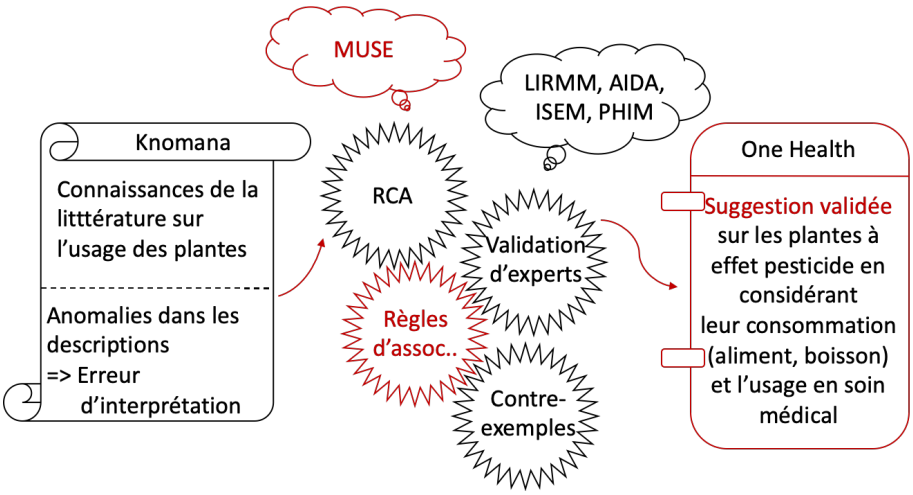
MUSE, Bourse de MASTER 2021

Détection et correction d'anomalies dans une base de connaissances sur l'usage des plantes à effet pesticide et antibiotique

Encadrants : Marianne Huchard, Pierre Martin, Pierre Silvie

Mails de contact : marianne.huchard@lirmm.fr, pierre.martin@cirad.fr, pierre.silvie@cirad.fr

Lieu du stage : LIRMM, 161 Rue Ada, 34095 Montpellier



Résumé :
Réduire l'utilisation des pesticides et des antibiotiques est un défi majeur pour l'environnement. Une des alternatives à ces produits de synthèse est l'utilisation des plantes. A cette fin, la base de connaissances Knomana rassemble 42 000 descriptions de l'utilisation de plantes, principalement du Sud, ayant des effets pesticides, antifongiques, antibactériens et antiparasitaires, pour la santé animale, végétale, humaine et publique (Silvie et al. 2021). L'ambition de l'équipe de chercheurs de l'UPR AIDA, de l'UMR ISEM, de l'UMR PHIM et de l'UMR LIRMM est de mettre à disposition des agriculteurs un logiciel d'exploration (incluant donc un mécanisme d'inférence) de Knomana qui propose des solutions adaptées aux besoins des utilisateurs et valorise la biodiversité locale tout en s'inscrivant dans l'approche One-Health. Par exemple, dans (Mahrach et al. 2020), nous avons développé un algorithme qui identifie les plantes pesticides de Knomana qui ne sont pas nocives pour l'homme, c'est-à-dire que la plante pesticide est non toxique si la culture est consommée (boisson ou nourriture) ou doit avoir une utilisation limitée et consciente si la plante pesticide est déjà utilisée en soin médical.
Les descriptions d'utilisation des plantes incluses dans Knomana sont extraites de la littérature scientifique. Certains algorithmes permettent déjà de corriger les erreurs typographiques (territoire, etc.) et d'autres de gérer les synonymes des noms d'espèces à partir de sites web de référence (par exemple PlantsOfTheWorld, CatalogueOfLife et FishBase). Cependant, il existe encore des anomalies dans ces descriptions, c'est-à-dire des informations comportant des valeurs incorrectes ou manquantes. Le problème causé par ces anomalies est qu'elles altèrent les résultats obtenus par l'algorithme d'exploration, qui est basé sur l'analyse des concepts relationnels (RCA ; Hacène et al. 2013). RCA est une méthode de classification non supervisée basée sur la théorie des treillis. C'est l'une des méthodes d'intelligence artificielle symbolique qui permet d'obtenir des résultats explicables. Vis-à-vis de Knomana, RCA permet de considérer la relation ternaire qui structure la description d'une plante utilisée pour protéger un système (parcelle cultivée, céréales stockées, piscine aquacole, être humain, etc.) contre un parasite (insectes, bactéries, champignons, etc.). La correction des anomalies est donc indispensable avant de fournir, aux utilisateurs finaux, un tel outil d'exploration, composé d'un algorithme d'exploration et de l'ensemble de connaissances.

