A new alignment method based on FoodOn as pivot ontology

Patrice Buche, Julien Cufi, Liliana Ibanescu,
Alrick Oudot, Magalie weber

12/10/2021

Objectives

- Use cases
  - Replace a product in a recipe by a similar product

    → Request many datasources
  - Sometimes nutritional information (like iron, vitamin B12..) are not present in a given datasource

    → Retrieve missing information from another datasource
- How ?
  - Use FoodOn as pivot to integrate various food product vocabularies

  → Determine for each product to integrate the closest « family » product in FoodOn

FoodOn

- FoodOn :
  - Ontology about food-processing
    - ~6300 food products
    - ~1200 food products families

Each FoodOn product is annotated with ...

- LangualL :
  - Descriptive food indexing system
  - Each food product is described with controlled terms grouped in facets.
  - Ex : pork (raw)@en
    - A0150 : Meat or meat product
    - B1136 : Swine
    - F0003 : Not heat treated

```
swine meat food product@en
├── pork liver (raw)@en
├── pork (raw)@en A0150 B1136 F0003
├── chitterling (raw)@en
└── swine cured meat food product@en
    ├── ham (cured)@en
    ├── ham (smoked)@en
    ├── pork (uncooked, cured)@en
    ├── country ham@en
    ├── pork shoulder (cooked, cured)@en
    ├── ham (cooked, cured)@en
    ├── pork cut (cured)@en
    ├── pork loin (cooked, cured)@en
```

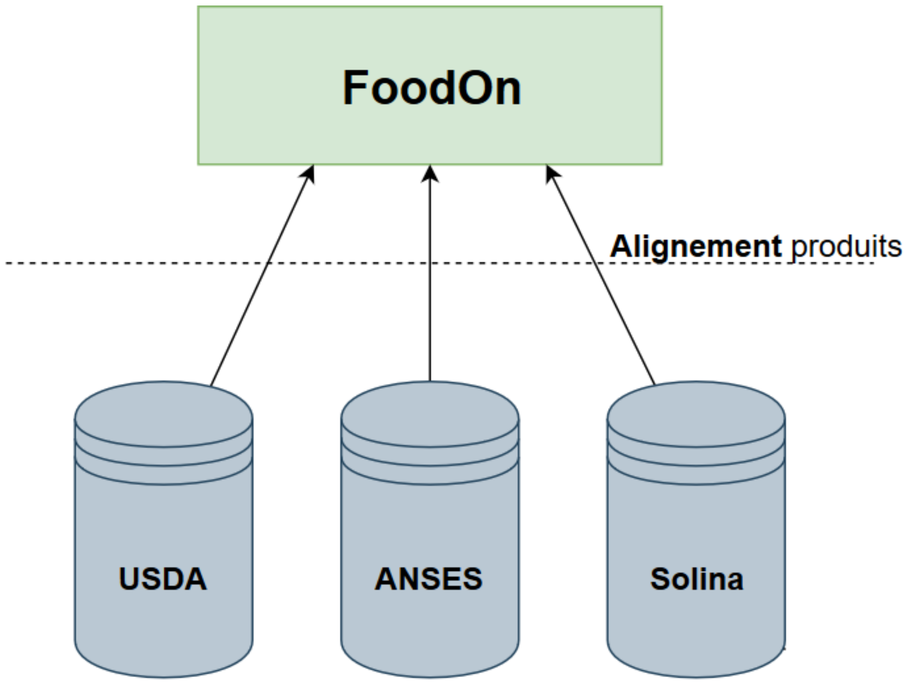
Nutritional information request

- Alignment

  Find for a ANSES or USDA product name the best family in the FoodOn hierarchy

- Amount of data
  - USDA :
    - ~ 8600 product names and associated data
  - ANSES :
    - ~ 3000 product names and associated data

