Numéro spécial
20 ans du GIS Sol

Le système d'information sur les sols de France

Capitaliser, analyser, diffuser, aller vers l'open data

C. Le Bas(1*), M. Brossard(2), J.-F. Brunet(3), L. Commagnac(4), A. Beaudou(5), S. Belbèze(3), M. Bellifa(4), H. Boukir(1), M. Dalmasso(4), N. Derrière(4), C. Lattelais(6), H. Le Martret(7), S. Lehmann(1), A. Pickel(1), A. C. Richer-de-Forges(1), N. Saby(1), A. Schellenberger(1), E. Taffoureau(3), B. Toutain(8), S. Wurpillot(4) et R. Yahiaoui(1)

1) INRAE, Unité Info&Sols, 45075 Orléans cedex 2, France
2) IRD, UMR 210 Eco&Sols, Campus SupAgro, 2 Place Viala, 34060 Montpellier cedex 2, France
3) BRGM, 3, Avenue Claude Guillemin, 45060 Orléans cedex 2, France
4) IGN, 73 avenue de Paris, 94165 Saint-Mandé cedex
5) Retraité, ancien pédologue IRD, France
6) INRAE, LESSEM, 38402 Saint Martin d'Hères cedex, France
7) Retraité, ancien informaticien et pédologue IRD, France
8) INRAE, Dynafor, 31326 Castanet Tolosan cedex, France

* Auteur correspondant : christine.le-bas@inrae.fr

RÉSUMÉ

Le Groupement d'intérêt scientifique sur les sols (GIS Sol) a été créé en 2001 avec la mission de concevoir et de coordonner l'inventaire et la surveillance des sols en France. Cette mission nécessite de capitaliser les données existantes ou nouvellement produites dans des bases de données, de les analyser, de produire de nouvelles connaissances et de restituer données et connaissances aux utilisateurs. Cet article décrit succinctement les bases de données et les outils associés aux grands programmes d'acquisition de données du GIS Sol, puis les travaux d'analyse de ces données pour produire des métriques et des prédictions sur les propriétés du sol dans l'espace et le temps, et enfin les outils et développements pour la consultation et la diffusion des données sur les sols. Ce panorama des outils actuellement disponibles ou en cours de développement montre l'importance des efforts réalisés sur ce plan par le GIS Sol depuis 2001.

Mots-clés
Base de données, système d'information, cartographie numérique des sols, statistiques, diffusion de données, web sémantique.

Comment citer cet article :
Le Bas C., Brossard M., Brunet J.-F., Commagnac L., Beaudou A., Belbèze S., Bellifa M., Boukir H., Dalmasso M., Derrière N., Lattelais C., Le Martret H., Lehmann S., Pickel A., Richer-de-Forges A. C., Saby N., Schellenberger A., Taffoureau E., Toutain B., Wurpillot S. et Yahiaoui R., 2024 - Le système d'information sur les sols de France : capitaliser, analyser, diffuser, aller vers l'open data - Étude et Gestion des Sols, 31, 59-73

SUMMARY

THE FRENCH SOIL INFORMATION SYSTEM: Capitalize, analyze, disseminate, move towards open data

The Scientific Interest Group on soil (GIS Sol) was created in 2001 with the mission of designing and coordinating the inventory and monitoring of soils in France. This mission requires capitalizing existing or newly produced data in databases, analyzing them, producing new knowledge and bringing data and knowledge to the attention of users. This article briefly describes the databases and their associated tools for the main data acquisition programmes of the GIS Sol. Then the work of analyzing these data to produce statistics and predictions on soil properties is presented. Finally, the tools and developments for consultation and dissemination of soil data are described. This panorama of tools currently available or under development shows the importance of the efforts made by GIS Sol since 2001.

Key-words

Database, information system, digital soil mapping, statistics, data dissemination, semantic web.

RESUMEN

EL SISTEMA DE INFORMACIÓN DEL SUELO FRANCÉS: Capitalizar, analizar, difundir, avanzar hacia los datos abiertos

El Grupo de Interés Científico del Suelo (GIS Sol) fue creado en 2001 con la misión de diseñar y coordinar el inventario y seguimiento de suelos en Francia. Esta misión requiere capitalizar datos existentes o recién producidos en bases de datos, analizarlos para producir nuevos conocimientos y devolver datos y conocimientos a los usuarios. Este artículo describe brevemente las bases de datos y sus herramientas asociadas para los principales programas de adquisición de datos del GIS Sol, luego el trabajo de análisis de estos datos para producir estadísticas y predicciones sobre las propiedades del suelo y finalmente las herramientas y desarrollos para la consulta y difusión de datos del suelo. Este panorama de herramientas actualmente disponibles o en desarrollo muestra la importancia del esfuerzo realizado por GIS Sol desde 2001.

Palabras clave

Base de datos, sistema de información, cartografía digital de suelos, estadísticas, difusión de datos, web semántica.

1. INTRODUCTION

Les sols sont au cœur de grands enjeux de l'humanité, qu'ils soient alimentaires, sanitaires ou environnementaux (Lehmann *et al*., 2020). Cette place prépondérante implique une bonne connaissance de leurs propriétés et de leurs évolutions, afin de pouvoir mettre en place des politiques publiques assurant leur préservation et la pérennité de leurs fonctions. Le Groupement d'intérêt scientifique sur les sols, le GIS Sol, a été créé, en 2001, dans le but de concevoir et de coordonner des programmes nationaux d'inventaire et de surveillance des sols (Arrouays *et al*., 2022). A cette époque, il était également nécessaire de développer un système d'information sur les sols de France permettant de capitaliser les données acquises, de les analyser, de produire de nouvelles connaissances et de restituer les données et les connaissances aux utilisateurs. Pour cela, le GIS Sol s'est appuyé sur des bases de données préexistantes comme DoneSol pour la France métropolitaine (Gaultier *et al*., 1993) ou Valsol pour l'Outre-mer (Beaudou et Le Martret, 2004). L'arrivée dans le GIS Sol de nouveaux membres comme l'IGN en 2019 et le BRGM en 2021 a fait entrer dans le champ du GIS Sol de nouveaux programmes sur les sols forestiers et les sols urbains. Il résulte de cet historique un ensemble de systèmes d'information développés pour les différents programmes du GIS Sol et dont cet article montre la richesse.

Parallèlement, depuis plus de 20 ans, le contexte relatif à la gestion des informations a également fortement évolué vers une ouverture de plus en plus grande des données (Rennes *et al*., 2023). De nombreux travaux ont été réalisés sur la manière de le faire. Comment les rendre faciles à trouver, accessibles, interopérables et réutilisables selon les principes des données FAIR (Findable, Accessible, Interoperable, Reusable) (Wilkinson *et al*., 2016) ? De nouveaux standards sont apparus comme ceux de l'Open Geospatial Consortium (OGC) pour les données spatiales (CSW en 1999, WMS en 2000, WFS en 2002). Ils ont été repris par la Directive européenne Inspire en 2007 qui vise à créer une infrastructure européenne des données géographiques sur l'environnement. Des standards ont également été définis par le World Wide Web Consortium (W3C) pour les données diffusées sur internet (par exemple format RDF en 2014). De nouvelles technologies et infrastructures se sont développées pour permettre l'interopérabilité des données. Le GIS Sol a investi dans ces nouvelles technologies et ces standards pour mieux diffuser ses données et en permettre l'interopérabilité.

