed in the alignment method of Algorithm 1 provides better results than the syntactic score to classify food concepts  $f \in O_{fs_1}$  into relevant  $O_{fs_2}$  families of concepts. Alignments at the same level of granularity (food match) between  $f \in O_{fs_1}$  and  $g \in O_{fs_2}$  assessed in Table 4 show that the method presented in Algorithm 1 and the syntactic score (Def.4) alone compared to state of the art methods (OnAGUI and AML both based on a syntactic similarity) give better results. Alignments at the same level of granularity (food match) with the method presented in Algorithm 1 and the syntactic score (Def.4) are less relevant than those obtained with the same method or the semantic score alone for family matches which is not surprising as it is more constraining and precise. It can be explained by the fact that there is not a unique way to annotate a food product using LanguAL. Nevertheless, both kinds of alignment may be very helpful for manual validation done by annotators. Indeed, we have experimented during data enrichment task presented in Section 4.2 that it is worth analyzing food matches because when they are correct, less

Table 5. Family matches results with  $GS_{181}$

Alignment method	first				best five			
	Threshold	P	R	F-M	Threshold	P	R	F-M
Combination Syntactic score (Def. 4)- Semantic score (Def. 7) (Algo. 1)	0.6	0.63	0.55	0.59	0.6	0.8	0.62	<b>0.63</b>
	0.9	0.81	0.12	0.21	0.9	0.64	0.11	0.19
Syntactic score (Def. 4)- Semantic score (Def. 7)	0.6	0.75	0.24	0.36	0.6	0.34	0.24	0.28
Semantic score (Def. 7)	0.6	0.81	0.63	<b>0.71</b>	0.6	0.51	0.66	0.58

food concepts  $g \in O_{fs_2}$  have to be analyzed during manual validation, which greatly reduces time spent for this task.

#### 4. PRESENTATION AND ASSESSMENT OF MULTIDB EXPLORER

MultiDB explorer has been designed in the framework of the French national research project MeatyLab involving academic and industrial partners. The need expressed by industrial partners was to be able to take benefit of external nutritional data sources to help them in their recipe formulation process. The first step is to identify the nutritional values for each ingredient either by its identification in an appropriate Food Composition DataBase (FCDB), or by designing a specific experimental analysis procedure. As the second solution is time-consuming and costly, common practice is to reuse existing data available in national FCDBs. Unfortunately, values associated with 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 commonly used by nutritionists to deal with incompleteness. The objective of MultiDB explorer is to facilitate this task. Consequently, the following system specifications have been determined during the MeatyLab project with industrial partners:

1. providing homogeneous access to a collection of national FCDBs of interest;
2. querying the collection of national FCDBs using English or national language terms denoting a given food;
3. providing access, for a given food, to its nutritional composition in terms of nutrients and associated quantity;
4. providing, for a given food, the list of similar foods in other FCDBs.

Specification 1 is explained by the fact that each national FCDB uses both specific data format and querying interface. For example, Ciqual and USDA databases provide access to their data sources using specific web querying interfaces or the Microsoft Access format. With this latter solution, users need to understand database structures which are quite different. In case when, for a given food, nutrient data of interest are not available in the FCDB of the country in which the nutritional quality must be assessed, specification 4 allows one to facilitate the search of this lacking value for a similar food in another FCDB.

In Section 4.1, the main functionalities of MultiDB explorer are briefly presented. Industrial partners provided a use case to assess the relevance of MultiDB explorer to manage incompleteness of a given FCDB. The results obtained using MultiDB explorer to deal with iron, Vitamin B12, Vitamin C nutrients incompleteness in Ciqual reusing USDA data source are presented in Section 4.2.

##### 4.1 Architecture and Functionalities

The first three system specifications are already implemented in EuroFIR FoodEXplorer tool (EUROFIR (2020)). Unfortunately, this tool must be used as it is and cannot be extended to be able to implement specification 4. However, EuroFIR provides, under license, copies of its database which integrates in a homogeneous XML file format the content of a large set of national FCDBs. A relational database with a schema equivalent to the EuroFIR XML schema has been created and fulfilled with data provided by EuroFIR. In MultiDB explorer, 6 FCDBs are currently included (see Table 6).