  …

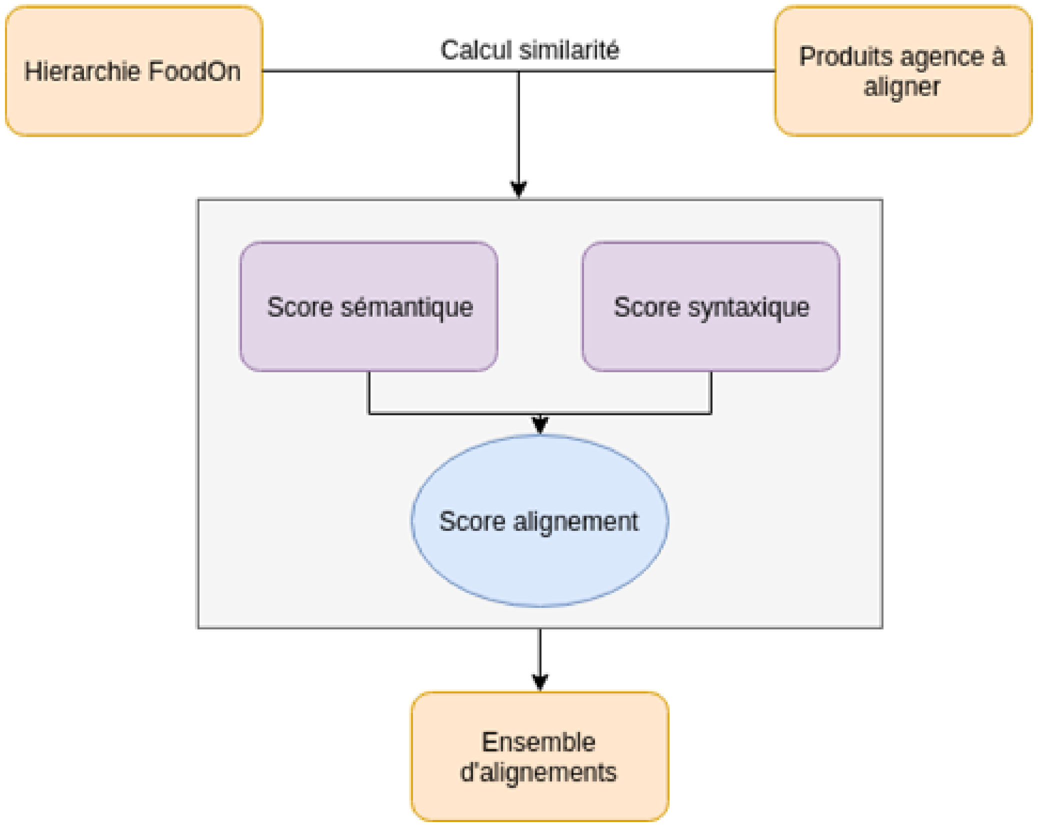


Alignment method

- Find the closest family in **FoodOn** for each product to add
- Problems

➔ We need an approach combining syntax (product name) and semantic (based on LangualL facets)

- Alignment method
- Compute **similarity score** between a product from USDA/ANSES and a product coming from *FoodOn*
  - **Semantic score** (LanguaL facets)
  - **Syntactic score**
  - Aggregation of both



Approach

- How to integrate a product in FoodOn ?
  - We compute a similarity measure between a food product p and a FoodOn food product p’
  - The proximity measure is the weighted sum of the similarity between each facets of both products
    - if two facets are the same the similarity is 1
    - if two facets are different the similarity depends on the length of the shortest path between these facets in the LanguaL hierarchy

Approach

- Example :
  - p has facets : A0001, B0011
  - p1 has facets : A0001, B0012
  - p2 has facets : A0001, B0021

```
Langual facets hierarchy
├── A
│   └── A0001
└── B
    ├── B0010
    │   ├── B0011
    │   └── B0012
    └── B0020
        └── B0021
```

SIM(p, p1) > SIM(p, p2)