L'objectif de cet article est de faire la synthèse des travaux réalisés dans ce domaine par le GIS Sol, en montrant le chemin parcouru depuis 2001. Nous présentons d'abord les différentes bases de données permettant de capitaliser les données, puis nous décrivons succinctement les outils développés pour les analyser, enfin, nous abordons les outils disponibles pour assurer leur visualisation et leur diffusion.

2. CAPITALISER LES DONNÉES SUR LES SOLS DE FRANCE

2.1. La base de données DoneSol : capitaliser les données des programmes IGCS et RMQS

Le coût d'acquisition des données sur les sols (y compris le temps passé) rend leur numérisation primordiale. C'est pourquoi, dans le cadre du programme Inventaire, gestion et conservation des sols (IGCS), la base de données DoneSol *(Figure 1)* a été constituée dès 1992 pour stocker, en un endroit unique et de façon harmonisée, l'ensemble des études pédologiques réalisées dans le cadre de ce programme (Gaultier *et al*., 1993). Son développement s'est appuyé, d'une part, sur des travaux antérieurs visant à décrire les unités cartographiques de sol, et d'autre part, sur le système standardisé de description des profils de sols et des résultats d'analyses de laboratoire STIPA (Système de Transfert de l'Information Pédologique et Agronomique, Bertrand *et al*., 1979). Cette base de données a évolué régulièrement avec une version 2 en 2001, l'intégration des données du programme Réseau de mesures de la qualité des sols (RMQS) en 2002 (Grolleau *et al*., 2004), et une version 3 en 2013 (Toutain, 2013). Ces différentes versions ont permis de faire évoluer le modèle de données afin d'intégrer de nouveaux types de données (données du RMQS, données sur les documents de la cartothèque, extension aux données de l'Outre-mer), d'améliorer leur description, d'améliorer l'efficience de la base de données et son évolutivité, en intégrant l'évolution des technologies.

Le modèle de données actuel de DoneSol (https://dw4.gissol.fr/fichiers/modele_physique_donnees_donesol310_2023.pdf) permet de stocker plusieurs ensembles de données dans une base de données relationnelle, gérée sous PostgreSQL® :

- les données sur les études pédologiques et les documents associés qui sont présents dans la cartothèque ;
- les données surfaciques décrivant les Unités cartographiques de sol (UCS), les Unités typologiques de sol (UTS) et leurs strates ;
- les données ponctuelles décrivant les profils et les horizons ;
- les données sur les sites RMQS incluant le suivi des interventions réalisées sur les sites ;
- les données sur les échantillons qui sont traités dans le Conservatoire européen d'échantillons de sol, notamment le traçage de leurs traitements ;
- les données sur les résultats d'analyse chimique ou physique.

Le schéma conceptuel de DoneSol inclut non seulement la description des objets (*i.e.* UTS) et de leurs propriétés (*i.e.* nom de l'UTS), mais aussi les relations entre ces objets (*i.e.* les horizons liés à un profil) et les contraintes d'intégrité (*i.e.* le pH est compris entre 0 et 14).

