Gérer et maintenir un outil de diagnostic des systèmes agri-alimentaires accessible en ligne : SI-BOAT

Pierre CASEL[1]
Myriam GRILLOT[1]

Correspondance
siboat@inrae.fr

Résumé.

Regrouper des bases de données nationales pour faciliter les diagnostics de systèmes et filières agricoles territoriaux ? C'est le projet de SI-BOAT, un système d'information créé et maintenu depuis 2019 par deux équipes de recherche : AGIR, unité mixte de recherche INRAE-INPT et le LESSEM, unité de recherche INRAE. Il est inspiré du Système d'Information Dédié au Territoire (SIDDT) qui permet des diagnostics socio-économiques. Leur fonctionnement est de i) regrouper des données de sources multiples généralement accessibles librement en ligne et couvrant le territoire national et ii) proposer une interface en ligne pour accéder à des indicateurs spécifiques pour un territoire infra-régional choisi par l'utilisateur. SI-BOAT regroupe et mobilise actuellement 20 bases de données publiques et libres d'accès.
Parmi la diversité d'enjeux autour de la gestion et maintenance de SI-BOAT, nous centrons cet article sur les objectifs suivants : proposer des données les plus à jour possible, appréhender au mieux leur l'évolution et, dans la mesure du possible, alléger le temps de traitement pour la maintenance. Nous présentons les procédures mises en place pour les cas d'ajout et/ou modification de bases et tables de données, ainsi que pour la mise à jour les indicateurs qui en découle. Ces procédures structurent l'import des données brutes depuis le web, leur prétraitement et stockage sur notre serveur de bases de données, jusqu'à la mise en forme des indicateurs avec leur diffusion sur l'interface web. La gestion des métadonnées est un élément clé sur lequel nous revenons. Elle permet de faciliter le travail des développeurs pour la mise à jour des données et, de manière quasi-automatique, celle des indicateurs existants. Elle facilite également l'échange et le suivi avec les thématiciens et l'information des utilisateurs de l'interface sur les données et indicateurs. Face à un engouement croissant pour ce type de système d'information, nous pointons l'importance de tenir compte de l'organisation, du temps et de la coopération nécessaire à leur gestion et maintenance. Les perspectives pour SI-BOAT seraient d'intégrer des retours utilisateurs afin de mieux identifier leurs utilisations de l'outil et les sujets traités. Cela nous permettrait de consolider notre ensemble de bases de données et d'indicateurs clés, tout en identifiant d'éventuels besoins pour l'ajout de nouvelles bases de données et d'indicateurs. Toutefois, malgré notre travail sur les procédures, le temps de travail reste la principale contrainte et SI-BOAT dépend de la disponibilité des développeurs et thématiciens.

Mots-clés
Maintenance d'un système d'information ; données accessibles publiquement ; multi-niveaux ; métadonnées ; interface web ; territoire

1 AGIR, Univ Toulouse, INRAE, Castanet-Tolosan, France.

Managing and maintaining an online diagnostic tool for agri-food systems: SI-BOAT

Pierre CASEL[1]
Myriam GRILLOT[1]

Correspondence
siboat@inrae.fr

Abstract.

Bringing together multiple national databases to facilitate the diagnosis of local agricultural systems and sectors—this is the goal of SI-BOAT, an information system created and maintained since 2019 by two research teams: the AGIR joint research unit (INRAE-INPT) and the LESSEM research unit (INRAE). SI-BOAT is based on the Système d'Information Dédié au Territoire (SIDDT), which is used for socio-economic diagnostics. The system operates by i) bringing together data from multiple sources, generally freely accessible online and covering the national territory, and ii) offering an online interface for accessing specific indicators for a sub-regional territory selected by the user. SI-BOAT currently consolidates and utilizes 20 public and freely accessible databases.
Among the wide range of issues surrounding the management and maintenance of SI-BOAT, this article focuses on the following objectives: ensuring that data is as up to date as possible, gaining better insight into its evolution, and, where feasible, reducing maintenance time.
We present the procedures implemented for adding and/or modifying databases and tables, as well as for updating the resulting indicators. These procedures structure the import of raw data from the web, its pre-processing and storage on our database server, and the formatting of indicators for dissemination through the web interface. Metadata management is a key element we highlight: it facilitates developers' work in updating data and, almost automatically, existing indicators. It also supports exchanges and follow-up with thematic specialists and provides interface users with essential information on data and indicators.
In view of the growing popularity of this type of information system, we emphasize the importance of considering the organization, time, and collaboration needed for their management and maintenance. Looking ahead, SI-BOAT aims to incorporate user feedback to better understand how the tool is used and the topics it covers. This would allow us to expand our set of databases and key indicators while identifying potential needs for adding new data and indicators. However, despite our work to streamline processes, time remains the primary constraint, and SI-BOAT depends on the availability of developers and thematic specialists.

Keywords
Information system maintenance; publicly available data; multi-level; metadata; web interface; landscape

1 AGIR, Univ Toulouse, INRAE, Castanet-Tolosan, France.

Introduction

Réaliser des diagnostics sur les systèmes et filières agricoles d'un territoire requiert d'accéder à une diversité de données (cheptels, cultures, spécifications, etc.) de sources diverses (recensements agricoles et de population, déclarations, etc.). Regrouper ces données est un travail qui peut s'avérer coûteux en temps et demande une bonne connaissance des différentes bases existantes et disponibles.

Depuis 2005, le Système d'information dédié au territoire (SIDDT[1]) est développé dans l'unité de recherche du LESSEM à INRAE. Accessible librement en ligne, il permet de réaliser des diagnostics de territoire sur une diversité de thèmes (infrastructures, économie, agriculture, etc.). SIDDT regroupe des données nationales multi-sources, elles-mêmes généralement accessibles librement en ligne.

Basé sur le même fonctionnement (libre accès, données nationales et multi-sources), SI-BOAT est un système d'information (SI) permettant spécifiquement de réaliser des diagnostics sur les filières agricoles et le système agro-alimentaire des territoires pour des utilisations alimentaires ou non (ex. énergétiques). Développé depuis 2018, il est maintenu conjointement par deux unités de recherche : AGIR (unité mixte de recherche INRAE-INPT) et le LESSEM (unité de recherche INRAE). Il est accessible en ligne depuis août 2023[2] à toute personne intéressée par la réalisation de tels diagnostics. Il est utilisé dans les milieux du développement territorial (ex. diagnostic pour les plans d'alimentation territoriaux-PAT), de la recherche (ex. description de cas d'étude) et de la formation (ex. travaux étudiants).