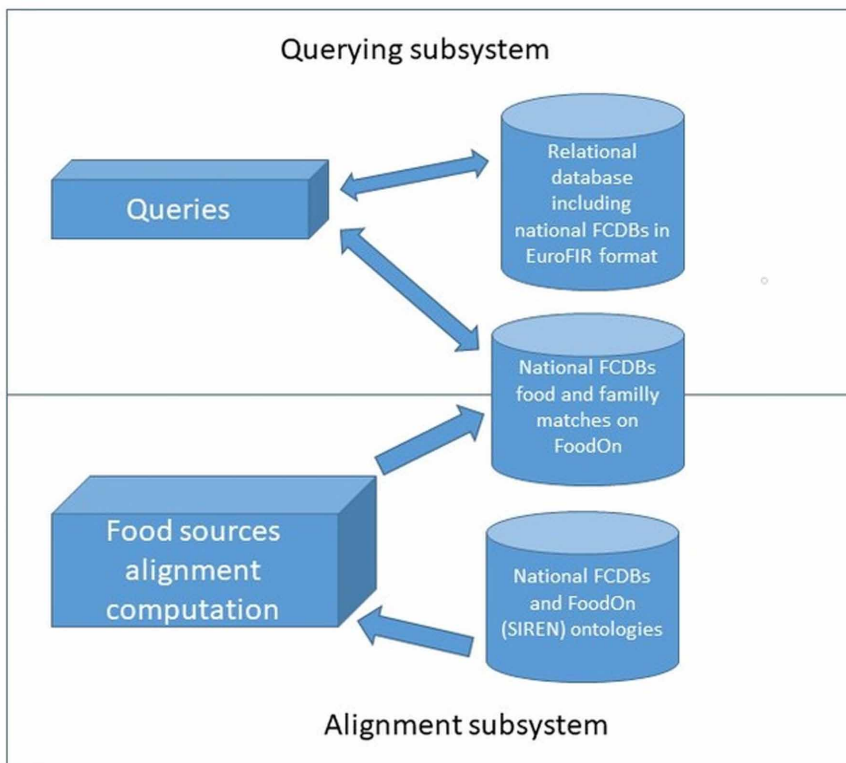
As presented in Figure 6, MultiDB explorer is composed of two subsystems. The first one is the querying subsystem which interacts with the relational database including EuroFIR data to retrieve nutritional values associated with searched foods and to provide access to the list of foods and family alignments associated with a given food. The second subsystem computes food sources alignments implementing the food sources alignment method presented in Section 3. The food sources alignment



Table 6. List of FCDBs available with current content using MultiDB explorer

FCDB name	# foods
Danish FCDB	1,049
Dutch FCDB	2,020
Swedish FCDB	2,056
USDA FCDB	8,618
French FCDB	2,807
British FCDB	2,897

Figure 6. MultiDB explorer architecture



engine must be run when a new version of EUROFIR XML file is integrated in the relational database to update food and family alignments. Implementation of specification 4 in MultiDB explorer, which is the original one compared to FoodEXplorer, is presented through illustrations in the following. Figure 7 presents MultiDB explorer search of the French term *Courgette, purée* which has been found in Ciqual FCDB. It may be noticed that the query may be done using the national language term or English term defined in the FCDB (here *Courgettes, puree*). A click on the result of the query, id 20264 in column Origfdb (see Figure 8), allows one to retrieve the LanguaL description associated with *Courgette, purée* and a list of USDA foods associated with the same FoodOn family. This list has been computed by MultiDB explorer thanks to the method presented in Section 4. MultiDB explorer