→ Facet B0011 is closer to B0012 than B0021 in Langual Facet hierarchy

→ p1 is closer to p than p2

Example on Ciqual

- Ciqual product :
  - « Cooked pork shoulder »
- Closest products in FoodOn :
  - **pork shoulder (cooked, cured)@en**
    - pork picnic (cooked, cured)@en
    - pork butt (cooked, cured)@en
- Closest family in FoodOn :
  - « swine cured meat food product@en »
    - Family of the closest product

```
swine cured meat food product@en
├── ham (cured)@en 7.85
├── ham (smoked)@en 7.9
├── pork (uncooked, cured)@en 7.0
├── country ham@en 7.85
├── pork shoulder (cooked, cured)@en 9.8
├── ham (cooked, cured)@en 8.0
├── pork cut (cured)@en 7.0
├── pork loin (cooked, cured)@en 8.0
├── pork product (cured)@en 6.0
├── pork butt (cooked, cured)@en 9.05
├── pork picnic (cooked, cured)@en 9.8
├── pork ham (uncooked, cured)@en 7.0
├── pork loin (uncooked, cured)@en 7.0
├── pork shoulder (uncooked, cured)@en 8.8
├── pork butt (uncooked, cured)@en 8.8
├── pork picnic (uncooked, cured)@en 8.8
└── bacon (whole cut or parts)@en
    ├── bacon (raw)@en 8.0
    ├── bacon (canned)@en 9.0
    ├── bacon (smoked)@en 8.0
    ├── bacon (baked)@en 9.0
    ├── bacon (pump-cured)@en 8.0
    ├── bacon side@en 8.0
    ├── bacon (made with dry curing material)@en 8.0
    └── bacon (immersion cured)@en 8.0
```

Method assessment

1. Define a sample based on Ciqual database (called Gold Standard)

2. Define a « Gold standard » for the given sample
   - Experts choose for each product the closest product and family product in FoodOn

3. Perform algorithm on the sample

4. Compare algorithm results with « Gold standard »

Method assessment

- Gold standard :
  - 181 Ciqual products fully described in Langual
  - Only 73 aligned with FoodOn products (14 modified thanks to algorithm suggestion)
  - All categorized in a FoodOn familly

- Results
  - Exact match : Same response from the algorithm and the expert
  - Near match : Expert's response is in the 5 best results suggested by the algorithm
  - Not found : Not present in the 5 best results

Alignment results

	# food matches	
similarity scores	exact	near
Syntactic score (Def. 4)	41	46
Semantic score (Def. 7)	25	49
Combination (Algo. 1)	38	50

Table 4. Food matches results with GS including 73 Ciqual food concepts

	# family matches	
similarity scores	exact	near
Syntactic score (Def. 4)	110	122
Semantic score (Def. 7)	124	131
Combination (Algo. 1)	125	135

Table 5. Family matches results with GS including 181 Ciqual food concepts

Table 4 and Table 5 show that the **best results with GS** are obtained using the **combination of syntactic and semantic similarity scores**. Best results are obtained for family near match (76 %), food near match being 68%.

INRA SCIENCE & IMPACT

Back to use case

- Replace a product in a recipe by a similar product
  - → Request many datasources
- **Sometimes nutritionnal information (like iron, vitamin B12..) are not present in a given datasource**
  - **→ Retrieve missing information from another datasource**

Nutritional data source incompleteness management

Idea: when a nutrient value is not available in the Food Composition Data Base (FCDB) of interest, search it in other FCBDs for similar foods

Method: Ciqual food concepts alignment on USDA food concepts using FoodOn as pivot ontology

Use case: finding in USDA values associated with nutrients vitamin C, vitamin B12 and iron when they are not known in Ciqual for a given food.

Assessed on GS-: 99 Ciqual terms from GS for which at least one of the values associated with the 3 nutrients is not known in Ciqual and at least one similar term can be found in USDA (should be better with supplementary FCDBs).

Incompleteness management task results

	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

Table 7. Results with GS- including 76 Ciqual food concepts

- **76 alignments have been considered relevant** (on 99 considered)
- For those 76 relevant alignments, values associated with the 3 nutrients of interest have been retrieved using Meatylab explorer. Detailed results are presented in Table 7: **91% of unknown values in Ciqual have been enriched by values from USDA** and **96% of known values in Ciqual have been completed by values from USDA**

Conclusion-perspectives

- Paper accepted in IJAEIS
- Possible valorizations:
  - CALIS infrastructure (Consommateur ALIment Santé)
- Reusing data for prediction of O_2/CO_2 solubility in food (postdoc 2021, in progress)
- Alignment method should be enhanced:
  - Learning from alignment corrections done by annotators and from GS
  - Reusing aligment method with FoodEx2 instead of Langual

Questions ?

