DATA4C+

Vers l'interopérabilité des bases de données pour être plus FAIR

Analyse des verrous informatiques à l'interopérabilité entre bases de données du Cirad, d'INRAE et de l'IRD sur le carbone du sol et les modes de gestion des sols et propositions de solutions

2022

Jean-Baptiste Laurent, Kenji Fujisaki, François Thévenin, Rachid Yahiaoui, Clément Lattelais, Antoine Schellenberger, Christine Le Bas, Etienne Lamy
Julien Demenois

N° ANR-19-DATA-0005-01

Rapport T2.1, T2.2, T2.3 et T3.2

1. État des lieux des bases de données informatiques et des technologies actuellement en cours au sein du Cirad, d'INRAE et de l'IRD sur les bases de données sur le carbone du sol (T2.1)

Les données sur le carbone des sols du Cirad, de l'INRAE et de l'IRD sont stockées dans 4 bases de données, réparties sur 4 serveurs informatiques dans des lieux différents. Les formats des données et leur structuration (métadonnées) sont différents pour chaque organisme. Le tableau ci-dessous présente la volumétrie et la complexité de structuration des données des trois organismes.

Institut	Nom de la base de données	Nombre de tables	Système	Accessibles de l'extérieur
Cirad	Lims labo d'analyse	39	Oracle	Oui
Cirad	Lims (données archivées)	18	Oracle	Oui
INRAE	Donesol	40	PostgreSQL	Non
IRD	Valsol	40	PostgreSQL	Non
IRD	Ithèque (*)	7	PostgreSQL	Non
IRD	2Carma(*)	11	PostgreSQL	Non

(*)Bases de données non retenues dans le cadre du développement du prototype du projet ANR DATA4C+

En complément, un travail de description systématique de toutes les informations stockées dans ces bases de données a été effectué dans l'OS1. Ce travail était indispensable pour concevoir des solutions d'interopérabilité.

2. Les verrous informatiques à l'interopérabilité (T 2.2)

Plusieurs verrous informatiques à l'interopérabilité ont été identifiés. Tout d'abord, chaque institut gère ses données sur le carbone des sols via des systèmes « maison ». Par conséquent, les bases de données ne sont pas nécessairement accessibles de l'extérieur des instituts (barrières technologiques de sécurité (Firewall)). De plus, les structures de bases de données sont très différentes et structurées différemment selon leur origine.

Par ailleurs, la description de ces structures de données (métadonnées) est hétérogène. En l'état, il est très difficile de croiser les données issues des trois organismes. Pour lever ces verrous informatiques, plusieurs solutions ont été explorées dans la T2.3.

3. Solutions informatiques permettant l'interopérabilité entre les bases de données sur le carbone du sol (T2.3)

Deux solutions d'interopérabilité ont été élaborées et testées dans le cadre du projet ANR DATA4C+ : le datamart (entrepôt de données) et l'API Sensor Things. Ces 2 solutions devaient permettre de franchir les barrières technologiques de sécurité d'une part, et d'autre part de s'affranchir de la

complexité des systèmes de gestion de bases de données (SGBD). L'expérience d'INRAE avec l'API Sensor Things a été déterminante dans le choix de cette solution d'interopérabilité entre les bases de données du Cirad, d'INRAE et de l'IRD. Elle peut être vue comme l'étape suivante à la mise en place d'un datamart.

Pour les 2 solutions, l'accès aux données depuis l'extérieur des instituts se fait via un système de portail Internet, avec identifiant et mot de passe. Il est donc restreint et sécurisé.

a. La solution d'un Datamart ou entrepôt de données

Un entrepôt de données (data warehouse en anglais) est une base de données qui regroupe les données fonctionnelles d'une organisation. Il est alimenté par les bases de données opérationnelles qui sont utilisées pour gérer et stocker les données afin de faciliter l'utilisation et la visualisation de ces données sans obérer les performances des bases de données opérationnelles. Un magasin de données (datamart en anglais) s'appuie sur l'entrepôt de données pour fournir des données répondant à une utilisation particulière. Un entrepôt de données peut ainsi contenir plusieurs datamarts pour répondre à différents types d'utilisations des données.