Concrètement, l'utilisateur se connecte sur l'interface web, choisit les limites de son territoire d'étude, puis SI-BOAT calcule, prépare et affiche les indicateurs qui caractérisent ce territoire d'étude. Lorsqu'ils sont disponibles, SI-BOAT indique les résultats aux niveaux départemental, régional ou encore national comme repères de comparaison. SI-BOAT regroupe et mobilise actuellement 20 bases de données publiques et libres d'accès, comme celles du Recensement Agricole, de l'Agence Bio, des statistiques agricoles annuelles ou encore du registre parcellaire graphique.

Un de nos objectifs, en tant qu'initiateurs de SI-BOAT, est de diffuser des données les plus à jour possible, car nous estimons que c'est un critère majeur pour son utilité et sa pérennité. Par conséquent, notre défi est de réussir à appréhender au mieux l'évolution des données mobilisées par SI-BOAT et d'anticiper le traitement de nouvelles données. Pour cela, nous avons mis en place des procédures permettant de répondre aux questions suivantes : comment mettre à jour ou ajouter de nouvelles bases de données ? Comment mettre à jour les indicateurs qui en découlent ? Comment structurer et maintenir à jour les métadonnées ? Pour y répondre, nous avons identifié trois étapes dans le traitement des données dans SI-BOAT : 1) l'import des données brutes depuis le web ; 2) leur prétraitement et stockage sur notre serveur de bases de données ; 3) la mise en forme des indicateurs pour leur diffusion sur l'interface web. Les procédures de mise à jour concernent chacune de ces étapes.

Cet article présente SI-BOAT et son flux de données, décrit les trois étapes de mise à jour des données et fait le point sur les avantages et les axes d'amélioration de cette organisation de maintenance.

Présentation de SI-BOAT

SI-BOAT : vue globale

Le fonctionnement de SI-BOAT repose sur trois composantes (Figure 1) : d'un côté, l'utilisateur qui cherche des informations sur un territoire donné, de l'autre, le serveur avec son **Système de Gestion de Base de Données Relationnelles** (SGBDR) et l'interface web, qui fait le lien entre les deux.

Pour utiliser SI-BOAT, l'utilisateur navigue sur l'interface web. Il doit d'abord définir une zone d'étude géographique délimitant le territoire pour lequel il souhaite obtenir des informations. Trois possibilités combinables sont proposées : 1) choisir des zonages administratifs existants (Établissements Publics de Coopération Intercommunale, petite région agricole, parc naturel régional) ; 2) faire sa sélection commune par commune ; 3) utiliser une liste de communes prédéfinie à partir d'un fichier externe. Ensuite, il choisit la thématique ou directement les indicateurs qui l'intéressent parmi ceux proposés. Les tables de données étant déjà chargées dans le SGBDR, le serveur exécute des requêtes sur ces tables pour calculer les indicateurs demandés sur le territoire sélectionné. Ensuite, il envoie les résultats pour affichage sur l'interface web. L'utilisateur peut alors visualiser les résultats sous forme de tableaux de données et de graphiques, et débuter son analyse en lien avec son questionnement initial autour des systèmes et filières agricoles. Dans cet article, nous nous concentrons sur le fonctionnement et la maintenance du SGBDR (données et indicateurs).

1 Accès à SIDDT : https://siddt.inrae.fr

2 Accès à SI-BOAT : https://siddt.inrae.fr/siboat

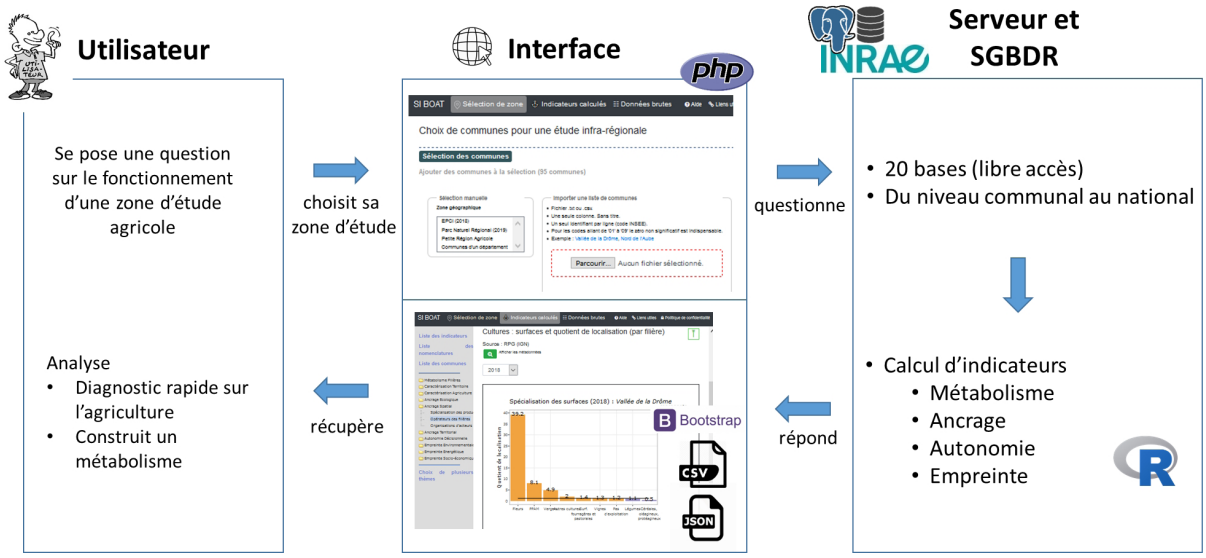

Figure 1. Chaine de traitement de l'information entre les 3 composantes de SI-BOAT