- Buche, P., Cufi, J., Dervaux, S., Dibie, J., Ibanescu, L., Oudot, A., & Weber, M. (2021). How to Manage Incompleteness of Nutritional Food Sources?: A Solution Using FoodOnas Pivot Ontology. International Journal of Agricultural and Environmental Information Systems (IJAEIS), 12(4), 1-26. http://doi.org/10.4018/IJAEIS.20211001.oa4

Merci !

Annexes

Approach

- How to integrate a product in FoodOn ?
  - We compute a similarity measure between a food product p and a FoodOn food product p’
  - The proximity measure is the weighted sum of the similarity between each facets of both products
    - if two facets are the same the similarity is 1
    - if two facets are different the similarity depends on the length of the shortest path between these facets in the LanguaL hierarchy

Zoom sur score sémantique

- **Score** (ou **similarité**) basé sur les facettes (descriptions) **LanguaL**
- La **similarité sémantique** entre 2 **produits** est la **moyenne pondérée** des mesures de similarité entre chacune des facettes de ces produits :

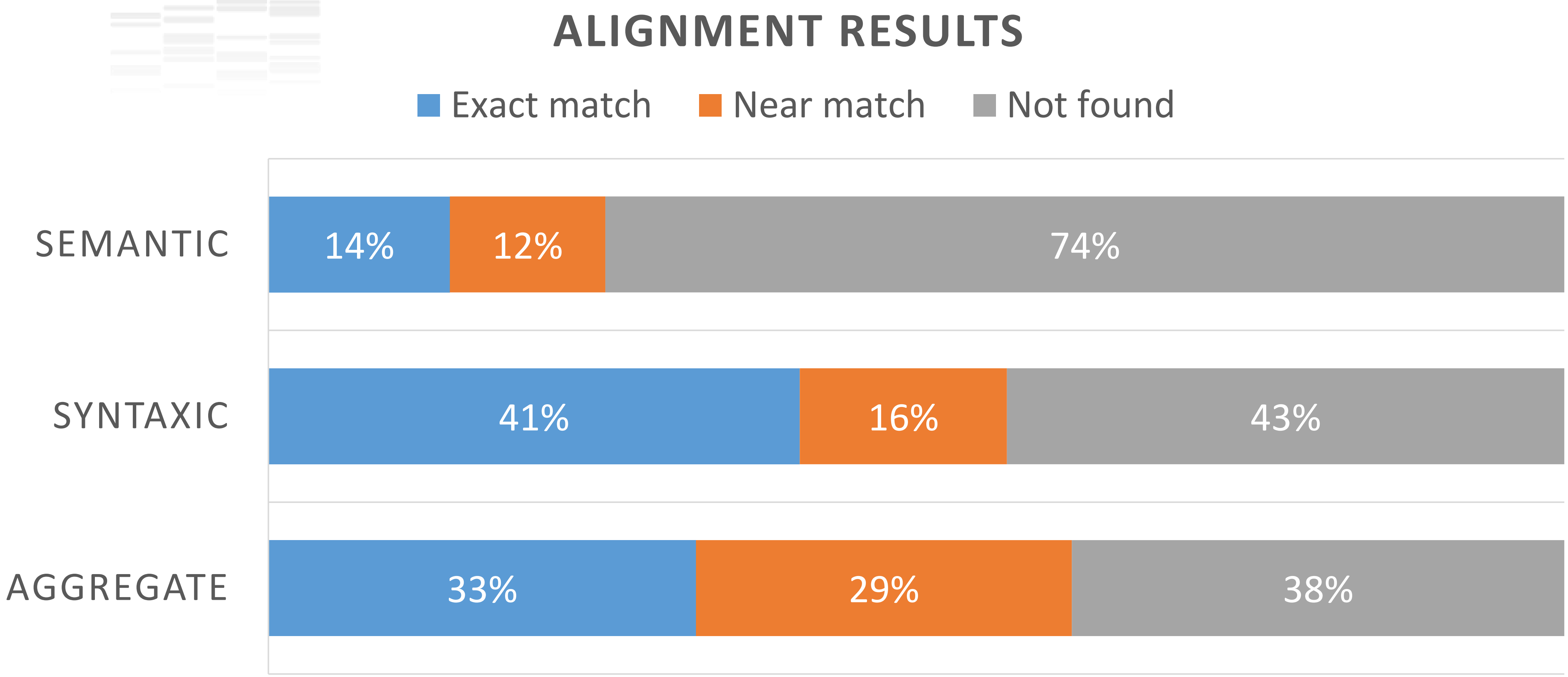
$$semanticSimilarity(P_1, P_2) = \frac{\sum_{(f_i, f_j) \in compFacets(P_1, P_2)} \omega(f_i) * sim(f_i, f_j)}{\sum_{(f_i, f_j) \in compFacets(P_1, P_2)} \omega(f_i)}$$