Le Cirad et l'IRD ont choisi d'alimenter un Datamart pour rassembler les données sources. Une fois que toutes les données des applications sont rassemblées sur une seule plateforme, elles pourront être utilisées dans des outils d'analyse pour identifier les tendances ou aider à la prise de décision. L'utilisation d'un Datamart doit viser les objectifs suivants :

- Rendre les données des institutions facilement accessibles.

  Le contenu du datamart doit être facile à comprendre. Les données doivent être parlantes et leur signification évidente pour l'utilisateur et pas seulement pour le développeur informatique ou l'expert.
  Le datamart entrepôt de données doit présenter l'information de l'organisation de manière cohérente et homogène.
- L'entrepôt de données doit être adaptable et résistant aux changements.
  Les données de l'entrepôt devront être conçues pour traiter les changements et l'évolution des bases de données source.
- Permettre l'historisation des informations et l'accès simple, rapide et intuitif.

La solution Datamart nécessite donc une remise à plat de la structure des bases de données et leur description précise. Cette étape de description a été réalisée dans le cadre du projet ANR DATA4C+ pour chacune des bases de données identifiées du Cirad, d'INRAE et de l'IRD. Elle est indispensable pour gérer les verrous informatiques liées aux structures des bases de données et l'éventuelle hétérogénéité des données. La solution Datamart peut constituer une étape préalable à la solution API Sensor Things.

L'alimentation du Datamart pour les données du Cirad.

Les données décrivant une analyse de sol sont réparties dans plusieurs tables. Certaines tables sont de grandes dimensions. Il est donc lourd de réaliser des jointures entre les tables chaque fois qu'on veut explorer des données dans la base de données. Il est par conséquent nécessaire de créer un seul tableau (ou table) à grande dimension contenant toutes les informations nécessaires à la description d'un échantillon de sol.