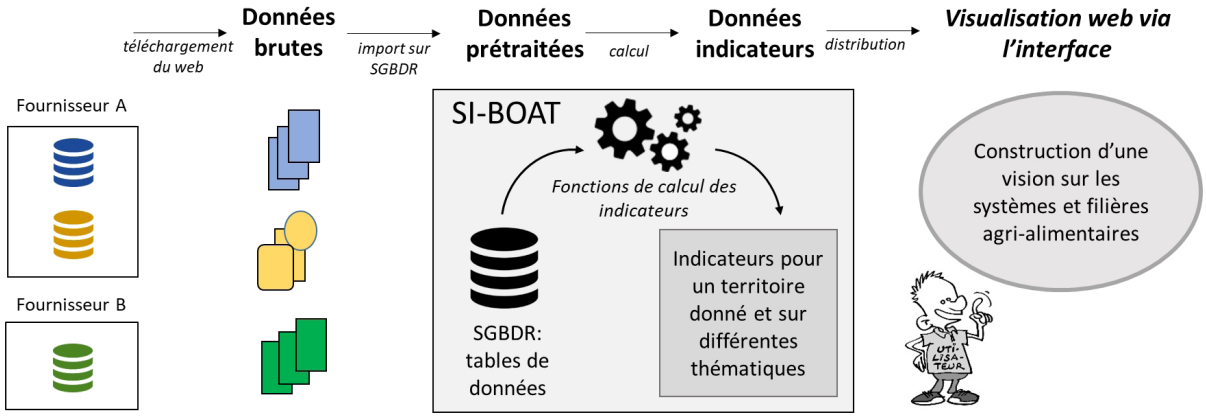

Figure 2. Flux des données

Le flux des données

Avant d'intégrer le SGBDR de SI-BOAT, les données sont produites et mises en forme par des fournisseurs. Nous les téléchargeons et obtenons alors ce que nous appelons les données brutes (Figure 2). Ces données sont importées sur le SGBDR et constituent alors les données prétraitées (sous forme de tables). SI-BOAT utilise ces données prétraitées pour calculer les données indicateurs relatives au territoire et aux thématiques choisis par l'utilisateur sur l'interface web. Ce sont ces données indicateurs que SI-BOAT compile et affiche via l'interface web. Nous détaillons ce flux de données dans les paragraphes suivants, ainsi que le système de métadonnées mis en place pour le suivi.

Hétérogénéité des bases de données mobilisées

Dans SI-BOAT, nous définissons une base de données comme un regroupement de données constitué avec un objectif défini et une méthode de collecte spécifique. Les bases de données mobilisées dans SI-BOAT doivent répondre à trois critères : informer sur les filières agricoles et systèmes agro-alimentaires ou sur leur contexte, être en accès libre et être disponibles pour l'ensemble du territoire français. En pratique, elles sont principalement constituées et diffusées par des organismes publics (ex. INSEE, Agreste). SI-BOAT mobilise actuellement 20 bases de données dont les principales caractéristiques sont montrées dans le tableau 1.

Les unités statistiques de collecte sont diverses et peuvent être plus fines que le niveau de diffusion des données. Par exemple, les données du recensement agricole sont issues d'enquêtes individuelles (l'unité statistique est la ferme) et leur diffusion libre d'accès est agrégée au niveau communal. Cela permet notamment la « secrétisation » pour les bases contenant des données personnelles, c'est-à-dire qu'il n'est pas possible d'identifier les individus à partir de ces données. Pour d'autres bases de données, notamment de type

Tableau 1 : Les 20 bases de données mobilisées dans SI-BOAT au 01/07/2024

Nom de la base	Diffuseur principal *	Unité statistique	Echelon géographique représentatif le plus fin	Fréquence de distribution	Nombre de tables structurellement differentes	Intervalle** des années représentées
Agriculture						
Agence Bio	Agence bio	établissement	commune	annuelle	6	[2010 ; 2022]
Certification à haute valeur environnementale	MAA	exploitation agricole	commune	annuelle	1	[2020 ; 2023]
DGAL	DGAL	commune	commune	annuelle	41	2023
INAO – Signes de qualité	INAO	polygone spécifique	entité spatiale	annuelle	2	[2019 ; 2023]
Recensement Général Agricole	CASD, Agreste	exploitation agricole	commune	décennale	6	[1988 ; 2020]
Registre Parcellaire Graphique	ASP, IGN (2015 et +)	ilot, parcelle	parcelle	annuelle	7	[2006 ; 2022]
Statistique Agricole Annuelle	Agreste	-	département	annuelle	23	[2000 ; 2023]
Environnement						
Espaces protégés	INPN	polygone spécifique	entité spatiale	annuelle	2	[2006 ; 2023]
ICPE	ICPE	commune	commune	annuelle	1	[2021 ; 2024]
ITDD	INSEE	commune	commune	annuelle	4	[1990 ; 2020]
NATURA2000	INPN	polygone spécifique	entité spatiale	ponctuelle	2	[2007 ; 2023]
Observatoire régional climat air énergie	ORCAE	exploitation agricole	commune	annuelle	19	[1990 ; 2023]
Registre des Emissions Polluantes	MTES	établissement			5	[2017 ; 2021]
SINOE® déchets	ADEME	entreprise	commune	annuelle	2	[2017 ; 2020]
Zone Vulnérable	Sandre - Eau France	polygone spécifique	entité spatiale	annuelle	1	2019
Socio économie						
Recensement général de la population	INSEE	foyer	commune	annuelle	3	[1990 ; 2021]
Sirene®: Système Informatique pour le Répertoire des ENtreprises et des Etablissements	INSEE	établissement	commune	quotidien	3	[2011 ; 2022]
Nomenclatures (support pour le système d'information)						
Code officiel géographique	INSEE	commune	commune	annuelle	8	[2010 ; 2023]
Nomenclature d'activités française	INSEE	activité économique	national	ponctuelle	1	2008
Projet BOAT***	INRAE	-	commune	ponctuelle	20	

(*) MAA - Ministère de l'Agriculture et de la souveraineté Alimentaire; INAO - Institut NAtional de l'Origine et de la qualité; CASD - Centre d'Accès Sécurisé aux Données; INPN - Inventaire National du Patrimoine Naturel; ORCAE - Observatoire Régional Climat Air Energie; MTES - Ministère de la Transition Ecologique et Solidaire; ADEME - Agence de la transition écologique; Sandre - Service d'administration nationale des données et référentiels sur l'eau.

(**) Intervalle continu ou discontinu selon les bases.

(***) Tables construites spécialement pour SI-BOAT, par ex. permettant de faire des appariements.

annuaire, les données individuelles peuvent être diffusées par les fournisseurs mais uniquement pour les entreprises qui l'ont accepté. Ces bases ne sont donc pas exhaustives.

