SUJET DE STAGE

NOM, prénom de la personne proposant le stage :

Michotey Célia

Francillonne Nicolas

Adresse Professionnelle : INRAE, Centre de recherche de Versailles, bat.18 RD10, Route de Saint Cyr 78026 Versailles France
Numéro de téléphone :

Adresse électronique : celia.michotey@inrae.fr et nicolas.francillonne@inrae.fr

Etablissement d'accueil / Entreprise :

NOM : Institut National de recherche pour l'agriculture, l'alimentation et l'environnement INRAE, Unité de Recherche en Génomique-Info (URGI)
Domaine d'expertise de l'entreprise/établissement : Développement d'outils et acquisition de connaissances sur la structure, l'évolution et le fonctionnement du génome.

Nom du représentant légal de l'établissement /entreprise : Anne-Francoise Adam-Blondon

Adresse postale et N° de téléphone : INRAE, Centre de recherche de Versailles, bat.18 RD10, Route de Saint Cyr 78026 Versailles France

Unité d'appartenance / Laboratoire :

Nom complet et code unité le cas échéant *:* Unité de Recherche en Génomique-Info (URGI) - UR 1164 INRAE Versailles
Nom de l'équipe d'accueil ou du service : URGI

Adresse du service d'accueil (si différente de celle de l'établissement/entreprise) :

Nom, prénom, statut et spécialité *(biologiste, informaticien, statisticien, bioinformaticien, biostatisticien, etc.)* **de la personne qui encadrera le stagiaire** : Célia Michotey bioinformaticien, analyse des génomes,

Numéro(s) de téléphone où l'on peut la joindre :

Adresse électronique : celia.michotey@inrae.fr

Autre(s) encadrant(s) : Nicolas Francillonne bioinformaticien, datamanager nicolas.francillonne@inrae.fr

Titre du stage :

Intégration de données génétiques/génomiques dans une base de connaissance graphe

Mots clés :

Graphe de connaissances, intégration de données, ontologies, génétique, phénomique, génomique

Langages et outil de développement : Neo4j, Python, RDF, conteneur Docker ou Singularity/Apptainer, GNU/Linux, Shell.

Description du sujet (1 page maximum)

Contexte :

Un des enjeux de la biologie végétale est d'étudier, voire de prédire la capacité des plantes à s'adapter à des stress en s'appuyant sur des données génétiques, génomiques et environnementales. Pour y parvenir les chercheurs étudient des collections de ressources génétiques représentant la diversité existante d'une espèce et souhaitent disposer de connaissances intégrées aussi exhaustives que possible sur chacune des accessions/ressources génétiques de ces collections.

L'URGI est une unité de recherche basée sur le centre INRAE de Versailles-Saclay, dont un agent est missionné au sein du CNRGV sur le centre INRAE de Toulouse. Elle développe des approches basées sur les graphes de connaissances permettant d'intégrer des données hétérogènes dans l'optique d'apporter un appui efficace et rapide en termes d'exploration de données à la communauté scientifique.

L'objectif du stage proposé sera de compléter le graphe de données existant qui permet pour le moment de collecter et interroger les données et métadonnées génomiques et génétiques d'un large panel d'espèces de plantes d'intérêt (Blé, vigne, peuplier, chêne …). Cette extension se fera sous deux angles :

- L'introduction dans le graphe de nouvelles données sur les accessions des collections (variétés et accessions[1] de plantes, synonymie, caractérisation primaire et phénotypique sur la base d'ontologie ou du thésaurus INRAE à faire évoluer, provenance des données) ainsi que l'intégration avec d'autres données disponibles dans le système d'information GnpIS (génotypage, phénotypage, génomes et annotations, données omiques) pour permettre une exploration de la diversité intra-spécifique. L'intérêt de représenter dans le graphe l'héritage ou l'apparentement des accessions sera étudié.
- La mise en place d'un processus semi-automatisé permettant de réaliser un catalogue des ressources associés aux accessions pour faciliter leur sélection dans de nouvelles recherches. Un

effort particulier sera attendu sur l'automatisation des processus de collecte, notamment sur la mise à jour et l'ajout de nouvelles données. Nombre d'entre elles sont disponibles dans les bases de données de référence (EBI/NCBI, Phytozome-JGI...), d'autres le sont sous des formats tabulés, enfin des références croisées permettent de lier vers d'autres ressources externes, en partie de la bibliographie.

Objectifs :

Intégration des données hétérogènes, dans une base de données de type « graphe » (Neo4j).
Ces données hétérogènes hébergées au laboratoire et dans des entrepôts publics devront être traitées pour être insérées dans une base pilote pour faciliter l'intégration, l'enrichissement des données et leur exploitation.
Le(a) candidat(e) devra enfin pouvoir proposer une automatisation de l'insertion des données en base et des visualisations permettant une interrogation accessible et reproductible.

Compétences techniques recherchées :

• Maîtrise des commandes UNIX (shell) et de la programmation python.
• Connaissance en SGBD, connaissance du NoSQL sera un plus.
• Connaissance de la technologie Docker souhaitable

Références bibliographiques (facultatif) :

Ce sujet constitue un premier pas vers un travail de thèse : Non

Date de début du stage et durée estimée du stage :

A partir de début 2025 de 6 mois.

Montant (brut mensuel) de la rémunération proposée :

Indemnité de stage selon barème en vigueur (environ 550 euros net par mois)

Date de la proposition de stage et date limite de candidature :

Proposition de stage fin septembre 2024
Date limite fin décembre 2024