Des outils de saisie et de consultation ont été développés

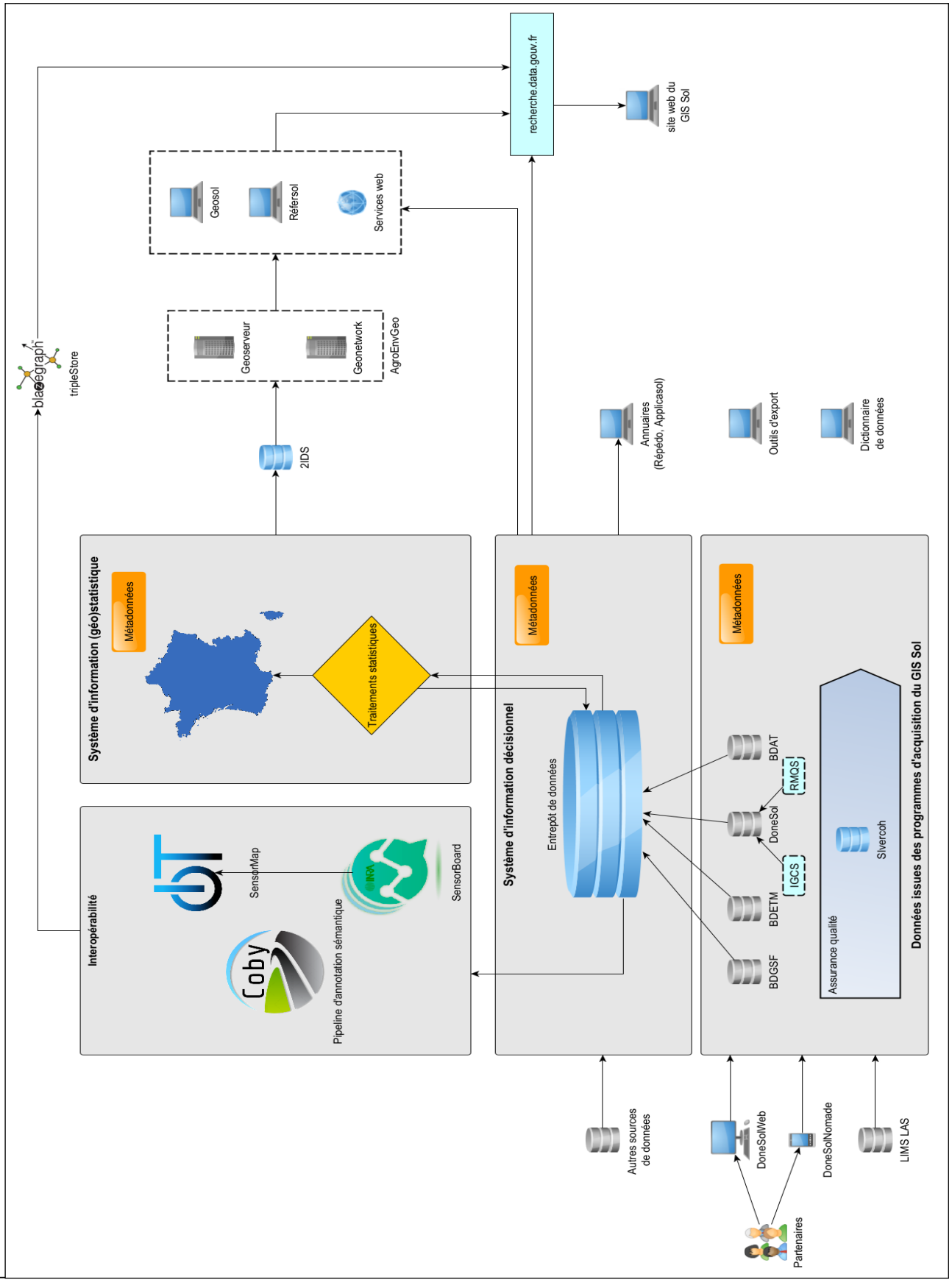

Figure 1 : Structure du système d'information sur les sols de France, SI Sol

Figure 1: *Structure of the French soil information system, SI Sol*

par INRAE et certains partenaires régionaux du programme IGCS dès les années 2000 (Grolleau *et al.*, 2004). Ils ont permis d'introduire des vérifications dès la saisie des données, améliorant ainsi leur qualité. En 2006, une interface web de saisie et de consultation a été développée, DoneSolWeb, permettant un accès facilité à la base de données, une connexion sécurisée et de meilleures fonctionnalités. Cette interface est en constante amélioration afin de mieux répondre aux demandes des utilisateurs et de s'adapter aux progrès technologiques.

Dans le cadre de la 2e campagne du RMQS (2016-2027), le développement d'un utilitaire de consultation et de saisie nomade pour le terrain a été réalisé afin de simplifier le travail de terrain en permettant i) la consultation des données de la première campagne du RMQS nécessaire au retour sur site et ii) la saisie directe des observations de la 2e campagne. Cet outil fluidifie également le transfert des données entre les équipes des partenaires régionaux qui interviennent sur le terrain et la base DoneSol, et limite la saisie manuelle de données, source d'erreurs. Cette application, DoneSolNomade, fonctionne avec tout type d'appareil (ordinateur, tablette ou smartphone) et prend en charge les contraintes du terrain en matière de connexion Internet. Une version de test a été déployée en 2023, sa mise en production sera effective en 2024.

Un premier niveau de vérification de la cohérence des données intervient au moment de la saisie, grâce à des tests au niveau de l'interface DoneSolWeb. Une vérification de la cohérence des données est aussi nécessaire *a posteriori*. Cette vérification s'effectue depuis 2013 (Chapuis *et al.*, 2013) grâce à l'outil Sivercoh (SI de Vérification de Cohérence). Développé dans un souci de généricité, Sivercoh est un système d'information permettant de vérifier la cohérence de données stockées dans des bases de données relationnelles. Cet outil permet de réaliser des contrôles complexes des jeux de données, comme par exemple i) d'importer des requêtes de vérification de cohérence, ii) de les organiser en « jeux » de requêtes, et iii) de les exécuter sur une base de données relationnelle. Un rapport dresse la liste des incohérences à vérifier pour améliorer la qualité des données. L'utilisateur peut alors soit corriger l'anomalie, soit la justifier si celle-ci correspond à une situation particulière.

Des formations à la saisie des données et à l'interrogation de DoneSol sont dispensées gratuitement depuis 2005 (Richer-de-Forges *et al.*, 2018). Cette offre de formation évolue régulièrement afin de répondre aux besoins des utilisateurs. Ainsi, depuis 2016, une formation dédiée à l'interrogation des données parcourt toutes les étapes depuis l'interrogation de DoneSol jusqu'à la réalisation d'une carte thématique (*i.e.* carte du Réservoir Utilisable). La formation régulière des opérateurs de saisie depuis 2005 a amélioré la qualité des données de façon appréciable. Cette offre de formation est complétée par un ensemble de vidéos en ligne dédiées chacune à un aspect spécifique de DoneSol (*e.g.* personnalisation des interfaces, format d'export de DoneSolWeb, etc.).

2.2. La base de données Valsol : capitaliser les données sur les Outre-mer

Alors que la base de données DoneSol a vocation à capitaliser les données du territoire national métropolitain, la base de données Valsol a pour objectif de sauvegarder et de valoriser les données acquises par l'Orstom puis l'IRD en zones tropicales. Même si Valsol a bénéficié de l'expérience d'INRAE avec la base de données DoneSol, plusieurs éléments ont conduit à l'élaboration d'un modèle conceptuel de données différent. Les données Orstom-IRD, nombreuses et variées, sont spécifiques des régions intertropicales ; le dictionnaire des données y est donc adapté. De plus, les travaux Orstom-IRD concernent généralement la reconnaissance et la cartographie des sols à des échelles très variées, allant du 1/20 000 (ex. Basse Terre de la Guadeloupe, Martinique) à des échelles continentales en Afrique (1/500 000, voire 1/1 000 000). L'approche morpho-pédologique des pédologues Orstom-IRD étant à la base de ces travaux de terrain, Valsol a donc suivi cette logique d'organisation de l'information. Elle a été développée à partir de 1997 (Beaudou et Le Martret, 2004) avec deux prérequis : premièrement, s'appuyer sur des logiciels libres (PostgreSQL®, etc.) afin de ne pas pénaliser les partenaires des pays tropicaux, et deuxièmement, la nature des données pédologiques existantes étant très variée, maintenir une souplesse qui permette des restitutions adaptées aux différents milieux.

Valsol est entrée en production en 2003, ouverte aux partenaires étrangers, mais n'a plus été maintenue à partir de 2014. En parallèle, des interfaces de visualisation des données géographiques (Le Martret *et al.*, 2008) ont été développées.

Les données sont regroupées selon deux grandes subdivisions :

- les données générales concernant les études, les auteurs, les cartes, les organismes, avec les tables correspondantes. Elles sont indépendantes des données thématiques et peuvent se rattacher aux métadonnées.
- les données thématiques séparées en deux ensembles : (1) les données morphologiques regroupant les caractéristiques des grands paysages, paysages et segments ; et (2) les données pédologiques (description des sols et des horizons, et leurs caractères analytiques).

Les données thématiques ont été subdivisées en deux groupes en fonction de leur nature : (1) les données réelles décrivant des objets observés sur le terrain, ponctuels ou linéaires (grands paysages ou transects, paysages ou séquences, segments, profils, horizons), spatialement localisés et représentés par des polylignes ou des points sur une carte ; (2) les données cartographiques correspondant à une représentation spatiale d'un ensemble de sols, d'un sol et de ses composants (unité de grand paysage, unité de paysage, unité de segment) et qui sont représentées par un polygone ; enfin (3) les données synthétiques, résultat d'un travail de

synthèse associant plusieurs sources d'informations. Elles sont construites à partir des données réelles (moyennes) et permettent de caractériser les sols (profils, horizons, analyses synthétiques).

Les travaux d'inventaires et de prospections pédologiques, puis leur déclinaison dans divers documents, sont anciens dans les Outre-mer. Ils s'échelonnent sur plusieurs territoires dès les années 1950, tant en milieu continental que dans les domaines insulaires volcaniques (Brossard *et al.*, 2023). Valsol a permis de sauvegarder des données originales non encore publiées ainsi que des données issues de travaux d'inventaires pédologiques publiés. Ainsi les données des Antilles couvrent environ 900 profils de sols, celles de la Guyane 250, celles de la Réunion 270 ; selon les études, les contenus analytiques sont variables.