Les échelons géographiques des bases incluses dans SI-BOAT vont de la parcelle à la région. Pour certaines bases, l'échelon géographique est spécifique à la base (ex. parc naturel régional, appellation d'origine géographique) et ne correspond pas à un découpage administratif « classique » (ex. commune, département).

Les bases mobilisées ont des fréquences de mise à jour et de distribution hétérogènes, allant de la distribution quotidienne à la distribution décennale. Dans le cas des tables mises à jour quotidiennement, SI-BOAT ne prend en compte qu'une mise à jour annuelle faisant l'état des lieux en début d'année civile.

Les données brutes

Une base de données peut contenir plusieurs informations (ex. pour le recensement agricole : surfaces agricoles, cheptel) et pour plusieurs années (millésimes), qui peuvent être diffusées dans un ou plusieurs fichiers. Les données brutes pour SI-BOAT sont les fichiers de données directement accessibles sur les sites des fournisseurs des bases de données. Nous n'avons pas mis en place de système reliant SI-BOAT aux différents fournisseurs et effectuons les téléchargements manuellement. Les données brutes sont téléchargées telles que fournies, sous forme de fichiers de type tabulaire (formats csv, txt, xlsx) et géographique (formats shapefile, GeoJSON).

Les données prétraitées

Identification des tables de données à charger

Une fois les données brutes téléchargées, nous identifions la structure des données à charger dans le SGBDR sous forme de tables de données. Le nombre de tables chargées est très variable d'une base de données à l'autre (Tableau 1). Il dépend notamment du nombre de fichiers fournis (ex. une table par variable et par année, versus une table regroupant l'ensemble des données et millésimes). Pour faciliter la maintenance à terme et être au plus près

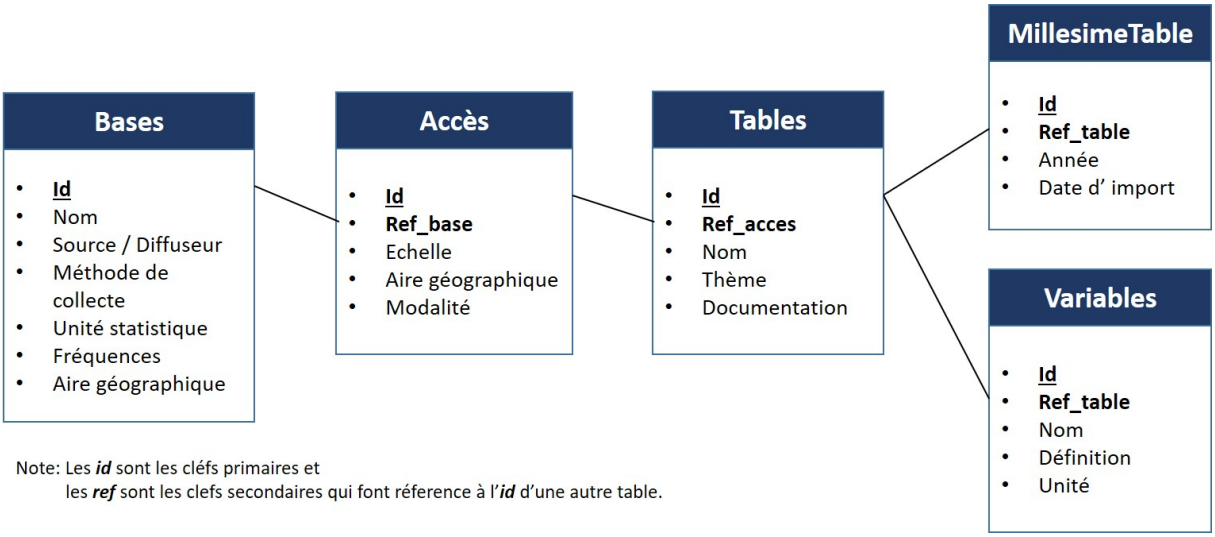

Figure 3. Métadonnées des bases de données (version simplifiée)

de la donnée brute, nous conservons généralement la structure en tables du fournisseur. Dans de rares cas où les fichiers contiennent une grande diversité d'informations, et pour réduire les temps de requête pour les indicateurs, nous les séparons en plusieurs tables. Les différents millésimes sont conservés pour permettre de travailler sur des évolutions.

Scripts de prétraitement et chargement sur le SGBDR

Les opérations de prétraitement visent à homogénéiser et structurer les tables provenant de sources diverses tout en modifiant le moins possible la donnée initiale. Ainsi, le prétraitement consiste en des modifications mineures : homogénéisation du format des noms des champs (ex. lettres minuscules, pas d'espace), explicitation des noms de champs (ex. remplacement de « LIBELLE » par « nom_departement »), création de clés pour faciliter l'appariement entre les tables. Une opération de prétraitement essentielle a par exemple consisté en l'homogénéisation des noms de champs présents dans plusieurs tables et contenant la même information. Ainsi, nous avons défini l'information du code commune INSEE avec le champ « depcom » dans toutes les tables où elle est présente. Nous avons utilisé ce champ comme clé principale permettant d'apparier les bases de données communales.

L'ensemble des opérations de prétraitement est réalisé et automatisé en utilisant des scripts rédigés dans le langage de programmation R[3] (logiciel sous licence libre). Ce choix d'écrire et de réutiliser des scripts vise à assurer autant que possible la pérennité du système d'information en facilitant sa maintenance et sa reproductibilité. Ces scripts peuvent cependant demander des adaptations en fonction des modifications effectuées par les fournisseurs de données (voir section 2.1 dans « Les différentes étapes de mise à jour des données »).

Métadonnées des bases et tables

Afin de documenter ces données prétraitées, nous avons mis en place un double système : 1) à visée des développeurs avec un commentaire directement sur chaque table du serveur. Il a une structure générique avec un bref descriptif des données de la table, le niveau de granularité, le fournisseur, l'adresse du site web de téléchargement, la date de mise à jour des données par le fournisseur et enfin la date d'import dans le SGBDR. 2) pour un suivi plus fin et détaillé permettant de remonter aux métadonnées des bases et des variables (ex. éléments de vigilance sur la méthodologie, relations entre les tables, etc.), nous avons constitué une base de métadonnées (Figure 3). Cette base est mobilisée sur l'interface web pour informer l'utilisateur sur les tables associées à l'indicateur qu'il consulte (ex. date de mise à jour, base de données source). Renseignée manuellement par les développeurs, cette base est constituée de 5 tables relationnelles :

- *Bases* : recense les informations relatives aux bases de données (1 ligne par base).
- *Accès* : recense les informations liées aux conditions d'accès des tables de données (1 à plusieurs lignes par base en fonction des modalités d'accès possibles). Une même donnée peut être diffusée à différents niveaux de confidentialité et de précision. Par exemple, les données du recensement agricole peuvent être « secrétisées au niveau communal et en accès libre » ou « détaillées par ferme en accès restreint ».