Figure 7. MultiDB explorer query for the Ciqual food Courgette, purée

The screenshot shows the MultiDB Explorer interface. At the top, it says "List of foods". Below that, it lists "Loaded databases" including Danish FCDB (1049 foods), NEVO-online 2013/4.0 (2020 foods), SE FCDB 2013-01-10 (2056 foods), USDA (8618 foods), IFIP (23 foods), French food composition table version 2017 (2807 foods), and CoF IDS 2015 (2897 foods). There are two search sections: "Name search" and "Langual search". The "Name search" section has a text input field containing "Courgette, purée" and a "SEARCH" button. The "Langual search" section has an "ADD" button, a text input field, and a "SEARCH" button. Below the search sections, it displays "1 results for 'Courgette, purée'" and "1 food displayed." Below this, there is a table with the following data:

Origidcd	Names	Origin database
20264	['Courgettes, puree@en', 'Courgette, purée@fr']	French food composition table version 2017

also provides the list of nutrient values associated with *Courgette, purée* in Ciqual: none of the 3 nutrients vitamin C, vitamin B12 and iron are known. Selecting the similar USDA food *Squash, winter, acorn, cooked, boiled, mashed, without salt* allows its list of nutrients values to be retrieved. Figure 9 shows an excerpt of this list which includes the value associated with vitamin C in USDA FCDB.

#### 4.2 MultiDB Explorer Assessment to Enrich CIQUAL Data Using USDA

A use case of the method designed to implement Ciqual FCDB enrichment task and associated results is presented in the following. This use case consists in finding in USDA food source values associated with nutrients vitamin C, vitamin B12 and iron when they are not known in Ciqual for a given food.

To achieve this task, the set of Ciqual food products defined in  $GS_{181}$  (see Section 3.3) has been reused. The subset of 99 foods called  $GS_{99}$  for which at least one of the values associated with the 3 nutrients is not known in Ciqual and at least one similar food can be found in USDA has been extracted from  $GS_{181}$ .

Then the alignment method between food sources presented in Section 4 has been used to align USDA on FoodOn. For each USDA food term, two alignments on FoodOn are provided: the best FoodOn food alignment and the best FoodOn family alignment. Consequently, as the same alignment has been done between Ciqual and FoodOn, it is possible to use FoodOn as pivot to determine the list of USDA foods which are similar to Ciqual foods at two levels of granularity: food level and family level.

Figure 8. MultiDB explorer excerpt of the answer obtained for the Ciqual food Courgette, purée (Langual description and USDA foods belonging to the same FoodOn family)

**Courgettes, puree@en**  
**Courgette, purée@fr**

Langual codes :

A0152	A0826	B1462	C0140	E0110	F0014	G0012	H0001	J0001
K0003	M0001	N0001	P0024	FULLY HEAT-TREATED [F0014]				

**FoodOn family :**  
[squash vegetable food product](#)

**Products in FoodOn family :**

- [Squash, Indian, raw \(Navajo\)@en \(USDA\)](#)
- [Squash, winter, spaghetti, cooked, boiled, drained, or baked, without salt@en \(USDA\)](#)
- [Squash, summer, all varieties, raw@en \(USDA\)](#)
- [Squash, winter, hubbard, raw@en \(USDA\)](#)
- [Squash, winter, acorn, cooked, baked, with salt@en \(USDA\)](#)
- [Squash, winter, acorn, cooked, boiled, mashed, without salt@en \(USDA\)](#)
- [Squash, summer, scallop, raw@en \(USDA\)](#)
- [Balsam-pear \(bitter gourd\), pods, cooked, boiled, drained, with salt@en \(USDA\)](#)
- [Squash, Indian, cooked, boiled \(Navajo\)@en \(USDA\)](#)
- [Squash, summer, all varieties, cooked, boiled, drained, without salt@en \(USDA\)](#)
- [Squash, winter, acorn, cooked, boiled, mashed, with salt@en \(USDA\)](#)
- [Squash, winter, butternut, raw@en \(USDA\)](#)
- [Squash, summer, scallop, cooked, boiled, drained, with salt@en \(USDA\)](#)
- [Squash, summer, scallop, cooked, boiled, drained, without salt@en \(USDA\)](#)
- [Squash, summer, crookneck and straightneck, canned, drained, solid, without salt@en \(USDA\)](#)
- [Squash, summer, crookneck and straightneck, raw@en \(USDA\)](#)
- [CAMPBELL'S Homestyle Butternut Squash Bisque@en \(USDA\)](#)
- [Squash, winter, spaghetti, cooked, boiled, drained, or baked, with salt@en \(USDA\)](#)
- [Squash, winter, all varieties, raw@en \(USDA\)](#)
- [Squash, winter, butternut, cooked, baked, with salt@en \(USDA\)](#)