En 2021, un travail méthodologique d'échanges de données entre Valsol et DoneSol a été initié. L'adaptation d'un dictionnaire Valsol qui permette la mise en correspondance des données Valsol avec DoneSol, sous forme de table, a été effectuée. Des tests ont été réalisés avec les données de Guyane. Les scripts de transfert sont aujourd'hui prêts à être appliqués sur des bases conformes au format Valsol vers DoneSol3. Il est donc possible d'intégrer dans DoneSol3 des données Valsol selon leur disponibilité. Ce travail a permis d'identifier les difficultés liées à la mise en conformité de données pédologiques, celle-ci nécessitant un long travail d'expertise.

2.3. Les bases de données à vocation statistique sur les sols agricoles : BDAT et BDETM

Depuis leur création, les programmes Base de données d'analyses de terre (BDAT) et Base de données sur les éléments traces métalliques (BDETM) ont pour objectif de capitaliser dans un système d'information national les résultats d'analyses de sols produits dans le cadre des activités agricoles (Baize *et al.*, 2006 ; Saby *et al.*, 2014). Une très grande majorité de ces analyses est réalisée par les laboratoires d'analyses de sols agréés par le ministère en charge de l'agriculture. Pour la plupart, ces analyses concernent les horizons de surface des sols cultivés et sont demandées par les agriculteurs pour piloter la fertilisation de leur parcelle (programme BDAT, données d'analyses agronomiques) ou pour conduire des plans d'épandage de boues de stations d'épuration (programme BDETM, données sur les éléments traces métalliques). Chaque analyse de sol est géoréférencée à la commune pour limiter la problématique juridique liée à la protection des données personnelles, ce qui rend leur exploitation plus complexe. Cette sécurité est nécessaire pour être conforme au règlement européen sur la protection des données (RGPD). Une procédure de traitement des fichiers d'analyse permet leur insertion dans une base de données relationnelle dédiée, sous postgreSQL® après vérification et validation de l'analyse.

2.4. La base de données BDSolU : capitaliser les données sur les sols urbains

La France s'est dotée d'un cadre législatif (Loi ALUR - Accès au logement et à un urbanisme rénové - 2014 (MTECT, 2022) et Loi NOTRE - Nouvelle organisation territoriale de la République (Vie Publique, 2015)) et de mécanismes financiers (Plan « France-Relance » (MEFSIN, 2022)) pour limiter l'étalement urbain et favoriser la restauration des friches, y compris celles potentiellement polluées. Du point de vue technique, les recommandations ministérielles en matière de gestion des sites (potentiellement) pollués, préconisent depuis longtemps le recours aux concentrations habituelles des sols en métaux, métalloïdes et substances organiques, pour diagnostiquer une éventuelle pollution ou pour gérer les terres excavées. Intervenant en appui à cette politique, l'Ineris et le BRGM éditent régulièrement des guides de bonnes pratiques de prélèvement, d'analyse, de caractérisation et de valorisation des sols (MEEM, 2018).

Or, les sols des zones urbaines résultent de plusieurs siècles d'activités humaines incluant remaniement, déblais, remblais et dépôts d'émissions atmosphériques provenant d'usines, de chaufferies et de trafic routier. La composition chimique des sols urbains, notamment pour des molécules indésirables, est une information mal connue, difficile à appréhender et pourtant de plus en plus recherchée. Elle est l'objet d'une demande sociétale croissante en vue de garantir la santé publique tout en économisant les ressources foncières urbaines et agricoles.

Pour un territoire donné, les concentrations naturelles d'éléments ou de substances chimiques dans le sol, en dehors de tout apport lié aux activités humaines, constituent le fond pédogéochimique naturel (FPGN). Le FPGN et les concentrations diffuses dues aux activités humaines représentent le fond pédogéochimique anthropisé (FPGA). Les concentrations diffuses proviennent des retombées des émissions atmosphériques, proches ou lointaines, des substances introduites par les pratiques agricoles ainsi que des zones de remblais d'origine naturelle. Le FPGA est donc compris entre des seuils, appelés valeurs de fond, qui le distinguent du FPGN, et des concentrations relevant des anomalies d'origine humaine.

En vue de déterminer ces valeurs de fond, l'Ademe et le BRGM recueillent et harmonisent depuis 2010, sur l'ensemble du territoire national, des analyses de sols représentatifs du fond pédogéochimique anthropisé ainsi que la description de leurs conditions d'obtention (localisation, méthodes de sondage, de prélèvement, d'analyse, etc.). Il s'agit du projet « Etablissement de valeurs de fond pédogéochimique - FGU ». Ces informations sont réunies dans la Base de Données des analyses de Sols Urbains, BDSolU (www.bdsolu.fr) développée sous postgreSQL®. Elles ne résultent pas de campagnes de prélèvements et d'analyses conduites spécifiquement pour cet objectif, mais de divers projets de recherche, d'aménagements

urbains ou de détermination de référentiels pédogéochimiques locaux (Brunet *et al*., 2023).

Depuis 2018, ce projet, conduit en partenariat avec INRAE et le bureau d'étude eOde avec l'appui de Mines Paris-Tech, comprend, outre la numérisation des données, trois volets destinés respectivement :

- à la mise au point d'une méthode reconnue de calcul des valeurs de fond ;
- à la mise au point de l'interopérabilité des données DoneSol et BDSolU ;
- au développement, à terme, d'un outil cartographique interactif pour la diffusion des résultats.

Il n'existe pas actuellement, de par le monde, une méthode de référence de détermination des valeurs de fond. Diverses méthodes sont mises en œuvre en fonction des domaines scientifiques, des choix méthodologiques et de chaque pays, des objectifs visés, voire des auteurs (Belbèze *et al*., 2023). La mise au point d'une méthode reconnue par les principales instances nationales fait donc partie des objectifs et se met en place progressivement en France.

L'interopérabilité des données DoneSol et BDSolU permettra une interrogation simultanée des deux bases de données afin :

- d'étendre l'effectif des populations servant à la détermination des valeurs de fond ;
- d'obtenir des résultats exploitables à l'interface entre zones agricole et urbaine.

Plusieurs challenges importants dépendent de la diffusion des référentiels de fond pédogéochimique au moyen d'un outil interactif (Brunet *et al*., 2023) :

- interroger les bases de données et réaliser les calculs en temps réel ;
- produire des calculs experts au moyen de méthodes statistiques et géostatistiques avancées ;
- fournir des cartographies représentant par interpolation les zones géographiques couvertes par les référentiels, intégrant des informations sur les incertitudes associées ;
- réaliser des représentations cartographiques en temps réel en préservant la confidentialité des lieux de prélèvement parfois souhaitée par les propriétaires des données.