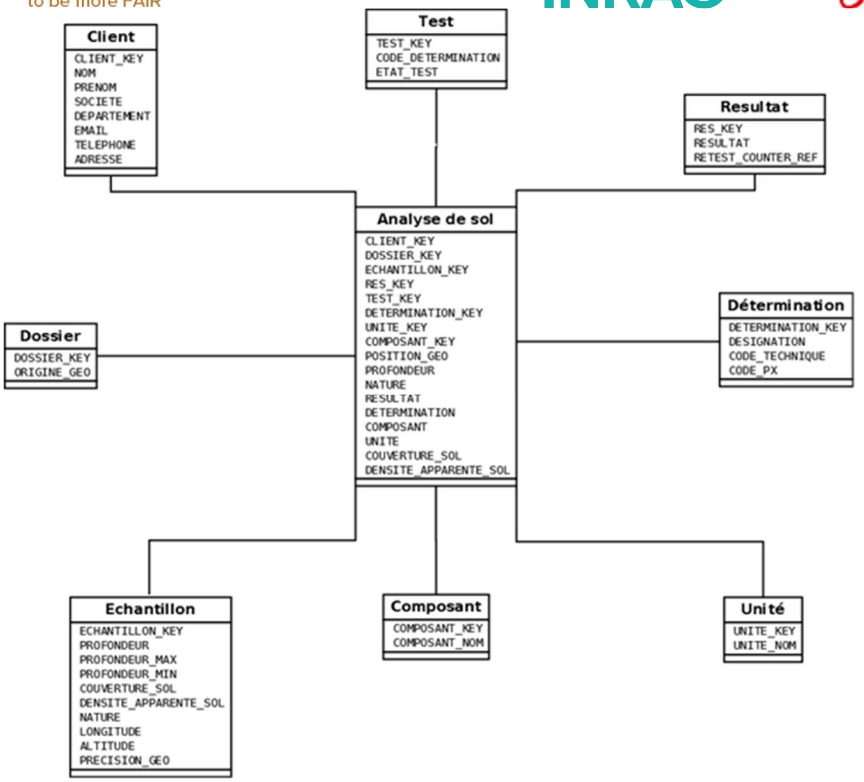


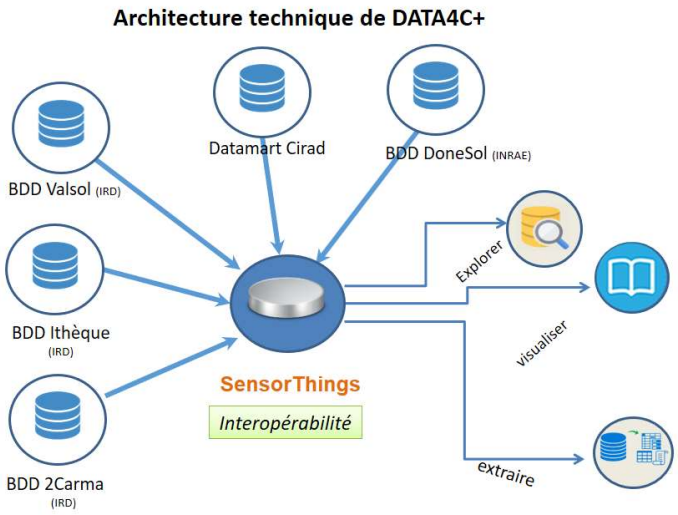
Schéma en étoile du datamart du Cirad

La description du Datamart du Cirad est présentée en annexe.

Alimentation du Datamart avec les données de l'IRD.

Un travail similaire d'alimentation d'un datamart avec les données issues de la base de données Valsol est en cours.

Les Datamarts des données du Cirad et de l'IRD sont ensuite connectés aux API SensorThings pour alimenter le prototype d'accès aux données (voir les § 3 et 4).



b. La solution API Sensor Things

INRAE a opté pour l'utilisation des API Sensor Things pour assurer l'interopérabilité avec la plateforme DATA4C+. En effet, cette solution a déjà été mise en place dans le cadre d'autres projets, notamment avec le BRGM (projet FGU III financé par l'Ademe).

Le standard OGC Sensor Things API permet l'échange de données entre serveurs distants. Ce standard se compose de deux parties: « Sensing » et « Tasking ». Voir la description générale des API sur http://docs.opengeospatial.org/is/15-078r6/15-078r6.html

- Sa partie « Sensing », est destinée à la détection et à la collecte d'observations à partir d'appareils de détection. Elle utilise des concepts établis à long terme dans la description des données de mesures issues de capteurs (Capteurs, Objets monitorés, Méthode de monitoring, Flux de données et Observations), tout en tenant compte des exigences modernes pour des interfaces efficaces.
- Sa partie « Tasking » (OGC 17-079r1) se concentre sur le contrôle des appareils connectés et sort du cadre des travaux du projet ANR DATA4C+.