[Get JSON data](#)  
[Go back to products list](#)

Automatic alignments of Ciqual foods on USDA foods have been manually assessed by two domain experts: 76 alignments have been considered relevant, which corresponds to 76% of  $GS_{99}$ . For those 76 relevant alignments, values associated with the 3 nutrients of interest have been retrieved using MultiDB explorer. Detailed results are presented in Table 7. For those 76 relevant alignments, 91% of unknown values in Ciqual have been enriched with values from USDA and 96% of known values in Ciqual have been completed by values from USDA.

$GS_{99}$  is available in the dataset <https://doi.org/10.15454/4XIBS9> stored in Excel format in the institutional INRAE dataverse. The same dataset also contains the list of automatic alignments of Ciqual foods on FoodOn foods, the list of automatic alignments of USDA foods on FoodOn foods, the list of alignments of Ciqual foods on USDA foods.

Figure 9. MultiDB explorer excerpt of the answer obtained for the USDA food Squash, winter, acorn, cooked, boiled, mashed, without salt (excerpt of the nutrients list including vitamin C value)

riboflavin	0.008000	milligram	per 100g edible portion	best estimate	imputed/estimated, generic	Analytical methods	File or Database
thiamin	0.100000	milligram	per 100g edible portion	best estimate	imputed/estimated, generic	Analytical methods	File or Database
vitamin C (ascorbic acid)	6.500000	milligram	per 100g edible portion	best estimate	imputed/estimated, generic	Analytical methods	File or Database
lycopene	0.000000	microgram	per 100g edible portion	best estimate	Analytical result(s)	Analytical methods	Article in Journal
beta-cryptoxanthin	0.000000	microgram	per 100g edible portion	best estimate	Analytical result(s)	Analytical methods	Article in Journal

## 5. COMPARISON WITH THE STATE OF THE ART

### 5.1 Why Choosing FoodOn?

As pointed out in Ireland and Moller (2016), there is general agreement on the importance of using a multifaceted approach for identifying foods in food databases. Moreover, the use of a common food classification and description system could provide a central linking system between existing systems to promote more accurate dietary exposure calculations in Europe and also at a worldwide level.

In agriculture, the Global Agricultural Concept Scheme (GACS) (2020) is a hub for concepts related to agriculture integrating three well known thesauri: AGROVOC Concept Scheme proposed by the Food and Agricultural Organization of the United Nations (FAO) (AGROVOC, 2020), the CAB Thesaurus (CABT, 2020), and NAL Thesaurus (NALT, 2020) of the National Agricultural Library of

Table 7. Results obtained for 76 Ciqual food concepts using GS<sub>99</sub>

	vitamin C	vitamin B12	iron
# missing values in Ciqual	37	64	27
# missing values completed with USDA	35	55	26
# known values in Ciqual	39	12	49
# known values completed with USDA	37	12	47

the USA (NALT, 2020). The three organizations selected the 10,000 concepts most frequently used in each respective thesaurus. These concepts were automatically mapped, mappings were checked by hand, and inconsistencies were resolved by discussion (Baker et al., 2016). However, the initial thesauri that were designed for general purposes in Agriculture are not precise enough for the task of food description and are not suitable for representing specific foods appearing in nutrient databases (Ireland & Moller, 2016).

In food domain, a long-term effort has been done by the International Network of Food Data Systems (INFOODS) including a template in Excel format to compile food nutritional composition using a standardized vocabulary for chemical components (Charrondiere & Burlingame, 2011; Murphy et al., 2016). Unfortunately, until now, each national agency, e.g. USDA or ANSES, indexes its own composition tables using its specific food terminology. The recent initiative of the FoodOn consortium supports the development of the FoodOn ontology (Dooley et al. 2018), aiming at proposing a standard food description vocabulary composed of term hierarchy facets. Reusing terms from LanguaL—the thesaurus used to index numerous agency databases and presented as a good potential choice for a central mapping structure—the FoodOn project provides a familiar and well-known hierarchy of concepts and associated terminology for professionals and researchers. This is the main reason why FoodOn was chosen as a pivot ontology in the proposed approach to map specific food terminologies used by national agencies.