- Pondération calculées pour correspondre à l'importance relatives des différentes familles de facettes

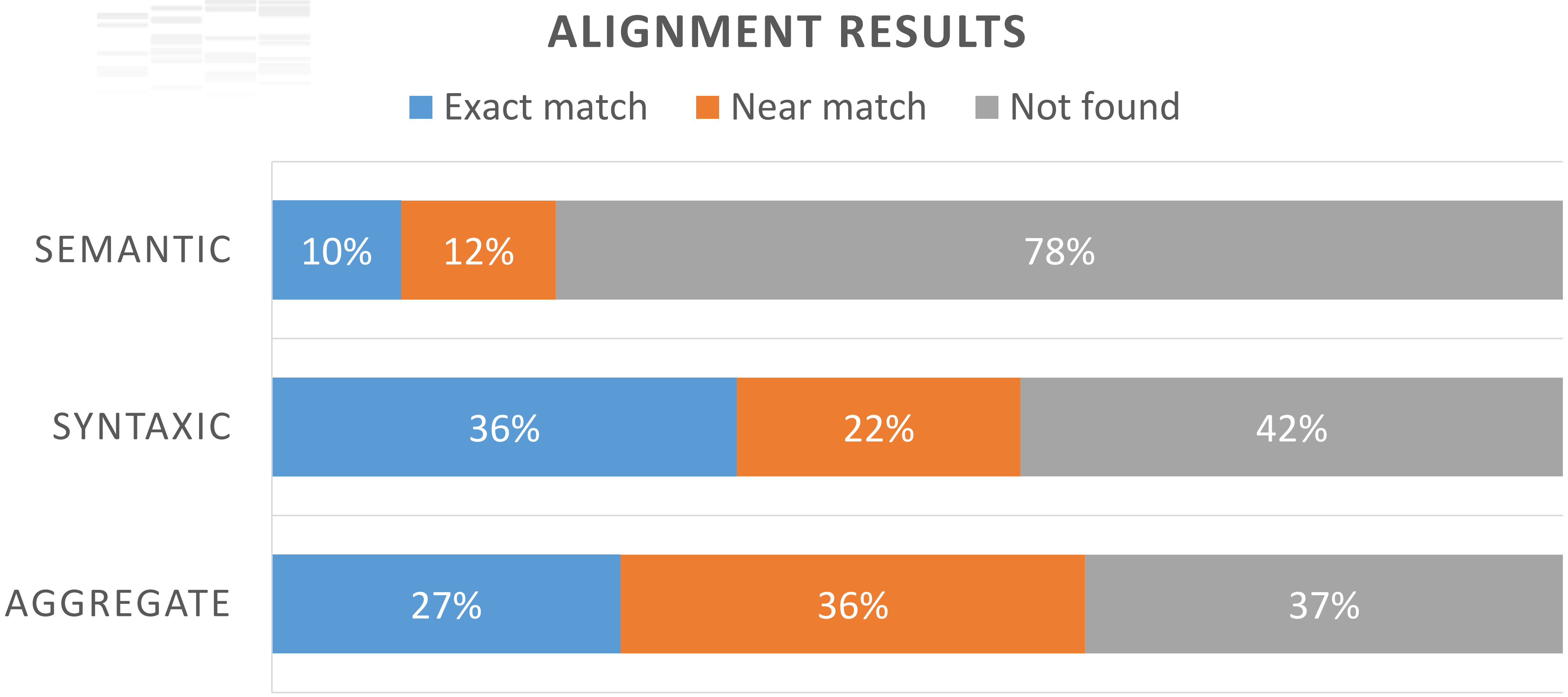
- La **similarité sémantique** entre 2 **facettes**
  - Si deux facettes sont identiques, la similarité est **1**
  - Si deux facettes n'appartiennent pas à la même famille, la similarité est **0**
  - Si deux facettes sont différentes (mais dans la même branche), on calcule leur similarité selon **Wu-Palmer** :