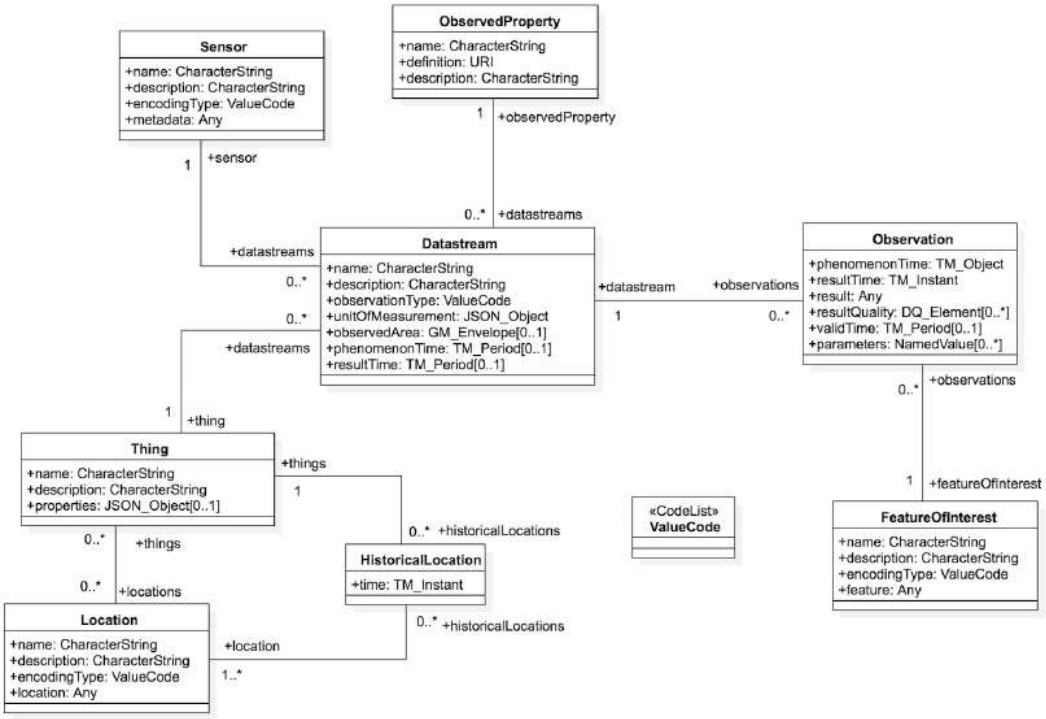


Schéma relationnel des objets de l'API Sensor Things

Différents outils ont été développés par l'Unité InfoSol d'INRAE pour permettre :

- le mapping de la base de données relationnelle DoneSol dans le modèle de données Sensor Things : outil Sensor Map. Ce travail repose sur une bonne connaissance de la structure de la base de données relationnelle DoneSol permettant avec des requêtes SQL d'alimenter les différents templates correspondants à chaque objet Sensor Things (Thing, Datastream, Sensor, etc.) incluant leur relation
- la visualisation des données selon le modèle Sensor Things. L'API Sensor Things fournit les données au format JSON. Des outils peuvent ensuite être utilisés pour, à partir du format JSON,

afficher les données de manière plus ergonomique et en permettre l'exploration (outils Sensor Board et Map Go). Le fichier JSON peut également être envoyé sur l'interface développée par le CIRAD pour DATA4C+.

4. Solutions de mise en œuvre pour l'interopérabilité (T3.2)

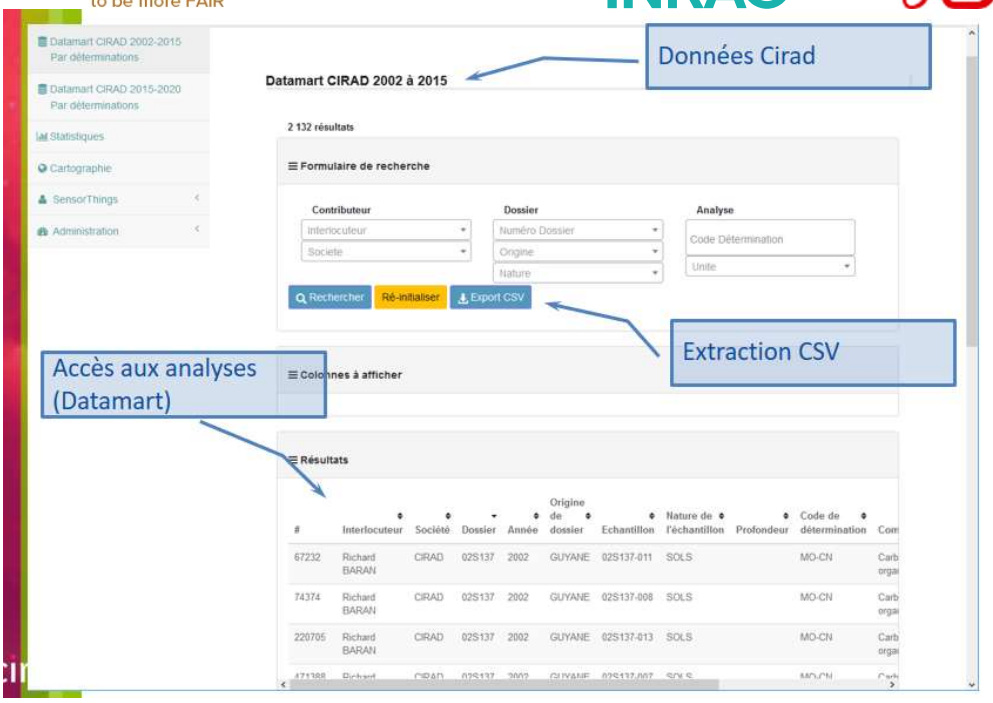
Les 2 solutions d'interopérabilité décrites précédemment ont été mises en œuvre dans une preuve de concept (T3.2.).