## 5.2 Why Designing a New Food Sources Alignment Method?

From the ontology alignment perspective, there is a clear need for specific reference knowledge in specific domains (Shvaiko & Euzenat, 2013; vanHage et al., 2010). Indeed, commonly used external knowledge sources, such as WordNet, fail to provide the semantic information that is needed to correctly discover the correspondences between domain specific concepts. More precisely, it has been shown in many publications (Aleksovski et al., 2006; Faria et al., 2013; Tigrine et al., 2015; Tigrine et al., 2016; Annane et al., 2018) that the ontology matching process can take benefit from the use of Background Knowledge. Our alignment method of a given food vocabulary with the FoodOn ontology can be viewed as an ontology matching problem using a background knowledge, encoded here in the LanguaL thesaurus itself included in FoodOn. A contribution of this paper is to propose (i) a way to represent this background knowledge in OWL to facilitate its reuse, (ii) a new semantic measure based on this background knowledge for food comparison.

Ispirova et al. (2017) proposes an alignment method dedicated to nutritional data food sources. They study different string matching comparison methods and conclude that POS tagging combined with probability theory provides best results. Unfortunately, they tested their method only with food nutrient English names which are less complicated to compare as they only include combinations of standardized chemical names, adjectives and numbers. Food product names are more complicated to compare as they may include different food product names specific to countries and verbs describing applied treatments.

One of the tasks of the Ontology Alignment Evaluation Initiative (OAEI) in 2006 and 2007 campaigns (Euzenat06 et al., 2006; vanHage et al., 2010) was named “food”. In fact, the data sets were containing AGROVOC and NALT, two thesauri for agriculture. Specific challenges identified for the food task were: i) large data sets alignment; ii) concepts alignment from many different domains not restricted to food; iii) alignment based on the weak semantic of the thesaurus structure requiring term disambiguation and good lexical matching strategies. When comparing this OAEI food task with the context of the proposed alignment method of a given food vocabulary with the FoodOn ontology using LanguaL description, one can notice the following differences: the data sets are smaller, concepts are all from food domain and defined using the multifaceted thesaurus LanguaL which requires a solution to consider this information which was not proposed in the OAEI food task. The aims of food sources alignment task addressed in this paper are different compared to the one proposed in the OAEI food task.

### 5.3 Why Designing a New Tool?

MultiDB explorer is a system tailored for querying food vocabularies and associated nutritional values through a pivot vocabulary. As pointed out in Section 4.1, MultiDB explorer implements some specification of the EuroFIR FoodEXplorer tool (EUROFIR, 2020) and extend it by providing, for a given food, the list of similar foods in other FCBDs using food sources alignments. Unfortunately, it was not possible to directly extend EuroFIR FoodEXplorer tool which must be used as it is.

Azzi et al. (2016) describes a methodology to automatize the assessment of a nutritional score to a recipe and its implementation. The core of this methodology is based on mapping between text corpora (cooking recipes) and structured data, extracted from Nutrinet, a food composition table associated with French study Nutrinet (<https://etude-nutrinet-sante.fr/>). A termino-ontological resource, not publicly available, is used in order to enhance the quality of the mapping and therefore allow a better nutritional qualification of the recipe. The approach proposed in MultiDB explorer brings a complementary functionality which allows one to retrieve in other FCDBs nutrients value required to compute the nutritional score of a recipe when this value is unknown in FCDB used (here Nutrinet).

Eftimov et al. (2017) and Popovski et al. (2020) propose a method to classify food product names in English in four main FoodEx2 categories (raw, derivative, simple composite food, and aggregated composite food). Moreover, the method proposes to determine a list of FoodEx2 codes which could be associated with food product names. This method is complementary to the one proposed in MultiDB explorer as it could be potentially adapted to LanguaL codes in order to help annotators to describe food product names using LanguaL.