3. •R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

- *Tables* : recense les informations sur les tables de données. Le détail du millésime n'est pas noté ici. Ainsi, les tables ayant la même structure avec uniquement des modifications liées au millésime sont recensées en une seule ligne/table avec le code '*xxxx*' à la place de l'année. Ce sont ce que nous appelons les tables génériques. Par exemple, les tables cultures_2019, cultures_2020 et cultures_2021 sont regroupées sous le nom cultures_*xxxx*
- *MillesimeTable* pour les informations uniques à chacune des tables chargées dans le SGBDR, comme par exemple l'année, les date de mise à jour et d'import (1 ligne par table).
- *Variable* pour la liste des champs des tables de type "Tables" (1 ligne par champ par table générique).

Les indicateurs

L'indicateur dans SI-BOAT

Pour l'utilisateur, un indicateur est une information ciblée accessible sur l'interface web sous forme d'un tableau de données et/ou d'un graphe associé (ex. les différents types d'occupation du sol). L'indicateur est basé sur les données prétraitées du SGBDR. Ces dernières peuvent être filtrées sur la zone d'étude choisie puis diffusées directement telles quelles ou utilisées pour calculer d'autres éléments de diagnostic (ex. à partir des différents types d'occupation du sol, un quotient permettant d'identifier les spécialisations des productions agricoles sur la zone d'étude).

Fonctions pour calculer les indicateurs (PL/R)

Les indicateurs sont calculés à partir de fonctions en langage R directement utilisées par le SGBDR : des fonctions en PL/R (R Procedural Language). Une fonction peut servir à calculer plusieurs indicateurs. Ses données d'entrée sont des paramètres fournis par l'interface web. Les choix des données prétraitées du SGBDR, des calculs à effectuer et des données de sortie sont prédéfinis par les développeurs de SI-BOAT. Les paramètres peuvent être spécifiques à chaque fonction. Le paramètre commun à chaque fonction est celui de la vue temporaire délimitant la zone géographique d'étude. Cette vue, associée à la session d'utilisation de l'interface, est créée sur le SGBDR lors de la sélection du territoire d'étude (cf. section « Vue globale »). Elle contient notamment la liste des communes du territoire. L'année pour laquelle l'indicateur doit être calculé est un paramètre fréquent. Les formats des sorties générées sont de type JSON pour les futurs tableaux et HTML ou PNG pour les graphiques. Ces formats de sortie sont reconnus et mis en forme dans l'interface web ; ils n'ont pas d'impact sur l'expérience utilisateur, qui peut ensuite choisir un tout autre format de téléchargement des données tabulaires (ex. csv).

Métadonnées des indicateurs

Les indicateurs sont référencés dans une base de métadonnées constituée de six tables relationnelles (Figure 4) permettant d'assurer leur suivi, mise à jour et lien avec l'interface web. Ainsi, lors de la modification d'un indicateur, seule la saisie ou la modification de ses métadonnées est nécessaire pour qu'il soit ensuite pris en compte dans l'interface web. Deux tables servent à renseigner la grille d'analyse pour classer les indicateurs et faciliter la navigation de l'utilisateur dans l'interface web (« Thèmes » et « Sous-Thèmes »). Les quatre autres tables servent à décrire l'indicateur :

- *Indicateurs* : recense les métadonnées générales pour chaque indicateur calculable à partir des données prétraitées dans le SGBDR et des fonctions PL/R existantes. Cela concerne notamment les données d'entrée et formules utilisées ainsi que des points d'aide à l'analyse sur l'interprétation des résultats ou de vigilance (ex. quand la source de données est non exhaustive).
- *Lien indic-grille* : renseigne à quel thème et sous-thème de la grille est rattaché chaque indicateur (un indicateur pouvant être associé à plusieurs sous-thèmes de la grille).
- *Fonctions* : renseigne sur la fonction PL/R utilisée par le SGBDR pour calculer l'indicateur, comme par exemple le nom de la fonction et les tables qu'elle utilise.
- *Indic détail* : précise les paramètres d'entrée nécessaires (par exemple l'année) et des informations spécifiques comme l'aire géographique pour laquelle l'indicateur est calculé.

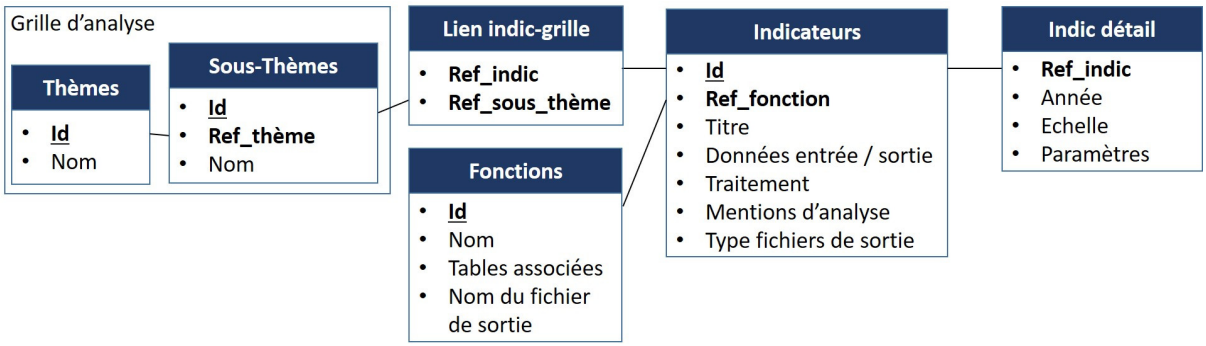

Figure 4. Métadonnées des indicateurs

Principaux packages utilisés avec R dans SI-BOAT

- Conway J, Eddelbuettel D, Nishiyama T, Prayaga SK, Tiffin N (2024). RPostgreSQL: R Interface to the 'PostgreSQL' Database System. R package version 0.7-6 https://CRAN.R-project.org/package=RPostgreSQL
- Wickham H, Hester J, Bryan J (2024). readr: Read Rectangular Text Data. R package version 2.1.5 https://CRAN.R-project.org/package=readr
- Wickham H (2023). stringr: Simple Consistent Wrappers for Common String Operations. R package version 1.5.1 https://CRAN.R-project.org/package=stringr
- Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4 https://CRAN.R-project.org/package=dplyr
- Wickham H, Vaughan D, Girlich M (2024). tidyr: Tidy Messy Data. R package version 1.3.1 https://CRAN.R-project.org/package=tidyr
- Pebesma E. & Bivand R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016
- Sievert C (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC, Florida.
- Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Technologies informatiques utilisées par SI-BOAT

