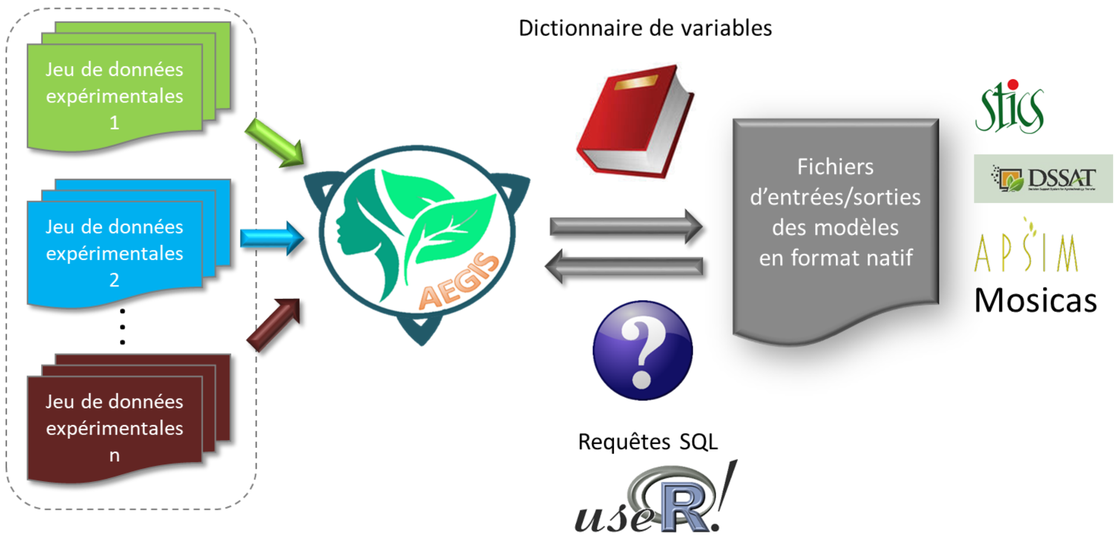
S. Auzoux[1], M.Christina, B. Heuclin[1], M. Roche[2], B. Ngaba[3]

[1]CIRAD UPR AIDA, Sainte-Clotilde, Réunion
[2]CIRAD, UMR TETIS, Montpellier, France
[3]Université Grenoble-Alpes, Grenoble, France

Du terrain à l'outil
Comment assurer le continuum données-modèles de culture ?

Problématique

- Chaque chercheur a son propre système d'annotation et de structuration des données
- Le système d'information AEGIS utilise un dictionnaire de variables basé sur du vocabulaire contrôlé
- Chaque modèle de culture a une terminologie propre pour décrire les paramètres de simulation



Cheminement de la données vers le modèle de culture

Mettre en relation des données qui ont la même signification mais décrites différemment

Méthodologie

1. Constituer 28 jeux de données normalisés au format AEGIS provenant d'essais canne à sucre sur les pratiques alternatives de lutte contre les adventices, réalisés de 2012 à 2021, sur le site expérimental de La Mare (eRcane – CIRAD).

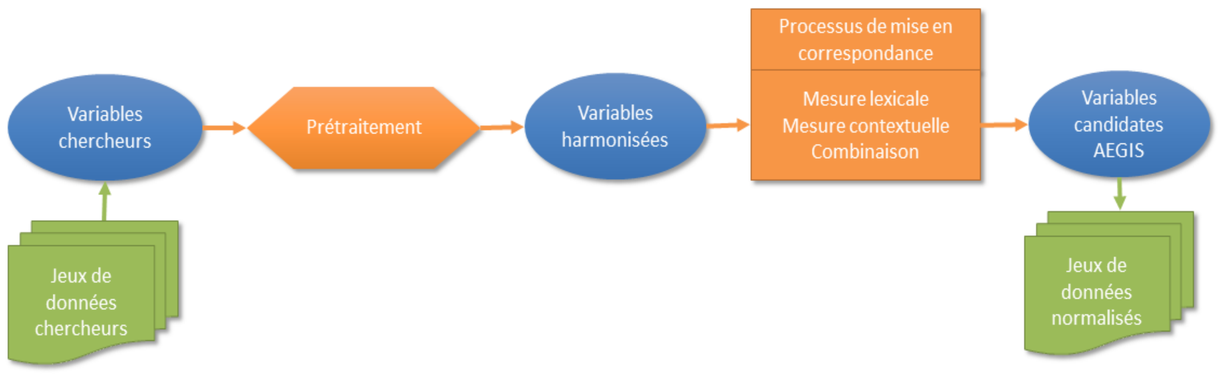


Figure 1 : Processus Global de traitement des variables

2. Mobiliser des méthodes de fouille de texte pour la normalisation des variables :
   - Mesure lexicale
   - Mesure contextuelle
   - Mesure globale : combinaison