## 6. CONCLUSION

The main objective of this paper is to deal with different heterogeneous food sources in order to be able to compute the nutritional value of a given recipe when some data are lacking. For that, a new food sources alignment method using LanguaL description resources as background knowledge on FoodOn considered as a pivot ontology has been proposed. To reach this objective, the contributions of this paper are: first to define how to transform into an ontology a food source with its associated LanguaL descriptions, second to compute similarity scores which were combined in a suitable way to take into account the specificities of LanguaL descriptions as a background knowledge, third to create an original gold standard to assess the proposed alignment method, fourth to propose a new method based on FoodOn as pivot ontology to semi-automatically link similar foods belonging to different FCDBs, and fifth to assess this method on a real use case provided by industrial partners.

To the best of our knowledge, it is an original contribution compared to the state of the art (see Section 5 for more details). Few works have addressed the problem of food products comparison for alignment purpose. Food products are complicated to compare as they may include different food product names specific to countries and associated cultures and verbs describing applied treatments. We propose to tackle this difficulty by (i) selecting a syntactic similarity method which provides better results than 2 state of the art methods tested in this paper, (ii) using the food description expressed in a standardized vocabulary, namely its LanguaL description to define a semantic similarity function. Assessments presented in Section 3 show that taking into account the LanguaL semantic description associated with the food product significantly enhances the results obtained with state of the art syntactic methods for family matches.

This paper proposes to perform this task using the OWL ontologies knowledge representation model. This choice has several advantages: (1) the use of international standards established by the W3C consortium (<https://www.w3.org/>); (2) the availability of numerous tools based on those standards (OWL ontology portals, editors, validators and alignment tools, triple stores to manage RDF databases annotated by OWL ontologies) to facilitate the reuse of ontologies in an Open Science perspective; (3) the accuracy of OWL language operators to describe knowledge compared to interchange data languages as JSON classically used.

A short-term initiative will be to disseminate those results to national agencies to incite them to validate the automatic alignments provided by the proposed method in order to facilitate FCDBs interoperability. Another perspectives are (1) to study in which extend the method could be extended to take benefit from manual alignment validations in an iterative learning process; (2) to propose for a next OAEI ontology alignment challenge organized at the international level (<http://oaei.ontologymatching.org/>) a new task dedicated to food products alignment based on the ontological resources which are also a contribution of this paper from the ontology community point of view.

## **ACKNOWLEDGMENT**

This work was supported by the FUI Metyl@b Project financed by BPI France.