a. Le portail d'accès aux données

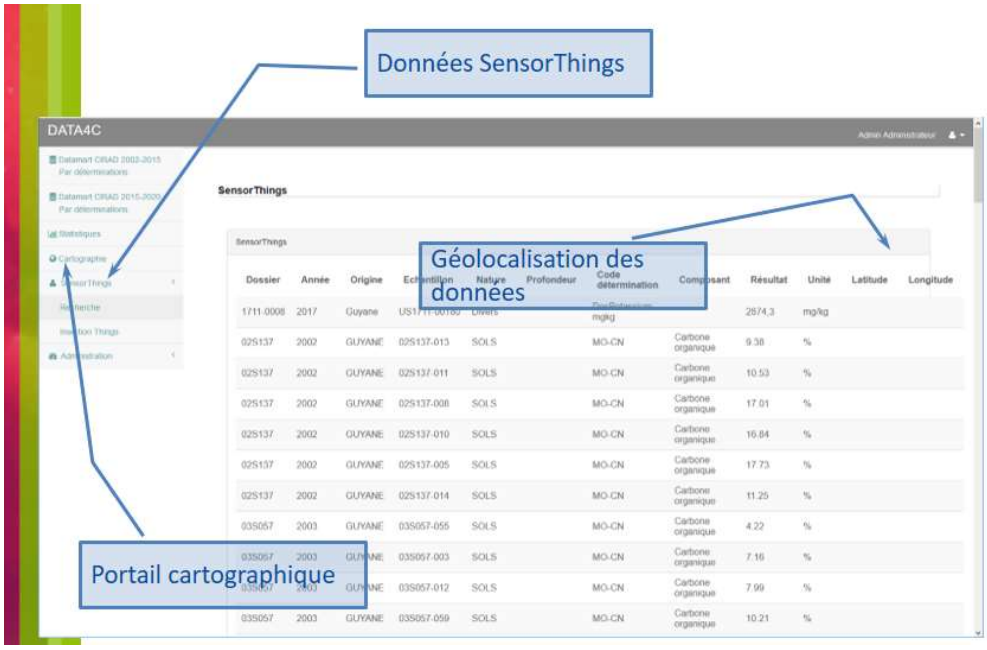
Nous avons développé une interface pour que les utilisateurs puissent explorer des données, ajouter des informations importantes telles que les coordonnées géographiques, la profondeur, la couverture de sol, …

Ce prototype d'accès aux données DATA4C+ permet aux chercheurs d'accéder (après authentification) aux données de Guyane des trois institutions de recherche (Cirad, INRAE, IRD). L'utilisateur peut filtrer, trier et télécharger des données sous forme d'un fichier CSV pour les analyser ou alimenter des modèles.

Page d'accueil du prototype DATA4C+



Accès aux données par institution (Guyane)