3. Créer une chaine de traitement pour récupérer les données dans AEGIS et réaliser une simulation de croissance de la canne à sucre à partir du modèle STICS

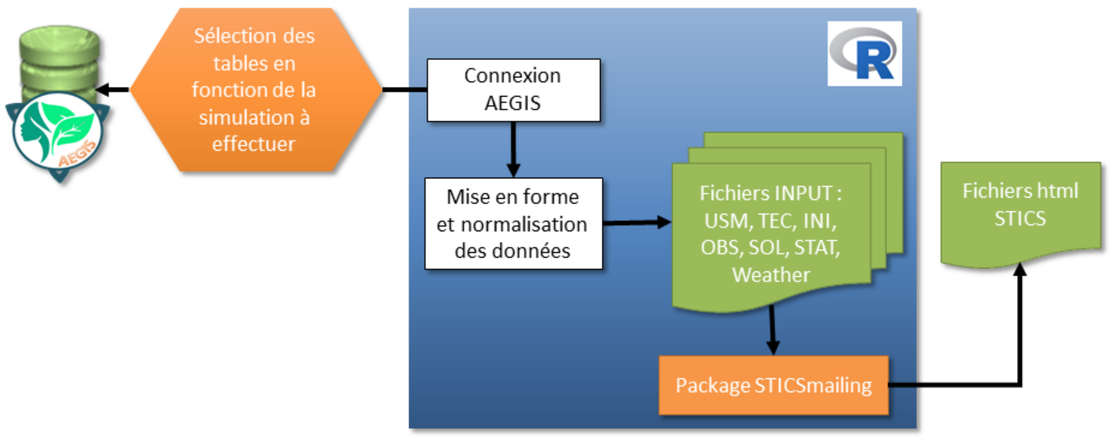
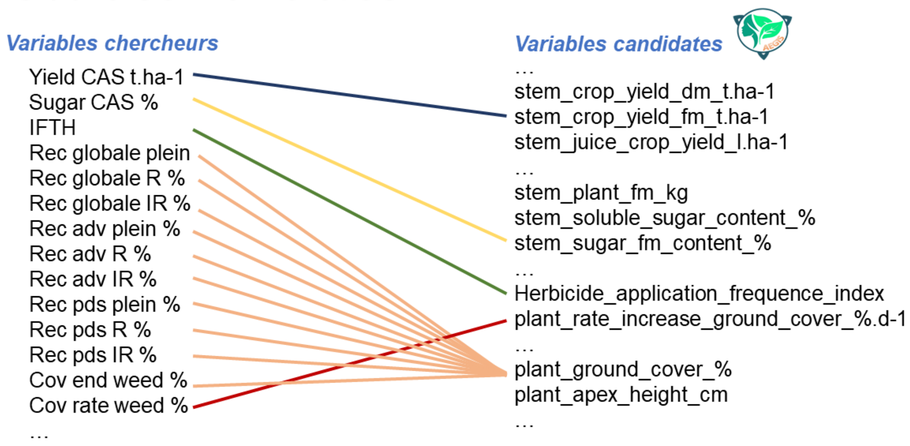


Figure 2 : Processus d'automatisation de la chaîne de traitement AEGIS - STICS

Résultats

- Dépôt et partage des données normalisées sur le dataverse du CIRAD https://dataverse.cirad.fr/dataverse/APEEDAIS
- 1er résultats encourageants : mise en correspondance de 73% des variables



- Dépôt sur Github des codes sources https://github.com/bilson98/STAGE_Cirad

Perspectives

- Améliorer la fouille de texte :
  - Exploiter les unités et échelles de valeurs des variables
  - Web crawling pour enrichir le contexte
  - Modèles de langues en deep learning
- Package R pour faciliter la normalisation des données et généraliser la chaine de traitement