2.5. La base de données de l'Inventaire forestier national : capitaliser les données sur les sols forestiers

Depuis les années 1960, l'Inventaire forestier national (IFN) acquiert des données sur les ressources forestières. Depuis son changement de méthode en 2005, l'IFN est un inventaire statistique par échantillonnage spatial systématique (selon une grille nationale de maille kilométrique déployée à l'échelle de la France métropolitaine) à caractère obligatoire. Il a obtenu le label d'intérêt général et de qualité statistique en juin 2017 auprès du Conseil national de l'information statistique (CNIS). Les données relevées concernent la forêt indépendamment des questions de propriété. C'est un inventaire anonyme et les coordonnées précises des placettes sont protégées par diverses législations (données personnelles, secret statistique, protection de l'environnement). Pour plus de détails sur la méthode d'inventaire, voir https://inventaire-forestier.ign.fr. Les données sols de l'inventaire forestier font partie des données écologiques relevées en forêt, et constituent une des plus longues séries temporelles de l'inventaire forestier. Elles sont relevées sans modification notable du protocole depuis 1992. Les données sur les sols forestiers, versées dans la base de données DoneSol en 2021, remontent à 1987. La description du sol sous forêt, dans le cadre du protocole IFN, nécessite d'étudier l'humus présent (donnée d'importance majeure pour les sols forestiers) identifié selon une clé de détermination réalisée d'après Jabiol *et al.* (1995). Une fosse pédologique de 40 cm de profondeur est ensuite creusée, puis un sondage à la tarière jusqu'à 1 mètre de profondeur est effectué, si possible. Des levés sont réalisés pour décrire au mieux le point d'inventaire. La fosse et le sondage sont réalisés dans une zone considérée comme représentative de l'ensemble de la surface de la placette (cercle de 15 m de rayon), en évitant au mieux les zones perturbées ou anthropisées.

L'étude de la fosse et du sondage permet de renseigner à dire d'expert (l'IFN ne réalise pas d'analyses chimiques ou granulométriques sur les points d'inventaire) près de 30 variables, caractérisant notamment l'humus, la charge en éléments grossiers, la texture des horizons, la profondeur d'apparition de la décarbonatation, la présence d'hydromorphie, etc. Après avoir renseigné ces éléments, le type de sol est identifié, selon une clé de détermination proche de la classification de Duchaufour (Duchaufour, 1991). Les données de l'inventaire sont stockées dans des bases de données gérées sous PostgreSQL®. Les données sol sont dans une table dédiée. Le lien avec les autres données acquises sur la placette (dendrométriques, écofloristiques, etc.), stockées dans d'autres tables, se fait grâce à un numéro d'identifiant unique de placette.

Pour l'Inventaire forestier, la capacité à analyser conjointement les données dendrométriques et pédologiques est essentielle, hier pour analyser les relations entre type de stations forestières et productivité des essences, comme aujourd'hui pour anticiper l'adaptation des forêts et des essences au changement climatique.

3. ANALYSER LES DONNÉES : STATISTIQUES ET PRÉDICTIONS

3.1. Développer un entrepôt de données : système d'information statistique et système d'information décisionnel

L'objectif principal de la base de données DoneSol est de stocker des données. Son modèle conceptuel a donc été développé pour optimiser leur stockage et assurer leur intégrité et leur unicité. Ainsi, DoneSol est une base de données relationnelle comprenant un nombre de tables important ce qui nécessite une bonne connaissance de sa structure et des compétences en requêtage SQL pour en extraire l'information recherchée. A cela s'ajoutent les enjeux de l'open data qui nécessitent de mettre en place des outils de consultation et de téléchargement de données, et de leur diffusion massive. Afin d'une part, de faciliter l'interrogation des données, et d'autre part, de disposer de jeux de données validés, faciles à mettre à jour et à diffuser, un système d'information décisionnel (SID) et un système d'information statistique (SIS) ont été développés *(Figure 1).*

Le système d'information décisionnel (SID) met en œuvre un ensemble de techniques et d'outils pour transformer des données issues de bases de données en données à valeur ajoutée, lisibles facilement et directement utilisables. Il est employé pour faciliter l'accès, l'interrogation et l'analyse des données en provenance des bases de données du système d'information opérationnel (DoneSol, les bases de données BDAT et BDETM) mais aussi pour faciliter leur croisement avec d'autres sources de données (données climatiques, topographiques, d'occupation du sol, etc.).

Les données nécessaires sont extraites des bases de données, puis transformées selon des règles prédéfinies, dites métier, incluant une connaissance fine de leur organisation (ce qui peut inclure des calculs) et enfin restituées sous forme d'informations exploitables (datamart ou magasin de données). Cela permet de capitaliser l'expertise acquise sur les données et les traitements mis au point par les ingénieurs chargés de l'analyse des données en les automatisant. Cela facilite également le suivi des différentes versions des jeux de données, en perpétuelle évolution du fait des corrections apportées aux données de base. Les datamarts ainsi produits constituent des supports de partage des données, dans le respect des droits de propriété intellectuelle, au sein de l'unité Info&Sols, auprès de partenaires, et dans le cadre de projets nécessitant la structuration, l'agrégation et le croisement de données multi-source. Ils sont également le support des applications de consultation et de téléchargement de données (services web, Geosol) pour celles qui peuvent être diffusées directement par INRAE. Cette architecture s'inscrit pleinement dans une démarche qualité : informations contrôlées, reproductibles, traçables et décrites par des métadonnées.

Le système d'information statistique (SIS), quant à lui, a été mis en place pour organiser et automatiser les traitements statistiques réalisés sur les données avant leur diffusion. Il prend la forme d'une chaîne de traitement des données qui sont stockées dans le système d'information décisionnel. Il peut produire trois formes de résultats : des tableaux de données, des cartes de type choroplèthe et des cartes de type grille issues de processus d'interpolation spatiale. L'ensemble de ces données est diffusable à travers des services web.

3.2. Prédire les propriétés des sols, calculer des bilans et des indicateurs

En disposant, dans le système d'information, des observations conjointes de propriétés des sols et de variables représentant certains de leurs déterminants, il est possible, en utilisant des modèles statistiques adaptés, de modéliser spatialement et temporellement ces propriétés. Les jeux de données ainsi que les méthodes utilisées sont adaptés à l'objectif visé par la modélisation, à savoir comprendre et modéliser les distributions spatio-temporelles des sols et de leurs propriétés, prédire spatialement ces propriétés (cartographie des sols par modélisation statistique (CSMS)) ou produire des indicateurs. Le choix des méthodes statistiques tient également compte de la nature particulière des données sols disponibles : distributions avec données anomaliques, données censurées, distributions log-normales, etc.

Dans le cadre de la prédiction de propriétés des sols, l'intérêt de ces méthodes, outre la production de jeux de données spatialement exhaustif à résolution fine (90 m pour le programme *GlobalSoilMap* (Chen *et al.*, 2022 ; Loiseau *et al.*, 2020)), est de fournir les incertitudes d'estimation associées. Ces incertitudes permettent d'informer l'utilisateur sur la précision de la prédiction et d'identifier des zones prioritaires pour l'acquisition de données. On peut citer, par exemple, les prédictions du réservoir utile des sols par Román Dobarco *et al.* (2019) ou les prédictions des éléments traces métalliques par Saby *et al.* (2018).