Le SGBDR de SI-BOAT est hébergé sur un serveur Debian accessible uniquement en intranet par VPN (Virtual Private Network). Il est implémenté sur PostgreSQL avec PostGIS, et les langages R et PL/R, et leurs packages associés (cf. encadré). Ces outils sont mis à jour régulièrement par les administrateurs système.

Les fichiers de données brutes sont sauvegardés sur un serveur de données répliqué (Nextcloud). Les scripts R, PL/R et SQL sont versionnés et hébergés sur GitLab. Ils sont classés avec des conventions de nommage selon leur base de données (Agence Bio, Statistiques Agricoles Annuelles, etc.) et selon leur rôle (chargement de données, test, calcul d'indicateur).

Les différentes étapes de mise à jour des données

Les actions de mise à jour du SGBDR de SI-BOAT peuvent concerner chacune des étapes du flux de données, c'est-à-dire les données brutes, les données prétraitées, les données indicateurs et les données de l'interface web (Tableau 2). Différents cas peuvent être rencontrés : l'ajout d'une nouvelle base, d'une nouvelle table (ex. nouvelle année) ou l'actualisation d'une table existante (ex. les données des statistiques annuelles agricoles sont consolidées au cours des années suivant leur diffusion initiale). Dans cette partie, nous présentons et détaillons les différentes actions effectuées lors des mises à jour à chaque étape et chaque cas (Tableau 2).

Mise à jour des données brutes

Les bases les plus utilisées dans SI-BOAT et standardisées, comme le Registre Parcellaire Graphique (RPG) ou le Code Officiel Géographique (COG), sont téléchargées à des périodes fixes liées à leur date de sortie, peu variable d'une année à l'autre. Plus généralement, pour maintenir SI-BOAT à jour, un travail de veille est mis en place. Au moins une fois par an, l'ensemble des sites des fournisseurs de bases de données mobilisées est visité et parcouru. L'objectif de la veille est de vérifier si des mises à jour sont proposées concernant les données elles-mêmes, leurs métadonnées (ex. description, méthodologie) ou leur diffusion (ex. d'éventuels changements de licence). Les modifications liées aux métadonnées peuvent servir à mettre à jour les points de vigilance pour l'utilisation des indicateurs (ex. modification du protocole de collecte de données d'une année à l'autre, rendant la comparaison caduque). De nouvelles bases de données sont parfois ajoutées au SGBDR pour compléter la liste d'indicateurs alimentant la grille d'analyse en fonction des besoins soulevés par les utilisateurs de SI-BOAT.

Mise à jour des données prétraitées

Gestion des scripts de prétraitement

Les scripts R permettent l'import des données brutes sur le SGBDR et leur prétraitement. Lors de l'ajout d'une nouvelle base de données, ces scripts doivent être créés. Pour la majorité des ajouts de nouvelles tables ou modifications, les scripts sont réutilisables tels quels. Seules la date de mise à jour de la donnée par le fournisseur et le nom de

Tableau 2 : Tableau d'actions à réaliser par les développeurs en fonction des cas de mises à jour

Données	Action	Action détaillée	Nouvelle base	Nouvelle table	Table existante
Données brutes	Télécharger les données brutes sur l'ordinateur	Télécharger depuis le site du fournisseur	x	x	x
Données prétraitées	Importer les données sur le SGBDR	Créer le script	x		
		Mettre à jour le script		x	x
		Importer sur le schéma de test	x	x	x
	Vérifier la structure	Comparer données brutes et importées	x	x	x
	Vérifier la cohérence de la structure et des modalités avec les tables existantes	Comparer table importée et années précédentes		x	x
		Comparer table importée et table remplacée			x
	Importer les données sur le SGBDR	Importer sur le schéma de production	x	x	x
	Gérer les dépendances	Créer/mettre à jour une vue, communaliser les données, etc.	x	x	x
	Renseigner les métadonnées de la table et de ses dépendances	*Autres métadonnées plus stables*	x	~	~
		MillésimeTable	x	x	x
Données indicateurs	Intégrer les indicateurs	Créer le script	x		
		Mettre à jour le script		~	~
	Renseigner les métadonnées des indicateurs	*Lien indic-grille*	x		
		Autres métadonnées plus stables	x	~	~
		Indic détail	x	x	x
Données interface	Vérifier les résultats	Tester les indicateurs sur l'interface	x	x	x

	non concerné
x	concerné
~	peu concerné

la table sont à changer manuellement (ex. quand l'année est dans le nom de la table). Dans certains cas, les scripts doivent être mis à jour plus en profondeur. Par exemple, les formats de fichiers bruts peuvent changer (ex. passage de .xls à .ods), les noms de champs peuvent changer, voire certains peuvent être ajoutés ou supprimés. Un arbitrage par les développeurs et thématiciens du projet peut s'avérer nécessaire pour choisir entre un passage à la nouvelle structure des tables ou le maintien de l'ancienne.

Tests de vérification de l'intégrité des données hors production

Avant un import définitif dans le SGBDR, les données sont pré-importées dans une structure dédiée à des tests de vérification. Ces tests permettent de s'assurer de la cohérence des données et de vérifier le fonctionnement des scripts d'import. Ils permettent également de décider si la donnée sera conservée, utilisée, diffusée avec un point de vigilance ou rejetée.

Pour chaque table importée dans le SGBDR, nous vérifions la cohérence entre la structure de la table importée et celle de son fichier de données brutes associé : homogénéité des noms de champs, des types de données (numérique, texte, etc.). Dans le cas d'ajout d'une nouvelle table, nous testons les possibles changements des modalités des variables qualitatives par rapport aux tables précédentes.