L'objectif du stage est donc de développer une méthodologie et des algorithmes pour la détection et la correction semi-automatique de ces anomalies. L'approche sera basée sur l'extraction d'un ensemble synthétique de règles d'association résumant les principales informations fournies par RCA sur Knomana. Les règles seront organisées en utilisant des modèles sémantiques et des mesures quantitatives d'intérêt (par exemple support, confiance et lift) permettant de mettre en évidence les anomalies et les incomplétudes, pour être présentées aux experts. En retour, les experts valideront certaines des règles (qui seront considérées comme acquises), en invalideront d'autres, proposeront des contre-exemples, des compléments ou des corrections qui amélioreront la qualité du jeu de données. Cette méthode permettra aux experts d'améliorer continuellement Knomana et d'accroître les connaissances acquises. L'originalité du travail réside dans l'utilisation des règles comme révélateur d'anomalies, alors qu'elles sont habituellement utilisées pour révéler des modèles de connaissances fréquents.
Le travail attendu pendant le stage comprendra quatre étapes : (1) déterminer les types de règles d'association les plus susceptibles de révéler des anomalies ; par type, nous entendons le respect de modèles et de mesures d'intérêt ; (2) produire un algorithme qui extrait les règles et leurs modèles sémantiques et mesures d'intérêt associés ; (3) développer une méthode de classement pour les présenter aux experts ; (4) développer un prototype de logiciel qui prend en compte les retours des experts pour corriger les anomalies reconnues. Ce travail est un prérequis pour le développement d'une méthode d'analyse de risque appliquée à la sélection de plantes à effets pesticides et antibiotiques répertoriées dans une base de connaissances.
Le travail attendu consistera en :
- la conduite d'une revue de la littérature sur les règles combinées avec l'analyse formelle de concepts (FCA) ou son extension RCA.
- le développement de l'algorithme d'extraction des règles sur la classification de Knomana basée sur RCA.
- la conception et la mise en œuvre d'un algorithme qui calcule des modèles pertinents et des mesures d'intérêt sur les règles les plus susceptibles de révéler des anomalies.
- la conception de la structure et du processus de retour d'information avec les experts d'AIDA, du PHIM et de l'ISEM.
- le développement et l'évaluation d'un prototype de logiciel qui corrige les anomalies sur la base des retours d'information.

L'algorithme et le prototype logiciel seront implémentés en utilisant le langage Java. Le calcul de la classification sera effectué à l'aide de FCA4J (http://www.lirmm.fr/fca4j), une bibliothèque logicielle consacrée à FCA et RCA.

Références

Hacene MR, Huchard M, Napoli A, Valtchev P. Relational concept analysis: mining concept lattices from multi-relational data. Ann. Math. Artif. Intell. 2013; 67(1):81-108.
Mahrach L., Gutierrez A, Huchard M, Keip P, Marnotte P, Silvie P, Martin P. Combining Implications and Conceptual Analysis to Learn from a Pesticidal Plant Knowledge Base. Proceedings of International Conference on Conceptual Structure (ICCS) 2021; LNCS, volume 12879; Springer; p 57-72 (2021)
Saoud J., Gutierrez A., Huchard M., Marnotte P., Silvie P., Martin P. Explicit versus Tacit Knowledge in Duquenne-Guigues Basis of Implications. Preliminary Results. RealDataFCA workshop at ICFCA, pp 6, https://seafile.unistra.fr/f/27f540acdef347ffb228/, 2021.
Silvie PJ, Martin P, Huchard M, Keip P, Gutierrez A, Sarter S. Prototyping a Knowledge-Based System to Identify Botanical Extracts for Plant Health in Sub-Saharan Africa. Plants. 2021;10(5):896. https://doi.org/10.3390/plants10050896