Afin de répondre à la demande sociétale et des pouvoirs publics, la production d'indicateurs, de statistiques ou encore de bilans sont également réalisés. Ces bilans ou statistiques portent, comme la cartographie des sols par modélisation statistique, sur différentes propriétés des sols qui peuvent être aussi variées que la quantité d'ADN microbien, le réservoir utile, le stock de carbone ou les besoins en fertilisation minérale. Cependant, à la différence de la cartographie des sols par modélisation statistique, il s'agit ici moins de prédire le plus finement possible la distribution spatiale d'une variable, que d'estimer sa valeur moyenne, sa variabilité, au sein d'unités spatiales, temporelles ou encore thématiques (*e.g.* l'occupation du sol) homogènes et

qui ont un sens pour les pouvoirs publics. On répond donc plus à un besoin de quantification qu'à celui de localisation (de Gruijter *et al*., 2006).

L'ensemble de ces travaux s'effectue dans un objectif de capitalisation et d'automatisation des traitements en s'appuyant sur l'entrepôt de données afin d'améliorer la diffusion des résultats.

3.3. Définir un workflow pour le calcul du fond pédogéochimique

La capitalisation de données sur les sols des zones urbaines dans la BDSolU a pour objectif de permettre la détermination des valeurs de fonds pédogéochimiques. Un workflow basé sur quatre étapes a été mis au point :

1. Une première étape i) prend en compte une zone géographique (agglomérations de communes, communes et quartiers), ii) une profondeur d'intérêt (prélèvements inférieurs ou supérieurs à 30 cm sous la surface du sol) selon que le FPGA recherché est destiné à un diagnostic de l'exposition sanitaire à des polluants du sol ou à la gestion de terres excavées (Ademe, 2018), et iii) évalue s'il existe un effectif suffisant d'analyses.
2. Une sélection des données disponibles dans la zone et à la profondeur choisie permet d'extraire les points de prélèvement et les échantillons les plus représentatifs du fond pédogéochimique recherché. En effet, l'importante hétérogénéité des données de la BDSolU liée aux origines diverses des données nécessite de vérifier leur représentativité vis-à-vis de l'objectif de l'utilisateur. Ceci est possible grâce aux divers champs qui décrivent les conditions d'obtention des données. Si la sélection n'atteint pas le nombre minimal requis d'échantillons pour la détermination de valeurs de fond, elle peut être élargie sur des critères moins stricts qui seront notifiés aux utilisateurs. Cette méthode dite des « indicatrices » a été transcrite dans un algorithme sous le langage R®.
3. Afin de compenser l'irrégularité de la répartition spatiale des points de prélèvement qui peut biaiser les résultats en donnant plus de poids à une zone géographique qu'à une autre présentant une moindre densité de prélèvements, une méthode simple de dégroupement est mise en œuvre au moyen d'un algorithme tenant compte de la densité des points de prélèvements dans une grille arbitraire de 100 cellules tracée sur la zone géographique considérée (méthode dite « à fenêtre » ; Sauvaget *et al*., 2022). Pour tenir compte des concentrations très faibles inférieures aux limites de quantification des méthodes d'analyse et ainsi éviter un biais statistique important (Helsel, 2005), une méthode de discrétisation d'une loi entre 0 et la limite de quantification est appliquée. Transcrite sous langage R®, elle vient compléter le workflow de traitement des données et permet de reconstituer la population d'analyses entre ces bornes (de Fouquet, 2021). Malgré la sélection opérée pour obtenir des jeux de données représentatifs et des populations aux résultats homogènes, quelques valeurs extrêmes (ou outliers) peuvent subsister en raison d'anomalies ponctuelles naturelles ou anthropiques. Normalement en faible nombre, ces valeurs ont peu d'impact sur les résultats des calculs de détermination des valeurs de fond à l'échelle choisie. De plus, il peut être délicat de les distinguer du reste de la population et leur suppression n'est pas recommandée (Helsel, 2005). Toutefois, cette suppression peut être intégrée à l'algorithme de manière judicieuse afin d'obtenir une population plus homogène, adaptée à un traitement statistique.
4. Le calcul des valeurs de fond est réalisé par la méthode fondée sur la limite supérieure interne définie par Tukey dans la méthode des vibrisses (Tukey, 1977) par souci de pragmatisme et de cohérence avec les travaux conduits jusqu'à présent par INRAE dans le cadre du GIS Sol. Ce calcul, associé à la partie gérant les valeurs inférieures aux limites de quantification, vient compléter le workflow qui, en présence d'un effectif inférieur à 30, délivre des statistiques de base, des histogrammes, des boîtes de Tukey, ainsi que des cartes localisant les points de prélèvement. Lorsque l'effectif dépasse 30, ces informations sont complétées par des propositions de valeurs de fond (Brunet *et al*., 2023).

3.4. Calculer des statistiques sur les sols forestiers

Bien que seuls les points présentant une occupation du sol boisée et disponibles pour la production de bois soient enquêtés, les estimateurs statistiques de l'Inventaire forestier national, notamment les estimateurs de surface, sont valides. Ainsi, il est possible de calculer, à partir des données de l'Inventaire forestier national, une estimation des surfaces des principaux types de sol avec un intervalle de confiance associé. De même, il est possible de calculer des moyennes par hectare, avec un intervalle de confiance associé, pour certains paramètres (indice de réserve en eau utile, profondeur de sondage, etc.) que ce soit par type de sol, par type de peuplement forestier ou selon d'autres critères levés ou calculés par l'IFN. Ces calculs sont réalisables grâce à un service de calcul, développé en interne à l'IGN, accessible sur internet (https://inventaire-forestier.ign.fr/spip.php?rubrique226), qui s'appuie à la fois sur les données de la base d'exploitation (collectées et calculées) et sur les informations de post-stratification. Ce service, point de passage unique pour le calcul de résultats, garantit la cohérence et la fiabilité de l'exploitation statistique des données collectées. Un nombre minimal de points est toutefois requis dans chaque jeu de données sélectionné pour obtenir des résultats statistiquement significatifs.

4. VISUALISER ET DIFFUSER LES DONNÉES : VERS L'OPEN DATA

La question de la diffusion des données a été de plus en plus prégnante ces dernières années, en raison des enjeux sur les sols mais aussi de l'évolution du contexte juridique vers l'ouverture des données. Le GIS Sol a affiché explicitement dans ses dernières conventions le souhait d'une large diffusion des données acquises dans le cadre de ses programmes. Cependant, la propriété des données du SI Sol est particulièrement complexe du fait du long historique de ces programmes. De plus, la législation qui s'applique est elle-même complexe (Rennes *et al.*, 2023). Par conséquent, selon les programmes, certaines données sont diffusées en open data par INRAE ou par les partenaires régionaux du GIS Sol (*e.g.* certains Référentiels régionaux pédologiques, données RMQS avec coordonnées théoriques), d'autres ne sont que consultables (*e.g.* données BDAT), d'autres enfin ne sont pas diffusables (*e.g.* données RMQS avec les coordonnées réelles par exemple).