Une attention particulière est portée sur les données géographiques. Comme pour les autres données, l'objectif est de modifier le moins possible la donnée originale. Ainsi, le système de projection des données originales est conservé sans transformation et renseigné comme métadonnée. Les vérifications consistent à s'assurer de la cohérence des géométries des données (ex. erreurs liées à des polygones mal formés) et de la localisation (ex. points localisés hors des polygones des communes auxquelles ils sont censés appartenir d'après la table attributaire).

Pour les modifications de tables existantes, nous effectuons une analyse des différences entre l'ancienne et la nouvelle version de la table. Nous avons automatisé ce test avec une fonction R (utilisée en externe du SGBDR) qui génère un rapport détaillé des différences. Ce rapport affiche d'une part les zones géographiques et les modalités des variables qualitatives qui ont été ajoutées ou supprimées, et d'autre part les lignes ayant au moins une valeur numérique qui a changé pour un même couple zone/modalité. Pour ces lignes, le rapport affiche les anciennes valeurs, les nouvelles, et les différences entre les deux. Ces vérifications nous permettent d'avertir nos utilisateurs qui auraient utilisé les données avant la mise à jour de modifications pouvant impacter leur analyse.

Mise à jour des dépendances (tables et vues)

Comme les données prétraitées sont peu modifiées par rapport aux données brutes, nous créons parfois des tables et vues pour faciliter le calcul des indicateurs. Pour diminuer les temps de calcul à venir, des tables avec un grand nombre d'entrées individuelles peuvent être communalisées. Par exemple, les surfaces des parcelles du Registre Parcellaire Graphique sont agrégées au niveau communal ou encore le détail des établissements de SIRENE® est regroupé en un compte d'établissements par commune pour les différents codes d'activités.

Pour faciliter la manipulation des tables relationnelles, des vues peuvent également être créées. Ce sont des requêtes pré-enregistrées dans le SGBDR qui, le plus souvent, nous servent à joindre les tables entre elles (ex. association des codes cultures des rendements à leur libellé) ou à les transformer (ex. passage des coordonnées GPS à une donnée spatialisée, transformation du système de projection). Ces dépendances (tables et vues) sont donc aussi à créer et/ou mettre à jour.

Import des données prétraitées en production

Les scripts R d'importation des données prétraitées et de mise à jour des dépendances sont compatibles pour l'import en production. Ainsi, lorsque les vérifications des données sont terminées et que leur import est validé, nous importons les données en production à l'aide de ces scripts.

Mise à jour des métadonnées

Lors de l'ajout d'une nouvelle base de données, l'ensemble des champs des cinq tables de métadonnées sur les bases et tables doit être renseigné (Figure 4). Dans le cas de l'ajout d'une nouvelle table ou d'une modification de table existante, seule la table « MillesimeTable » doit être mise à jour. Le lien web d'accès direct au téléchargement de la donnée brute peut aussi nécessiter une mise à jour. Dans le cas moins fréquent où la documentation ou les champs de la table ont été modifiés, les tables « Tables » et « Variables » respectivement sont également à mettre à jour.

En fonction des modifications apportées sur les nouvelles tables ou tables existantes, nous pouvons également compléter les éléments relatifs à la documentation de la table.

Mise à jour des indicateurs et des données spécifiques pour l'interface web

Mise à jour des indicateurs

Les tables de données existantes sont mobilisées pour les calculs des indicateurs via les fonctions PL/R. Ces fonctions sont créées de sorte à pouvoir intégrer de nouvelles tables annuelles en modifiant uniquement les paramètres d'entrée. Sur l'interface web, l'utilisateur a accès aux années renseignées dans les métadonnées des indicateurs, qui servent de paramètres d'entrée (Figure 5, table « Indic détail »). Ainsi, pour l'ajout de nouvelles tables année ou pour la mise à jour d'une table existante, si la structure (ex. noms ou types des champs) n'a pas changé, seule une modification de la table « Indic détail » suffira à ajouter une nouvelle année pour l'indicateur ou à le mettre à jour sur l'interface.

Lors de l'ajout d'une nouvelle base de données, nous créons un ou plusieurs indicateurs ainsi que la ou les fonctions PL/R qui leur seront associées. L'ensemble des métadonnées associées doit être renseigné dans les métadonnées des indicateurs (lien à la grille d'analyse, détails pour l'analyse, fonction et paramètres).

Lorsque des indicateurs ont été modifiés ou ajoutés, nous vérifions manuellement sur l'interface web leur bonne prise en compte et fonctionnement (affichage, résultats, etc.).

Mise à jour de données spécifiques pour l'interface web

Nous n'entrerons pas dans le détail du fonctionnement de l'interface web de SI-BOAT. Cependant, comme mentionné dans la section « 2.1 SI-BOAT : vue globale », pour permettre à l'utilisateur de définir sa zone d'étude, des zonages existants lui sont proposés. Or, ces zonages administratifs évoluent régulièrement (ex. fusions de communes, ajout/suppression de communes dans les parcs naturels régionaux, etc.). Les données associées à ces zonages sont mises à jour annuellement pour garantir aux utilisateurs de SI-BOAT de travailler sur des délimitations de zones d'études telles qu'elles existent au moment de l'analyse.

Retours sur les éléments clés de notre démarche

Apport à la démarche scientifique

SI-BOAT est conçu pour faciliter la préparation des données et permettre à l'utilisateur de consacrer davantage de temps à l'analyse approfondie. Les objectifs de transparence et de reproductibilité sont centraux dans notre démarche. Nous souhaitons que les utilisateurs s'approprient les données et soient sensibilisés aux incertitudes associées à leur manipulation. Dans le développement et la maintenance de SI-BOAT, un travail conséquent est engagé pour renseigner les métadonnées des bases de données et des indicateurs. Les formules de calcul, les données d'entrée, ainsi que les éléments de vigilance et d'aide à l'analyse sont fournis afin de diminuer l'effet de « boîte noire » des bases de données. Les sources sont systématiquement citées. En homogénéisant leur présen-

tation dans une même trame pour chaque base et indicateur, nous souhaitons permettre à l'utilisateur d'accéder rapidement aux points de vigilance associés aux données mobilisées, afin de prendre du recul.

Délai des mises à jours et informations de l'utilisateur

Nous considérons que la mise à jour de SI-BOAT est cruciale pour assurer une bonne expérience utilisateur, des données obsolètes mettant potentiellement son travail en péril. Dans le même souci de transparence que pour les métadonnées des bases de données mobilisées et les indicateurs calculés, nous mettons en place un système qui permettra aux utilisateurs de suivre les mises à jour effectuées dans SI-BOAT. Côté développement, il s'agit d'une table recensant les modifications, et côté utilisateur, d'un affichage dans l'interface web.