## REFERENCES

- AGROVOC. (2020). *Concept Scheme proposed by the Food and Agricultural Organization of the United Nations (FAO)*. <http://aims.fao.org/fr/agrovoc>
- Aleksovski, Z., Klein, M., ten Kate, W., & van Harmelen, F. (2006). Matching unstructured vocabularies using a background ontology. In *Managing Knowledge in a World of Networks* (pp. 182-197). Springer Berlin Heidelberg. doi:10.1007/11891451\_18
- Annane, A., Bellahsene, Z., Azouaou, F., & Jonquet, C. (2018). Building an effective and efficient background knowledge resource to enhance ontology matching. *Journal of Web Semantics*, 51, 51–68. doi:10.1016/j.websem.2018.04.001
- Azzi, R., Despres, S., Guezennec, G., & Nobécourt, J. (2016). Utilisation de ressources sémantiques pour l'automatisation du calcul d'un score nutritionnel. CNIA 2016 – Conférence Nationale d'Intelligence Artificielle, 17-24.
- Baker, T., Caracciolo, C., Doroszenko, A., & Suominen, O. (2016). GACS core: Creation of a global agricultural concept scheme. *Metadata and Semantics Research - 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*, 311-316.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (1st ed.). O'Reilly Media, Inc.
- CABT. (2020). *CAB Thesaurus*. <https://www.cabi.org/publishing-products/cab-thesaurus/>
- Charrondiere, U. R., & Burlingame, B. (2011). Report on the FAO/IN-FOODS Compilation Tool: A simple system to manage food composition data. *Journal of Food Composition and Analysis*, 24(4), 711-715.
- Ciqual. (2020). *French food composition table*. <https://ciqual.anses.fr/>
- Dooley, D. M., Griffiths, E. J., Gosal, G., Buttigieg, P. L., Hoehndorf, R., Lange, M., Schriml, L. M., Brinkman, F. S. L., & Hsiao, W.W. L. (2018). Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Science of Food*, 2, 23.
- Eftimov, T., Korosec, P., & Korusic, B. (2017). Standfood: Standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. Nutrients.
- English Pen Treebank Tagset. (2020) <https://cs.nyu.edu/grishman/jet/guide/PennPOS.html>
- EUROFIR. (2020). *European Food Information Resource*. <https://www.eurofir.org/>
- Euzenat, J., Mochol, M., Shvaiko, P., Stuckenschmidt, H., Svab, O., Svatek, V., van Hage, W. R., & Yatskevich, M. (2006). Results of the ontology alignment evaluation initiative 2006. In *Proceedings of the 1st International Workshop on Ontology Matching (OM-2006) Collocated with the 5th International Semantic Web Conference (ISWC-2006), volume 225 of CEUR Workshop Proceedings*. CEUR-WS.org.
- Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I. F., & Couto, F. M. (2013). The AgreementMakerLight ontology matching system. *On the Move to Meaningful Internet Systems: OTM 2013 Conferences - Confederated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE 2013*, 527-541. doi:10.1007/978-3-642-41030-7\_38
- GACS. (2020). *Global Agricultural Concept Scheme*. <http://agroportal.lirmm.fr/ontologies/GACS>
- INFOODS FAO. (2020). *International Network of Food Data Systems*. <http://www.fao.org/infoods/infoods/fr/>
- Ireland, J. D., & Moller, A. (2016). Food classification and description. In *Encyclopedia of Food and Health* (pp. 1–6). Academic Press.
- Ireland, J. D., & Moller, A. (2000). Review of international food classification and description. *Journal of Food Composition and Analysis*, 13(4), 529–538. doi:10.1016/B978-0-12-384947-2.00307-X
- Ireland, J. D., & Moller, A. (2010). LanguaL food description: A learning process. *European Journal of Clinical Nutrition*, 64(S3), 44–48. doi:10.1038/ejcn.2010.209 PMID:21045849



Ispirova, G., Eftimov, T., Seljak, B., & Korošec, P. (2017). Mapping food composition data from various data sources to a domain-specific ontology. *KEOD 2017*.

LanguaL. (2020). *The International Framework for Food Description*. <https://www.langual.org/>

LanguaL Indexed Datasets. (2020). *The LanguaL indexed Datasets*. [http://langual.org/langual\\_indexed\\_datasets.asp](http://langual.org/langual_indexed_datasets.asp)

MultiDB Explorer. (2020). *MeatyL@b Data and Ontology Explorer*. <https://ico.iate.inra.fr/meatylab/>

Murphy, S. P., Charrondiere, U. R., & Burlingame, B. (2016). Thirty years of progress in harmonizing and compiling food data as a result of the establishment of Infoods. *Food Chemistry*, 193, 2–5. doi:10.1016/j.foodchem.2014.11.097 PMID:26433279

NALT. (2020). *National Agricultural Library Thesaurus of the USA*. <http://agroportal.lirmm.fr/ontologies/NALT>

OWL Language. (n.d.). <https://www.w3.org/TR/owl-features/>

Pehrsson, P., & Haytowitz, D. (2016). Food composition databases. In B. Caballero, P. M. Finglas, & F. Toldra (Eds.), *Encyclopedia of Food and Health* (pp. 16–21). Academic Press. doi:10.1016/B978-0-12-384947-2.00308-1

Popovski, G., Ispirova, G., Hadzi-Kotarova, N., Valencic, E., Eftimov, T., & Korošec-Seljak, B. (2020). Food data integration by using heuristics based on lexical and semantic similarities. In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020)* (vol. 5, pp. 208-216). SCITEPRESS. doi:10.5220/0008990602080216

Shvaiko, P., & Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 158–176. doi:10.1109/TKDE.2011.253

Tigrine, A. N., Bellahsene, Z., & Todorov, K. (2015). Light-weight cross-lingual ontology matching with LYAM++. *On the Move to Meaningful Internet Systems: OTM 2015 Conferences - Confederated International Conferences: CoopIS, ODBASE, and C&TC 2015*, 527-544.

