

Interrogation unifiée de données nutritionnelles hétérogènes

Julien Cufi¹, Patrice Buche¹, Brett Choquet¹, Stéphane Dervaux², Juliette Dibie²,
Liliana Ibanescu², Alrick Oudot¹ et Magali Weber³

¹INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2,
LIRMM, Montpellier, France

²UMR MIA-Paris, AgroParisTech, INRA, UniversitéParis-Saclay, 75005 Paris, France

³IBIA, INRA, France

Dans le cadre du projet Meatylab nous disposons de plusieurs bases de données nutritionnelles provenant d'agences et de pays différents. Chaque agence a sa manière de décrire un produit, que ce soit en terme de labellisation ou de catégorisation dans des référentiels différents les uns des autres.

Notre objectif est de proposer un outil de requêtage multi-bases basé sur une ontologie de permettant de les unifier.

A titre d'exemple cet outil devra être capable à partir d'un concept du vocabulaire (par exemple 'Porc'), de pouvoir récupérer l'ensemble des produits (et leurs données nutritionnelles) y faisant référence (par exemple 'Épaule de porc, crue' dans la base française et 'Varkensschouderlappen rauw' dans la base néerlandaise).

Pour y parvenir, nous disposons, pour certains produits, d'une connaissance supplémentaire. Les agences nationales ont utilisé un vocabulaire contrôlé, LanguaL, pour décrire les aliments sur plusieurs aspects (ou facettes). Ces descriptions fournissent un enrichissement sémantique que nous allons utiliser pour lier les produits entre eux. Cependant seule une partie de ces produits étant décrite avec ce vocabulaire nous devons donc disposer d'une méthode supplémentaire prenant également en compte les labels des produits.

Le pivot de tous ces vocabulaires est FoodOn, une ontologie sur le domaine agroalimentaire intégrant un ensemble de produits décrits avec LanguaL. Cet ensemble de produits sera notre noyau d'intégration pour les différents produits.

L'objectif de notre approche est donc de pouvoir aligner les produits de nos différentes bases sur FoodOn, en se basant sur les descriptions en facettes LanguaL (approche sémantique) en plus des labels des produits (approche syntaxique).

Nous utilisons une combinaison de ces deux approches pour pallier à la fois à l'absence de description en facettes pour certains produits, mais aussi aux lacunes d'une comparaison purement syntaxique. Les résultats obtenus seront ensuite comparés à différents jeux de données de référence établis par des experts du domaine afin d'avoir une idée de la qualité de nos outils.

Dans cette présentation nous présenterons les différentes approches, les résultats obtenus ainsi qu'un exemple d'interrogation exploitant ces résultats.