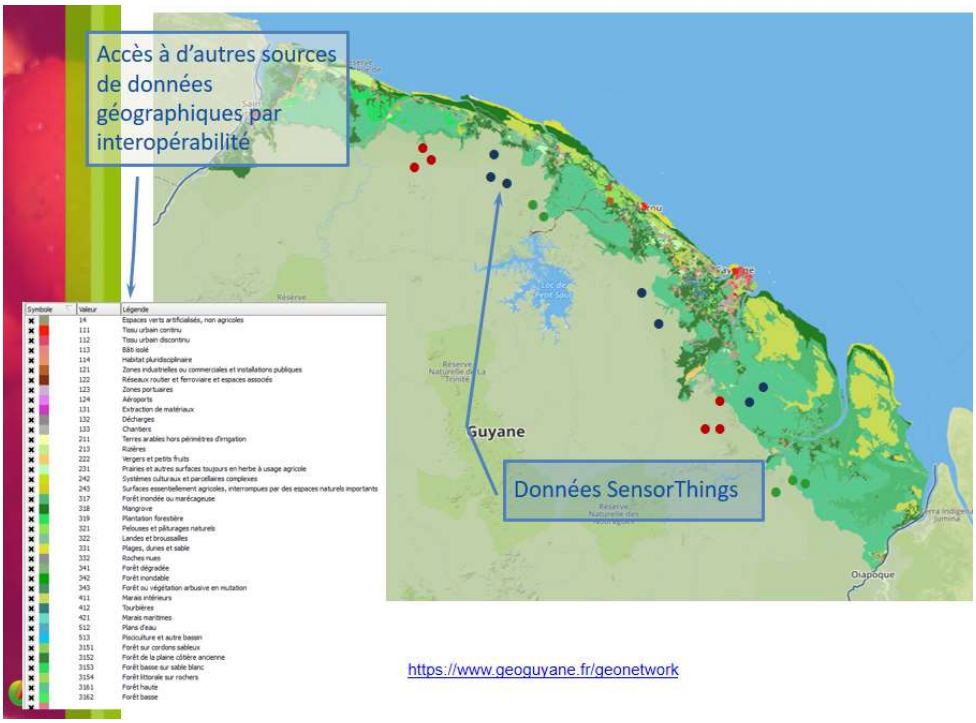
Accès aux données par institution (Guyane)

Logiciel/outil	Description	Version	Licence	observation
PostGis	Gestionnaire de base de données	2.4.1	PostgreSQL License	Licence logiciel libre proche de GNU-GPL General Public License
Lizmap	Diffusion de cartes en ligne et gestionnaire WMS	3.1	Mozilla Public License 2.0	Licence logiciel libre compatible GPL.
Symphony	Framework php	4.4	MIT Licence	Licence compatible avec la licence GNU-GPL General Public License
Leaflet	Interface cartographique	1.7.1	BSD License	Les licences BSD sont une famille de licences de logiciels libres permissive (adaptation et évolution autorisée avec clause de non-responsabilité)
API SensorThings	Interopérabilité entre objets à composantes spatiales	1.0	MIT License	

L'ensemble des programmes sources du prototype DATA4C+ sont diffusés sur la forge https://gitlab-ecosols.cirad.fr et accessibles sur demande sous licence CC BY-SA 2.0 FR.

b. Interopérabilité avec d'autres sources de données spatiales

Le couplage du portail DATA4C+ avec un serveur cartographique QGis Server permet, via les API Leaflet, d'afficher la localisation des échantillons de carbone avec des cartes thématiques de Guyane.



Représentation des points issus des différentes sources de données

D'autres sources de données de Guyane sont également disponibles. Elles sont stockées et diffusées au format géographique sur des serveurs cartographiques. Ainsi, le portail GéoGuyane (www.geoguyane.fr), mis en place par l'Agence d'Urbanisme et de Développement de la Guyane (AUDeG), diffuse un nombre important de données géographiques. Le portail DATA4C+ respecte les normes de l'Open Geospatial Consortium d'échange de données par interopérabilité. Il est donc possible d'afficher sur la même carte des données issues de serveurs tiers comme GéoGuyane.