Tigrine, A. N., Bellahsene, Z., & Todorov, K. (2016). Selecting optimal background knowledge sources for the ontology matching task. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge Engineering and Knowledge Management* (pp. 651–665). Springer International Publishing. doi:10.1007/978-3-319-49004-5\_42

USDA. (2020). *National Nutrient Database for Standard Reference*. <https://ndb.nal.usda.gov/ndb/>

USDA SR Legacy. (2020). *USDA National Nutrient Database for Standard Reference, Legacy Release*. <https://data.nal.usda.gov/dataset/usda-national-nutrient-database-standard-reference-legacy-release>

van Hage, W., Sini, M., Finch, L., Kolb, H., & Schreiber, G. (2010). The OAEI food task: An analysis of a thesaurus alignment task. *Applied Ontology*, 5(1), 1–28. doi:10.3233/AO-2010-0072

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, ACL '94* (pp. 133-138). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/981732.981751

## APPENDIX. EXCERPT OF THE OWL DEFINITION

An excerpt of the OWL definition of the concept which corresponds to the food *Cooked pork shoulder, choice* presented in Table 2 is given below:

```
<owl:Class rdf:ID="FoodConcept"/>
<!-- an excerpt of the specialisation hierarchy -->
<!-- associated with facet A -->
<owl:Class rdf:ID="A_PRODUCT_TYPE">
  <rdfs:subClassOf rdf:resource="#FoodConcept"/>
</owl:Class>
<owl:Class rdf:ID="CURED_MEAT_US_CFR">
  <rdfs:subClassOf rdf:resource="#A_PRODUCT_TYPE"/>
</owl:Class>
<!-- an excerpt of the specialisation hierarchy -->
<!-- associated with facet B -->
<owl:Class rdf:ID="B_FOOD_SOURCE">
  <rdfs:subClassOf rdf:resource="#FoodConcept"/>
</owl:Class>
<owl:Class rdf:ID="SWINE">
  <rdfs:subClassOf rdf:resource="#B_FOOD_SOURCE"/>
</owl:Class>
<rdf:ObjectProperty rdf:ID="has_Facet">
  <rdfs:domain rdf:resource="#FoodConcept"/>
  <rdfs:range rdf:resource="#FoodConcept"/>
</rdf:ObjectProperty>

<owl:Class rdf:ID="Cooked_pork_shoulders_choice"/>
  <rdfs:label xml:lang="en">Cooked pork shoulder, choice</rdfs:label>
  <rdfs:subClassOf rdf:resource="#FoodConcept"/>
<!-- Description of facet A -->
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#has_Facet"/>
      <owl:someValuesFrom rdf:resource="#CURED_MEAT_US_CFR"/>
    </owl:Restriction>
  </rdfs:subClassOf>
<!-- Description of facet B -->
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#has_Facet"/>
      <owl:someValuesFrom rdf:resource="#SWINE"/>
    </owl:Restriction>
  </rdfs:subClassOf>
```

Let us notice that we can identify, for a given food concept, the LanguaL descriptors of a given facet type querying *in SPARQL* simultaneously the specialization hierarchy associated with this facet type (by example facet B) and the SomeValuesFrom restriction associated with the food concept.

Let us notice that we can identify the descriptors of a given facet by the specialization hierarchy to which belongs the concept associated with the SomeValuesFrom restriction.

Let us also notice that, to facilitate understanding of the Appendix, actual URI associated with Languag descriptors in FoodOn have been replaced by an URI based on its english label. By example, the actual URI associated with #CURED\_MEAT\_US\_CFR is [http://purl.obolibrary.org/obo/FOODON\\_03400279](http://purl.obolibrary.org/obo/FOODON_03400279).