4.1. Les outils de consultation associés à DoneSol

Plusieurs outils de consultation sont disponibles à partir du site web du GIS Sol. Ils utilisent des informations décrivant les études pédologiques, à savoir :

- le répertoire des organismes intervenant en pédologie, Répédo https://annuaires.gissol.fr/repedo
- le répertoire des applications thématiques réalisées à partir des bases de données cartographiques sur les sols, Applicasol (Girot *et al.*, 2017) https://annuaires.gissol.fr/applicasol
- le répertoire des études pédologiques, Refersols https://webapps.gissol.fr/georefersols/.

Ces outils sont en accès libre depuis le site web du GIS Sol https://www.gissol.fr/.

En 2020, un important travail a été réalisé avec le RMT Sols et Territoires (Messant *et al.*, 2021) pour la création de la carte des sols dominants de France consultable sur le Géoportail https://www.geoportail.gouv.fr/donnees/carte-des-sols. Cette carte est issue de la compilation des référentiels régionaux pédologiques au 1/250 000ème disponibles sur la France métropolitaine. Elle permet une interrogation des polygones avec affichage des informations descriptives des UCS et des UTS présentes. Pour des raisons juridiques, cette carte est en consultation et ne peut être téléchargée.

Dans le cadre du programme IGCS, les référentiels régionaux pédologiques (RRP) sont diffusés par les partenaires régionaux les ayant réalisés. Certains d'entre eux ont créé des outils de consultation de leur RRP. On peut citer par exemple pour la Bretagne le site http://www.sols-de-bretagne.fr/. Le projet Websol (Vinatier *et al.*, 2013), soutenu par un financement CASDAR entre 2006 et 2009 puis repris par le RMT Sols et Territoires, a également développé la visualisation des RRP. Plusieurs plates-formes Websol ont ainsi été déployées en région comme, par exemple, sur la Bourgogne https://solsdebourgogne.fr/ ou sur Rhône-Alpes http://rhone-alpes.websol.fr/.

Un outil de consultation des statistiques réalisées sur le RMQS a également été développé en utilisant le logiciel de statistiques R®. L'intérêt de ce type de technologie est la facilité et la rapidité de développement pour un résultat ergonomique et interactif https://traitementinfosol.pages.mia.inra.fr/statistiquesrmqs/.

4.2. L'outil de consultation des statistiques issues de la BDAT : Geosol

Geosol est une application web développée pour la consultation des données agrégées issues du programme BDAT (https://webapps.gissol.fr/geosol/). Elle permet d'afficher des cartes à l'échelle nationale, des statistiques sur les propriétés mesurées et stockées dans la BDAT, calculées pour différentes entités administratives. L'application s'appuie sur le système d'information décisionnel. Les données agrégées sont calculées par le système d'information statistique de manière semi-automatique *(cf. 3.1)*. De nouvelles données étant incluses dans la BDAT chaque année, ce système permet une mise à jour fluide des données consultables. Les couches graphiques sont, quant à elles, gérées sur l'outil Geoserver de l'infrastructure de données spatialisées agroenvgeo *(cf. 4.3)*.

4.3. Diffuser des données spatiales

La diffusion de données spatiales sur les sols est régie par le code de l'environnement et plus précisément par la directive européenne Inspire. Cela implique de s'appuyer sur une infrastructure de données spatiales (IDS) conforme aux exigences de la directive. Pour cela, les données du GIS Sol, diffusées dans ce cadre par INRAE, utilisent l'IDS agroenvgeo https://agroenvgeo.data.inra.fr/geonetwork/srv/fre/catalog.search#/home mise à disposition par INRAE. Elle est basée sur le logiciel Open source geOrchestra https://www.georchestra.org/fr/ qui respecte les standards de l'Open Geospatial Consortium (OGC). Elle comprend deux outils, Geoserver pour la diffusion de données spatiales et GeoNetwork pour la diffusion des métadonnées associées. Celles-ci sont moissonnées par le Géocatalogue https://www.geocatalogue.fr/, point d'accès aux métadonnées publiées par les autorités publiques françaises. Cette IDS publie des services web de consultation et de téléchargement. Une chaîne de traitement a été développée pour automatiser leur publication depuis le système d'information décisionnel.

Il existe actuellement 104 fiches de métadonnées associées à plus de 250 couches graphiques. Les couches graphiques sont affichables sur le visualiseur mais aussi dans un SIG par

l'utilisation de services web au format WMS. L'outil permet également le téléchargement des données au format shapefile ou TIFF selon la nature des données, ou leur importation directe dans un SIG par l'utilisation de services web au format WFS. Les données de la première campagne du RMQS ainsi que les référentiels régionaux pédologiques dont INRAE est propriétaire sont ainsi consultables dans l'IDS.

La diffusion de certains RRP par les partenaires régionaux utilise également ce mode de diffusion au travers d'IDS régionales. On peut citer, par exemple, la diffusion de RRP d'Aquitaine par Bordeaux Science Agro sur l'IDS régionale Pigma (https://portail.pigma.org/) ou ceux de la région Grand Est par la Chambre régionale d'agriculture sur l'IDS régionale DataGrandEst (https://www.datagrandest.fr/portail/fr).

4.4. Outil de consultation des données sols de l'IFN

Le site internet de l'Inventaire forestier national, accessible à l'adresse https://inventaire-forestier.ign.fr/, permet d'accéder à de nombreuses données sur la forêt et les écosystèmes forestiers selon 3 principaux vecteurs d'information :

- Les tableaux standards ;
- Les résultats ou tableaux personnalisés (outil OCRE) ; visualisation de la distribution cartographique des placettes répondant à certains critères (*via* les données brutes, outil DataIFN) et/ou le téléchargement des données brutes.

Les résultats d'inventaire le plus souvent demandés sont disponibles sous la forme de tableaux standards. Ils concernent essentiellement le volume de bois, les surfaces forestières et les flux : production, mortalité et prélèvement. Un outil permet d'obtenir des résultats personnalisés, sous la forme de tableaux, accessibles aux personnes extérieures à l'IGN. Dans sa version « grand public » (https://inventaire-forestier.ign.fr/?rubrique226), il est possible d'effectuer des tris des données forestières (volume, surface terrière, etc.) selon des paramètres pédologiques (type de sol, type d'humus, type de roche-mère, niveau hydrique, niveau trophique calculés à partir de la flore) ou des indicateurs de sensibilité du sol issus du projet INSENSE (Augusto *et al.*, 2018)) : indicateur de sensibilité du sol à l'export de biomasse ; indicateur de sensibilité du sol à l'export de calcium, de magnésium, de potassium, de phosphore et d'azote. Il est possible de calculer de nouveaux indicateurs pour répondre à d'autres besoins. Seuls les indicateurs ayant une validité nationale sont en accès grand public. Une version « Pro » de l'outil, réservée aux professionnels, permet un accès plus large et personnalisé en matière de données ou de périodes d'investigation. Son utilisation nécessite une formation, notamment pour accompagner l'utilisateur sur la validité et l'interprétation des résultats obtenus.