Annexe - Description du Datamart du Cirad

column_name	data_type	description
cl_key	integer	clé primaire interlocteur (plus d'utilité)
cl_interlocuteur	character varying	nom interlocuteur
cl_societe	character varying	société interlocuteur
cl_departement	character varying	département interlocuteur
cl_service	character varying	service interlocuteur
cl_adresse_1	character varying	adresse interlocuteur
cl_adresse_2	character varying	adresse interlocuteur
cl_adresse_3	character varying	adresse interlocuteur
cl_code_postal	character varying	code postal interlocuteur
cl_ville	character varying	ville interlocuteur
cl_pays	character varying	pays interlocuteur
cl_telephone	character varying	Téléphone interlocuteur
cl_fax	character varying	fax interlocuteur
cl_email	character varying	emaul interlocuteur
do_key	integer	dclé primaire dossier
do_dossier	character varying	numéro de dossier
do_origine_geo	character varying	origine géographique dossier
do_latitude	double precision	coordonnées géo dossier
do_longitude	double precision	coordonnées géo dossier
do_nature_echant	character varying	nature de l'échantillon au niveau dossier
do_nb_echant_recu	integer	nombre d'echantillons reçus dossiers
do_clkey_ref	integer	clé primaire
do_date_demande	timestamp without time zone	date demande dossier
do_precision_geo	character varying	précision coordonnées géo dossier
s_key	integer	clé primaire échantillon
s_echantillon	character varying	numéro échantillon
s_nature	character varying	nature échantillon
s_profondeur	character varying	profondeur échantillon
s_sample_disposition	character varying	disposition échantillon
s_sample_state	character varying	etat de l'échantillon
s_dokey_ref	integer	clé primaire
s_couverture_sol	character varying	couverture du sol échantillon
s_densite_apparente_sol	double precision	densité apparente
s_profondeur_min	double precision	profondeur minimum
s_profondeur_max	double precision	profondeur maximum
s_latitude	double precision	coordonnées géo échantillon
s_longitude	double precision	coordonnées géo échantillon
s_precision_geo	character varying	précision geo
rt_id	integer	clé primaire
rt_key	integer	clé primaire

rt_code_determination	character varying	code détermination
rt_test_disposition	character varying	disposition du test (PASS)
rt_test_validation_disp	character varying	validation du test
rt_test_state	character varying	état du test
rt_pxkey_ref	integer	clé primaire
rt_urkey_ref	character varying	clé primaire
rt_tgtkey_ref	integer	clé primaire
rt_skey_ref	integer	clé primaire
px_key	integer	clé primaire
px_code_analyse	character varying	code analyse prélèvement
px_code_technique	character varying	code technique prélèvement
px_test_code	character varying	code test prélèvement
px_code_px	character varying	
px_designation	character varying	désignation prélèvement
res_key	integer	clé primaire
res_assay_result	character varying	résultat
res_result_tag	character varying	
res_tekey_ref	integer	référence
res_tekeyv_ref	integer	référence
res_comp_num_ref	integer	référence
res_unit_number_ref	integer	référence
res_rtkey_ref	integer	référence
res_rep_number_ref	integer	référence
res_retest_counter_ref	integer	référence
res_result_count	integer	
res_actual_unit_ref	integer	référence
tcu_exp_format	character varying	
tcu_tekey_ref	integer	référence
tcu_tekeyv_ref	integer	référence
tcu_comp_num_ref	integer	référence
tcu_unit_number	integer	
tcu_unit_key_ref	integer	référence
unit_key	integer	clé primaire
unit_name	character varying	Nom de l'unité
tc_comp_number	integer	
tc_tekey_ref	integer	référence
tc_tekeyv_ref	integer	référence
tc_comp_key_ref	integer	clé primaire
cn_comp_name	character varying	Composant nom
cn_comp_key	integer	clé primaire
ur_key	character varying	clé primaire
ur_unite	character varying	
observation	character varying	
observateur	character varying	

id	integer	
cl_code_cirad	integer	code cirad interlocuteur
share	boolean	
s_date_prelevement	date	date prélèvement échantillon
s_origine_geo	character varying	origine géographique échantillon
s_localisation_unite	character varying	localisation échantiilon
s_utm_zone	integer	coordonnées géo échantillon
s_utm_hemisphere	character	coordonnées géo échantiilon
s_localisation_autre	character varying	Autres données localisation echantillon
s_usage_sol	character varying	usage sol échantillon
do_share_conditions	USER-DEFINED	Dossier partagé
demandeur_nom	character varying	nom du demandeur
demandeur_prenom	character varying	prénom du demandeur
demandeur_email	character varying	email demandeur
demandeur_organisme	character varying	organisme demandeur
chercheur_nom	character varying	nom chercheur
chercheur_prenom	character varying	préniom chercheur
chercheur_email	character varying	email chercheur
chercheur_organisme	character varying	organisme chercheur
do_projet	character varying	projet dossier
do_pratiques	boolean	paratiques dossier
do_climat	boolean	climat indiqué dans le dossier
s_fraction_volumique	character varying	Fraction volumique échantillon