Par ailleurs, malgré ce travail de maintenance de l'outil, il est illusoire de penser que les données mobilisées seront toujours à jour. Personne ne travaillant à plein temps sur ce projet, le temps associé à la mise à jour dépend également de nos disponibilités. En raison du choix de ne visiter les sites des fournisseurs de données qu'une fois par an, il peut y avoir un décalage. Ce décalage est parfois à notre avantage, car il laisse au fournisseur de données le temps de corriger certaines erreurs et de consolider ses données. Avec le temps, la veille régulière permet de réduire les incertitudes sur les tables de données les plus anciennes.

Nous avons travaillé sur les conditions générales d'utilisation de SI-BOAT pour informer les utilisateurs que nous ne garantissons pas que les données soient toujours parfaitement à jour, et que leur réutilisation est sous leur responsabilité.

Une automatisation partielle des procédures, un besoin de temps humain incontournable

Nous avons cherché à automatiser autant que possible les différentes étapes de mise à jour de SI-BOAT jusqu'à leur intégration dans l'interface web. Cependant, certaines étapes manuelles restent nécessaires. D'une part, la veille sur les mises à jour et les modifications des fournisseurs est incontournable et doit pouvoir se poursuivre. Les données sont importantes, mais les informations associées, comme la licence d'utilisation des données, le sont également. Il nous est arrivé de voir des licences évoluer vers plus de liberté dans la réutilisation des données, ou à l'inverse, vers davantage de restrictions. Nous veillons à suivre ces évolutions pour en informer les utilisateurs, voire prendre des mesures si nécessaire (ex. suppression de l'utilisation des données).

D'autre part, du temps de développement informatique reste nécessaire pour les cas particuliers de mise à jour (ajout d'une nouvelle base de données, modification des protocoles de téléchargement, évolution des données brutes ou des librairies permettant de les manipuler). De plus, les tests de vérification sont essentiels et demandent également du temps pour assurer l'intégrité des données de SI-BOAT.

Compétences interdisciplinaires

Maintenir un système d'information tel que SI-BOAT requiert de multiples compétences. Des connaissances thématiques sont nécessaires pour sélectionner les bases de données et indicateurs à intégrer. Des compétences en développement informatique sont ensuite indispensables pour assurer le fonctionnement et la maintenance du SGBDR, suivant les étapes décrites dans cet article. Une coopération entre thématiciens et informaticiens est essentielle pour effectuer des arbitrages en cas de modification des données (ex. lors d'un changement de modalités pour une typologie), ou pour évaluer l'importance des changements dans les données (ex. dans le cas d'une mise à jour d'une table existante). Enfin, des compétences en administration de systèmes sont requises pour les mises à jour et la maintenance des serveurs de base de données PostgreSQL et web. Cette diversité de compétences doit être prise en compte, et constitue selon nous un argument en faveur de la création d'une dynamique collective, impliquée pour entretenir les échanges interdisciplinaires.

Conclusion

Depuis sa création en 2019, SI-BOAT contribue à faciliter les travaux de diagnostic des systèmes et filières agricoles, en mobilisant 20 bases de données libres d'accès. Dans cet article, nous avons présenté la démarche que nous avons mise en place pour maintenir et faire évoluer ce regroupement de bases, elles-mêmes en constante évolution. Nous avons également montré l'importance de la structuration des métadonnées des bases de données mobilisées et des indicateurs. Grâce à cette organisation, nous facilitons à la fois le travail de mise en ligne des indicateurs et l'appro-

priation d'une diversité de données et d'indicateurs par des utilisateurs non spécialistes.

Nous tenons à souligner que la construction d'un tel système d'information n'est pas anodine, et que sa maintenance exige une organisation, du temps et une coopération ; trois éléments qui doivent être anticipés par les concepteurs. Pour assurer la pérennité du système, nous insistons sur l'importance de disposer de personnes ayant du temps dédié à sa gestion et sa maintenance. Comme nous l'avons montré, étant donné la diversité des sources de données, une expertise humaine (technique et thématique) est indispensable pour certaines mises à jour, les tests, et les arbitrages nécessaires.

L'étude des besoins est également cruciale. À ce jour, à part un questionnaire de retours proposé sur l'interface web, nous n'avons pas mis en place de système systématique de recueil de retour d'expérience. Une réflexion sur cette question permettrait d'identifier les bases de données prioritaires pour les mises à jour, ou les données et indicateurs manquants pour traiter des filières agricoles et systèmes agro-alimentaires. ■

Remerciements et financement

SI-BOAT a été créé grâce aux financements du projet BOAT, appel à projets Graine de 2016 de l'ADEME. Il est maintenu par les unités de recherche AGIR (unité mixte de recherche INRAE-INPT) et LESSEM (unité de recherche INRAE). De nouveaux indicateurs ont été développés dans le cadre du projet Scalable, financé par l'appel à projets Graine 2020 de l'ADEME. Les auteurs remercient leurs collègues du LESSEM : Frédéric Bray et André Torre pour leur soutien sans faille dans le développement de SI-BOAT, Sophie Madelrieux pour sa prise de recul et apports thématiques, Sylvain Duchêne pour ses développements pour l'interface web, Eric Maldonado pour son appui système. Nous remercions également Julien Quénon pour sa relecture.

Glossaire

Données brutes

Par données brutes, nous entendons des données que nous avons téléchargées auprès des fournisseurs de bases de données.

Données prétraitées

Par données prétraitées, nous entendons les données générées à partir des données brutes et qui sont stockées sur le SGBDR. De légères modifications peuvent avoir été apportées aux données brutes, notamment sur le nom des champs. Ces données sont majoritairement utilisées pour le calcul des indicateurs. Certaines ne sont pas encore utilisées.

Données indicateurs : données traitées et diffusées

Par indicateurs, nous entendons des données préparées et mises en forme pour l'utilisateur qui sont diffusées via l'interface web. Ces données sont générées à partir des données prétraitées et par l'intermédiaire de fonctions permettant de les mettre en forme, voire de les compléter avec de simples calculs (ex. proportions des types de culture, etc.). Ils sont calculés à partir de fonctions PL/R.