Proposition de stage de Master 2 : début février à juin/juillet 2020 (dates indicatives)

Nom de l'entreprise ou du laboratoire : URGI, INRA

Adresse où se déroulera le stage : INRA, Route de Saint-Cyr, 78000 Versailles

Responsable du stage (personne qui sera contactée par les candidats):
- Nom, Prénom : Pommier Cyril
- Statut (Ing, chercheur, DR, MCF, Pr, autre) : Ingénieur
- Coordonnées (mél, tél) : cyril.pommier@inra.fr

Titre du stage : **Automatisation et généralisation d'intégration de données issues de fouille de texte dans un système d'information**

Mots clés résumant les méthodes et techniques à utiliser au cours du stage :

- Pipelines et containerisation : Python, Nextflow, VRE, Docker, GNU/Linux, Shell.
- Interfaces et visualisation : Java, Elasticsearch, Web services REST, Spring Boot, Javascript, Angular

Résumé du projet de stage :
Un des enjeux de la biologie végétale est d'étudier, voire de prédire, le phénotype d'une plante et sa capacité à s'adapter à des stress à partir de données génétiques, génomiques et environnementales. Ce lien entre génotype et phénotype s'effectue via des approches de détection de QTL ou de GWAS, souvent complétées par une approche gène candidat. GnpIS est une base de données développée par l'INRA qui permet de stocker des jeux de données de génétique et de génomique chez les plantes (https://urgi.versailles.inra.fr/gnpis/). En complément, des données de ce type se trouvent de façon non structurée dans la bibliographie. Une preuve de concept a été réalisée sur le blé pour développer un pipeline de fouille de texte permettant de collecter ce type d'information dans des articles, de les annoter sémantiquement et de donner accès au corpus bibliographique dans un portail de recherche (e.g. https://urgi.versailles.inra.fr/wheatis/ avec la recherche « yield rust ») aux côtés de jeux de données expérimentales. L'URGI propose un stage de Master visant à automatiser l'extraction des données depuis la littérature scientifique en s'appuyant sur ces premiers résultats.
Le premier objectif du stage sera donc d'automatiser le processus utilisé pour la preuve de concept en utilisant des technologies de type pipeline et containerisation (Nextflow, VRE, Jenkins, Docker, …). Ce travail se fera en collaboration avec l'équipe Bibliome de l'UMR MaIAGe. Cela permettra de générer un jeu de données à jour et de le publier dans le portail de

GnpIS de façon régulière. Cet objectif inclut la mise à jours des interfaces web de GnpIS pour améliorer la visualisation du text mining (les technologies utilisées dans GnpIS sont Elasticsearch, Spring Boot et Angular).
Le deuxième objectif consistera à généraliser ce processus à d'autres espèces en s'appuyant en particulier sur l'ontologie référençant les traits étudiés chez le blé (http://agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE) et sur le modèle de connaissance développés par l'équipe Bibliome. Un premier essai sera fait en utilisant la vigne comme cible. Un nouveau corpus bibliographique sera ainsi extrait des ressources internationales (PMC, Web of Science, Europmc, …). Parallèlement, l'ontologie de traits pour la vigne (https://urgi.versailles.inra.fr/ontology#termIdentifier=CO_356) sera utilisée pour enrichir la Wheat Phenotype Ontology qui sera ainsi étendue pour traiter plusieurs espèces. Une attention particulière sera portée sur les traits en lien avec la phénologie et la résistance aux maladies.
Ces travaux permettront de poser les bases de l'utilisation des outils disponibles actuellement en fouille de texte et alignement de vocabulaires nécessaire à l'extraction des connaissances dans la littérature. Cette infrastructure permettra d'enrichir les données déposées dans GnpIS avec des connaissances issues du *text mining* et permettant d'établir dans le futur des liens entre des gènes ou des régions génomiques et la variation de caractères phénotypiques.

Montant des indemnités de stage

Approximativement 500€.

Modalités de candidature

Les candidatures (CV + lettre de motivation) doivent être adressées jusqu'au 31/12/2019 par courriel à cyril.pommier@inra.fr avec l'objet suivant : [2020-stage-text-mining]