Séminaire du réseau InOvive du 21 septembre 2021
Ce réseau, démarré en 2011, a pour objectif de partager et de collaborer autour de méthodes et outils pour intégrer et gérer des masses de données hétérogènes, en utilisant les ontologies, dans un contexte général de la Science Ouverte, avec des applications en sciences du vivant, de l'agronomie et de l'agro-alimentaire.

Annotation des Bulletins de Santé du Végétal

Anna Chepaikina (TSCF) , Robert Bossy (MAIAGE), Catherine Roussey (TSCF), Stephan Bernard (TSCF)

Le projet ANR “Des Données aux Connaissances en Agronomie et Biodiversité” (D2KAB) (www.d2kab.org) vise à faciliter l’accès aux données scientifiques d’agronomie et de biodiversité en les promouvant en connaissances Facile à trouver, Accessible, Interopérable et Réutilisable (FAIR) formalisées avec les technologies Web sémantique. Une des tâches du projet se focalise notamment sur le développement d’un navigateur web augmenté pour les bulletins d'alertes agricoles français, intitulés Bulletins de Santé du Végétal (BSV).

Dans le cadre de cette tâche, nous avons élaboré un plan d’annotation sémantique des bulletins récoltés à partir des sites web des DRAAF. Les bulletins sont préalablement transformés de leur format initial (pdf) dans un format textuel (html) afin de faciliter le traitement automatique de leur contenu. Différents corpus de BSV ont été construits et sont disponibles sur le web de données liées. Les corpus se différencient par les types de cultures mentionnées (grandes cultures, vignes, etc…) ou leur mode de construction (manuelle ou tirage aléatoire) .

L’annotation sémantique se réalise par le biais du moteur AlvisNLP, développé par l’équipe Bibliome. Elle consiste à projeter la partie terminologique de ressources sémantiques sur la version textuelle des BSV. Deux ressources sémantiques ont été projetées sur les corpus des BSV : le thesaurus « French Crop Usage » et les bases de connaissances peuplant l’ontologie « BBCH based Plant Phenological Description Ontology ».
Plus particulièrement, nous optons pour l’utilisation des méthodes suivantes de projection:
(1) Construction du dictionnaire à partir des ressources sémantiques (rdf) et association des étiquettes des concepts du dictionnaire aux entités nommées extraites des textes ;
(2) Classification des entités nommées en comparant la structure syntaxique de l’entité et des étiquettes des concepts ;
(3) Application des patrons syntaxiques, basée sur les règles d’une grammaire régulière.

Lors du séminaire nous présenterons les chaînes de traitements nécessaires à la tâche d'annotation ainsi que les premiers résultats de l'annotation sémantique des bulletins. Nous aborderons également certains aspects de la nature des données et comment elle influence le choix des processus de traitement automatique de la langue. Les annotations finales sont modélisées à l'aide de l'ontologie Web Annotation Data Model du W3C. Nous illustrerons l'usage de cette ontologie par plusieurs exemples d'annotations possibles des bulletins agricoles.