$$Wup(f_1, f_2) = 2 * \frac{depth(lcs(f_1, f_2))}{depth(f_1) + depth(f_2)}$$

Alignement results : Products – 300 products



Alignement results : Products – 100 meat products

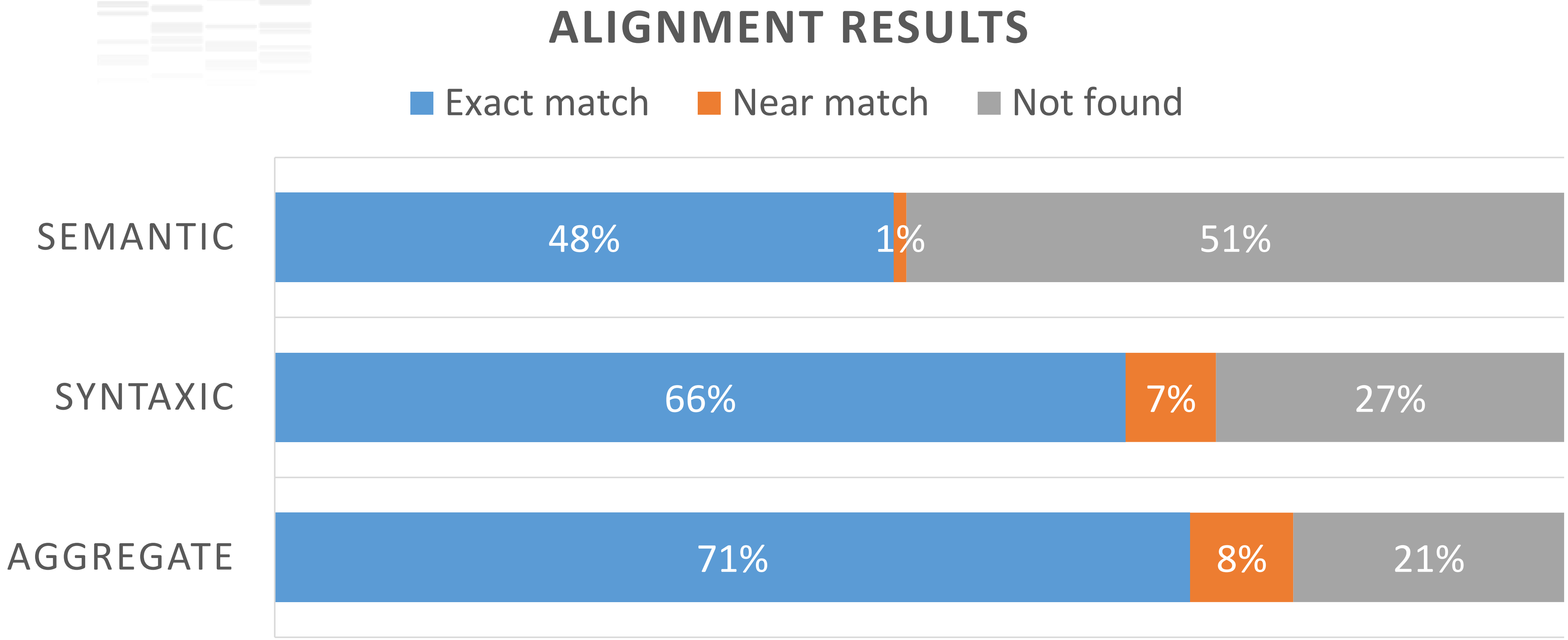


Results

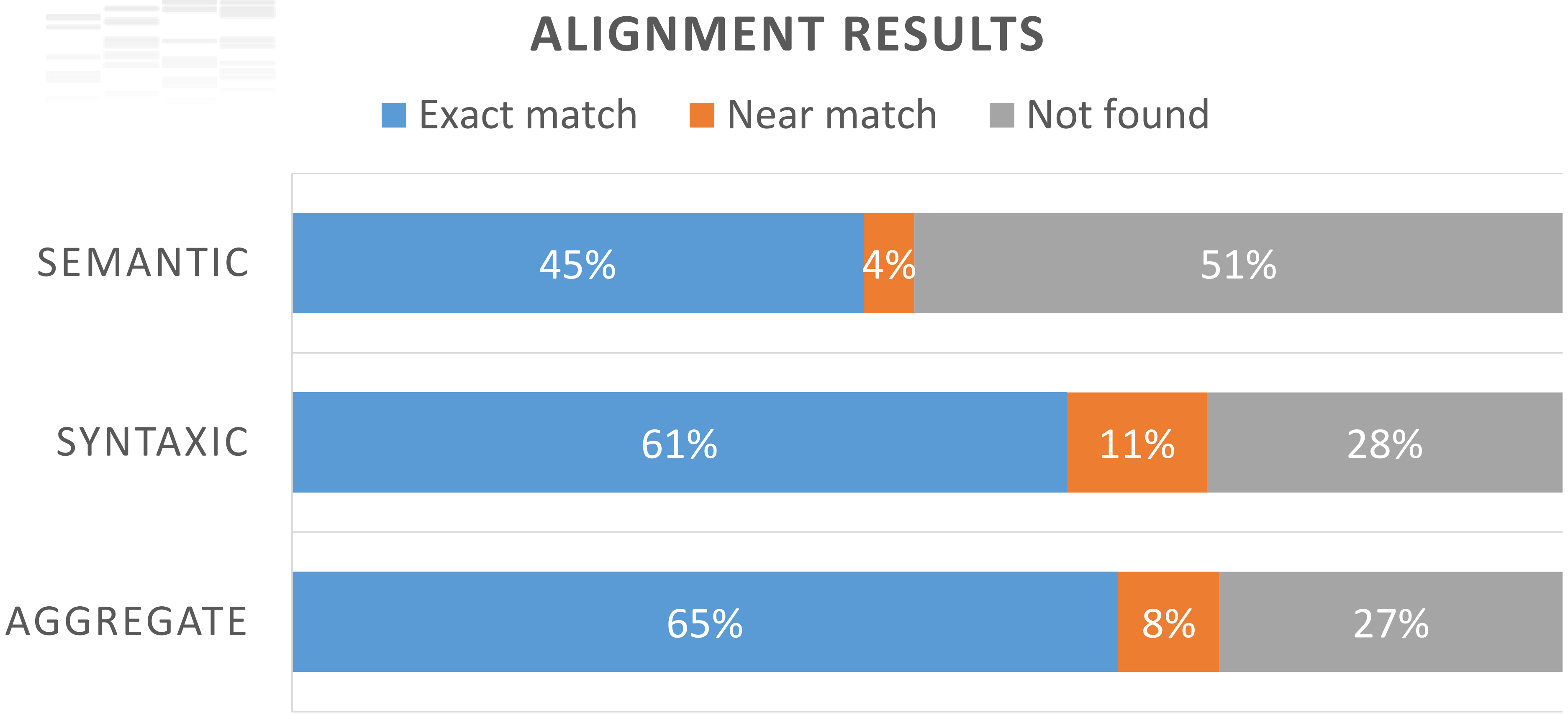
- Remarks
  - For 33% of products names only one Langual facet is present
  - Results are better if we take into account products with more than one facet : ~80% instead of 48% for the semantic approach on the previous slide

  → Good Langual is important annotation

Find the closest FoodOn family – Based on 100 meat products



Find the closest FoodOn family – Based on 300 products



INRA SCIENCE & IMPACT