Proposition de sujet de stage de Master 2

# Normalisation automatique de variables issues de bases de données en agroécologie

## 1. Contexte

Les études agro-écologiques génèrent de nombreuses bases de données hétérogènes en termes de structure et de contenu, qui sont difficilement exploitables et nécessitent une curation pour être mobilisées dans des approches statistiques ou de modélisation. La curation consiste à sélectionner les données les plus pertinentes et les enrichir de métadonnées nécessaires à leur compréhension pour pouvoir les rendre accessibles, partageables et réutilisables (principes FAIR).

Pour annoter les données et augmenter la précision des termes utilisés, un collectif interdisciplinaire de chercheurs du CIRAD a construit un dictionnaire des variables (Auzoux *et al*, 2018). Une variable est constituée de termes sémantiques issus des connaissances expertes et d'ontologies de référence. La liste des variables du dictionnaire a été définie pour faciliter la comparaison et l'analyse des données, et les liens avec les modèles de culture.

Un premier travail exploratoire sur la curation de bases de données en agroécologie, constituées à partir de 28 expérimentations sur la canne à sucre à La Réunion, a été réalisé lors d'un stage de Master 2 (Ngaba, 2022). Il a permis de tester et de valider une approche de fouille de textes pour automatiser la normalisation des variables créées et utilisées par les chercheur.e.s pour décrire leurs données.

## 2. Objectifs du stage

L'objectif de ce stage est d'automatiser la labellisation des variables hétérogènes des chercheur.e.s issues des bases de données en agroécologie à partir d'une liste de variables standardisées (dictionnaire des variables). Plusieurs méthodes de fouille de texte seront mobilisées pour proposer les variables du dictionnaire les plus en phase avec les variables des bases de données :

- des mesures de proximité lexicale (Maedche *et al.*, 2002),
- des méthodes de proximités contextuelles (Salton *et al*, 1988) fondées sur la description des variables issues des bases de données,
- des méthodes de proximités contextuelles fondées sur des corpus : des contextes seront constitués à partir de corpus textuels et de méthodes de plongements de mots (Mikolov *et al.,* 2013) et de modèles de langues issus des méthodes d'apprentissage profond (Devlin *et al.,* 2019).

Au-delà d'une extension de la méthode en proposant des méthodes originales de fouille de texte, un objectif important de ce stage consiste à proposer une approche générique pour labelliser les données et faciliter l'interopérabilité des bases de données en agroécologie.

## 3. Démarche :

Ce stage se déroulera en 3 grande étapes :

Etape 1 **: Appropriation des données et codes**

- Prise en main du dictionnaire des variables et familiarisation avec les données issues d'expérimentation en agroécologie à La Réunion
- Prise en main du code développé dans le cadre du précédent stage et développement d'une interface utilisateur de base

Etape 2 : **Préparation de nouveaux jeux de données pour étudier la généricité de l'approche**

- Mise en forme et nettoyage des données provenant des essais mis en place par le CIRAD et ses partenaires
- Évaluation des méthodes mises en œuvre à l'étape 1 sur ces nouveaux jeux de données.

Etape 3 : **Extension de l'approche de mise en lien de variables**

- Constitution du corpus de variables à labelliser à partir de ces jeux de données.
- Utilisation et combinaison de méthodes de plongements de mots et de modèles de langues à partir de ces corpus pour constituer des contextes à associer aux variables. Ces derniers pourront être utilisés pour améliorer les méthodes de mise en lien des variables des chercheur.e.s aux variables du dictionnaire.

Dans le cadre d'une démarche science ouverte, les codes sources et les données seront mises à dispositions sur la forge logicielle et le Dataverse du CIRAD. Les résultats de ce stage pourront donner lieu à deux publications scientifiques (Data paper et article scientifique).

## 4. Mots-clés

Science des données, fouille de texte, base de données, agroécologie, sémantique, python

## 5. Profil du candidat

Le profil que nous recherchons, est un informaticien (Master 2 ou école d'ingénieur) ayant une formation en science des données ayant une maitrise des bases de données, des méthodes de fouille de texte et d'analyse de données. Une ouverture sur l'interdisciplinarité est indispensable pour pouvoir dialoguer avec les experts métiers.

## 6. Outils de développement envisagés

SGBD PostgreSQL, R studio, Python

## 7. Conditions de réalisation du stage

- Accueil à l'UMR TETIS à la Maison De la Télédétection sur le campus Agropolis de Montpellier
- Encadrement : Deux unités de recherche de #DigitAg (UR Aïda et UMR TETIS) sont impliquées dans cet encadrement. Le stagiaire évoluera dans une équipe pluridisciplinaire composée de deux informaticiens (Sandrine Auzoux et Mathieu Roche), un biostatisticien (Benjamin Heuclin), et deux agronomes (Aude Ripoche et Mathias Christina).
- Période de stage : de février/mars à juillet/août 2023 (6 mois)
- Rémunération : indemnité au tarif en vigueur : 600 euros/mois x 6 mois = 3600 € + tickets restaurant
- 1 mission sera réalisée à La Réunion en milieu de stage pour présenter les premiers résultats et pour discuter plus en détail avec les encadrants et les partenaires réunionnais de la généricité de l'approche.

## 8. Contact pour les candidatures

Envoyer un CV, les derniers relevé de notes et une lettre de motivation avant le 15/12/2022, à Sandrine Auzoux (sandrine.auzoux@cirad.fr, Tél : +262 2 62 72 78 66) et Mathieu Roche (mathieu.roche@cirad.fr , Tél : +33 4 67 55 86 12).

## 9. Bibliographie


S. Auzoux, M. Christina, F.-R. Goebel, A. Mansuy, D. Marion, A dictionary of variables to harmonize data from agro-ecological experiments on sugarcane, ISSCT, 2018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4171–4186, 2019.

A. Maedche, S. Staab, Measuring similarity between ontologies, Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 251–263.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013

B. Ngaba, M. Christina, A. Mansuy, J. Chetty, M. Soulé, M. Schwartz, B. Heuclin, S. Auzoux, Agroecological Practices to reduce wEED infestAtion In the tropicS, 2022. URL: https://dataverse.cirad.fr/dataverse/APEEDAIS.

B. Ngaba, Couplage d'un modèle de culture avec une plate-forme de capitalisation des données issues d'agroécosystèmes à La Réunion, 2022. URL: https://agritrop.cirad.fr/601877/.

G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24, 1988, 513–523.