

ECOFI : une nouvelle base de données générique pour faciliter la modélisation et l'analyse des jeux de données issues des expérimentations en agro-écologie

Auteurs :

Sandrine Auzoux, Jean-François Martiné, Romain Loison, Christophe Poser, Pascal Marnotte, Régis Goebel, Julie Dusserre (UR Aïda), Lauriane Rouan, Myriam Adam, David Pot (UMR AGAP).

Pour répondre à enjeux globaux tels que la sécurité alimentaire, la réduction des impacts environnementaux et le changement climatique, des approches agroécologiques sont développées qui visent à concevoir des systèmes de production performants, qui utilisent mais préservent les fonctionnalités des écosystèmes.

Les études en agroécologie génèrent des données complexes difficiles à gérer, à partager et à confronter. Il y a autant de données que d'unités d'observation (champs, parcelles, plantes, bioagresseurs, mauvaises herbes...) combinées avec des unités de temps (année, saison de culture, date spécifique...) et d'espèces végétales (cultures et plantes associées). Les jeux de données sont multidisciplinaires et transdisciplinaires (agronomie, entomologie, malherbologie et sciences économiques, culturelles, politiques et sociales). Ils comprennent des niveaux multiples d'échelles spatiales et temporelles d'observations. En agroécologie, le nombre d'observations n'est pas fixe et de nouvelles observations peuvent être effectuées au cours de l'expérimentation. De ce fait, les jeux de données ont des structures hétérogènes. Au niveau sémantique, ils contiennent des homonymes et des synonymes. Il existe un réel problème de structuration et de standardisation des jeux de données expérimentales. De plus, les données sont habituellement recueillies dans des cahiers de terrain ou dans des fichiers informatiques mal documentés et non sécurisés de formats différents et stockés sur plusieurs périphériques de stockage. Les erreurs de manipulation des données et la perte de données sont fréquentes. Cette gestion des données ne convient pas à la modélisation des systèmes de culture et n'est pas adaptée pour les analyses statistiques. Elle a révélé ses limites en matière d'accessibilité, d'interopérabilité et de réutilisation de l'information.

La technologie des bases de données fournit une solution pour gérer les données issues d'expérimentations sur le terrain. Elle résout le problème de sécurité et de stockage permanent des données, normalise la description des données, offre un contexte multi-utilisateur et facilite l'accès aux données, l'exploration et l'analyse des données.

Pour résoudre ces problèmes de gestion des données provenant d'expérimentations menées dans le domaine de l'agroécologie, une base de données générique, appelée ECOFI, a été développée en appliquant la technologie des métadonnées. Cette technologie est pertinente car elle permet d'ajouter autant d'éléments de données sans modifier le schéma de la base de données. Un autre avantage important est qu'elle minimise le nombre de tables, de colonnes et de cellules vides. Elle améliore également la performance des requêtes de la base de données. L'intérêt de la généricité est d'éviter de développer une base de données spécifique pour chaque étude et de limiter le nombre de bases de données. C'est un gain de temps pour les gestionnaires de projet puisqu'il n'y a pas à redévelopper de bases de données. Pour les scientifiques utilisateurs de bases de données, cela facilite considérablement l'exploitation de leurs données.

Les données recueillies dans [®]ECOFI sont annotées à l'aide des termes issus d'ontologies. Les ontologies facilitent l'harmonisation de la saisie des données, permettent de nettoyer la base de données et assurent l'interopérabilité avec d'autres bases de données agronomiques. Afin d'améliorer la précision des termes utilisés par les chercheurs pour caractériser leurs données expérimentales, les vocabulaires contrôlés conformes à l'ontologie des plantes (PO) et à l'ontologie des cultures (Crop Ontology) ont été utilisés. Ainsi, les données sont à disposition sous le même format et utilisent une sémantique commune.

L'implémentation d'[®]ECOFI a été réalisée sous le système de gestion de base de données Open Source PostgreSQL. Une approche multidisciplinaire a permis de soulever des questions pertinentes au cours de la phase de modélisation de la base de données. L'équipe de travail était composée de statisticiens pour la conception expérimentale, de généticiens et de sélectionneurs pour la définition du matériel végétal, d'agronomes pour définir les observations, les notations et les interventions techniques, d'informaticiens pour la conception de la base de données, la définition des métadonnées et l'ontologie.

Nous avons utilisé SCRUM, un cadre de développement logiciel agile dans lequel les équipes multidisciplinaires développent des produits de manière itérative, incrémentielle et adaptative.

Les applications d'ECOFI :

De nombreuses opérations sont nécessaires pour extraire des jeux de données expérimentales et pour créer des fichiers d'entrée de modèles de système de culture. La manipulation des données est fastidieuse et difficile à automatiser pour la modélisation. Dans le cadre du projet Gloofoods DATA2050, une application conçue pour faciliter la création de fichiers d'entrée de modèles (DSSAT, APSIM, STICS, SARAH) a été développée. Il s'agit de DataMill. Par des requêtes simples, ®ECOFI alimente la base de données de DataMill et permet ainsi, de créer automatiquement les fichiers de saisie des modèles de culture. ®ECOFI est également utilisé pour récupérer les données des fichiers de sortie. ®ECOFI fournit directement le modèle de simulation de croissance de la canne à sucre, MOSICAS.

®ECOFI contient 20 ans de données expérimentales sur la physiologie de la canne à sucre dans l'île de la Réunion, qui permettent facilement d'implémenter les fichiers d'entrée du modèle de croissance de la canne à sucre Mosicas®. La base de données a aussi été utilisée dans le cadre du projet SYPECAR (système de production énergétique à base de canne à la Réunion) pour gérer des données expérimentales sur la canne fibre. Dans le cadre du projet Data2050 du programme GloFoods, ®ECOFI a permis de recueillir des données sur le Sorgho à partir d'expériences multi-sites en Afrique. Dans le cadre du projet BFF (Biomass For the Future) elle est utilisée pour gérer, explorer et analyser les données phénotypiques sur le Sorgho. Pour faciliter l'utilisation et l'accès à ®ECOFI, un système d'information appelé DAPHNE est en cours de développement.