Les données brutes de l'Inventaire sur les sols forestiers peuvent être visualisées par l'intermédiaire de l'application DataIFN : https://inventaire-forestier.ign.fr/dataifn/. Cette application permet la visualisation de la distribution spatiale

Figure 2 : Placettes à caractère podzolique pour les campagnes d'inventaire de 2017 à 2021

Figure 2: *Podzolic plots for the surveys from 2017 to 2021*

N.B. : Les coordonnées géographiques fournies sont celles du centre de la maille kilométrique de la grille d'échantillonnage la plus proche du point (les coordonnées exactes de la placette d'échantillonnage se situent à 700 mètres maximum de ce centre).

des placettes répondant à certains critères. Les données correspondantes peuvent être téléchargées au format csv. A titre d'exemple, la *figure 2* montre les placettes qui présentent un sol à caractère podzolique pour les campagnes d'inventaire de 2017 à 2021.

Enfin, il est possible de télécharger l'ensemble des données brutes de l'Inventaire forestier depuis 2005, y compris les données écologiques et pédologiques, *via* un lien disponible sur la même page que celle de l'application DataIFN. Les coordonnées exactes des placettes, protégées par diverses législations et le poids statistique des points dans l'échantillon ne figurent pas dans ces fichiers. Les indices calculés par l'IFN à partir des données pédologiques et floristiques sont disponibles à l'adresse suivante : https://inventaire-forestier.ign.fr/?rubrique262. Outre les niveaux trophiques et hydriques, calculés à partir de la flore, les indices d'hydromorphie, de texture et de réserve utile sont aussi proposés.

4.5. Diffuser les données sur les sols urbains

Les premiers résultats issus du workflow de traitement des données BDSolU seront prochainement disponibles. Ils seront préétablis par le BRGM pour des zones géographiques présentant un nombre suffisant d'analyses pour permettre le calcul statistique des FPGA (Brunet *et al.*, 2023).

Les informations accessibles comprendront pour chaque élément ou substance analysé : des statistiques de base (moyenne, médiane, écart-type, quantiles…) ; des histogrammes et des boîtes de Tukey (Tukey, 1977) représentant la distribution de la population après sélection lors du workflow ; des cartes représentant les points de prélèvement à une échelle ne permettant pas leur localisation exacte (pour des raisons de confidentialité) ; des valeurs de fonds pédogéochimiques (sous réserve d'une population d'effectif supérieur à 30).

Une interface interactive permettra à terme de diffuser les valeurs de fonds pédogéochimiques anthropisés (www.bdsolu.fr). Il est prévu de fournir des valeurs de FPGA déterminées en temps réel : par zone géographique, selon diverses échelles possibles (quartier, ville, agglomération urbaine) ; par profondeur, selon les objectifs de l'étude conduite (enjeux sanitaires en cas d'exposition directe, valorisation de terres excavées, etc.), par paramètre (métaux, métalloïdes, composés organiques persistants). L'obtention en temps réel de ce type de carte fait appel à des méthodes d'interpolation statistiques et/ou géostatistiques et à des algorithmes experts qui sont en cours de développement.

4.6. Diffuser les données en open data : vers l'interopérabilité

Dans le cadre de sa politique de science ouverte, INRAE s'est doté, en 2018, d'un portail de partage de données pour faciliter la gestion, le partage et la recherche des données et répondre aux obligations réglementaires françaises en matière d'ouverture des données. Ce portail repose sur la technologie dataverse. Il est, depuis 2022, inséré dans le portail des données de la recherche française (https://entrepot.recherche.data.gouv.fr).

Le GIS Sol a créé, dès 2018, une collection, dossier dans lequel on publie des jeux de données ou datasets (https://entrepot.recherche.data.gouv.fr/dataverse/gissol). En août 2022, 21 jeux de données y ont été publiés, un jeu de données pouvant contenir un ou plusieurs fichiers ou seulement indiquer un lien vers des fichiers (comme par exemple vers l'IDS).

Un des avantages de ce portail est la création automatique de DOI (*digital object identifier*) qui aident à mieux identifier et citer les jeux de données (publications scientifiques, projets opérationnels...) et tracer leurs réutilisations. Cependant, pour les données géographiques, la publication dans la collection et dans l'IDS pose des problèmes d'hétérogénéité entre les deux modes de diffusion, due à des problématiques organisationnelles et techniques : hétérogénéité dans les formats de diffusion (par exemple GeoTIFF dans la collection GIS Sol et standards WMS/WFS dans l'IDS), hétérogénéité dans les métadonnées et entre les versions des jeux de données. Un travail d'homogénéisation est en cours ainsi que le développement d'un workflow de publication qui permettra de publier de manière homogène les données et les métadonnées sur la collection GIS Sol et sur l'IDS.

Mettre des fichiers en téléchargement ne permet pas de remplir toutes les obligations relatives à l'ouverture des données. En effet, il est attendu que ces données répondent aux principes FAIR : facile à trouver, accessible, interopérable et réutilisable. Pour cela, il est nécessaire d'utiliser des technologies qui permettent une réelle interopérabilité. L'IDS permet l'interopérabilité des données spatiales mais il reste du travail pour y parvenir pleinement, notamment sur la description des propriétés observées. Cette préoccupation rejoint celle de la diffusion de données tabulaires. Deux pistes sont explorées actuellement par l'Unité Info&Sols d'INRAE :

- l'utilisation de l'API SensorThings (OGC, 2022) qui permet de diffuser des données selon le standard Observations and Measurements de l'OGC.
- l'utilisation des technologies du web sémantique selon les standards du W3C où sont repris des outils (pipeline Coby) développés dans le cadre de l'infrastructure nationale sur les analyses et expérimentations sur les écosystèmes continentaux (AnaEE France).

L'API SensorThings a été développée pour apporter une structure standardisée permettant l'interconnexion des données issues de capteurs, de données ou d'applications web. Elle facilite également l'intégration des données dans des infrastructures de données spatialisées ou des systèmes d'information géographique préexistants. Pour publier les données du GIS Sol *via* l'API SensorThings, l'Unité Info&Sols d'INRAE a développé ses propres outils, libres et open-source, restant ainsi dans la philosophie de l'open data *(figure 1)* :

- SensorMap : pour structurer les données issues de la base de données DoneSol au format requis par l'API. Les données sont extraites de la base, transformées et publiées vers un serveur Frost sous forme de JSON, utilisable par des machines.
- SensorBoard : pour visualiser les données issues du serveur Frost dans un format lisible et intelligible par les humains, et en permettre une représentation cartographique, une édition, création ou suppression de données.
- MapGo : pour fédérer des données au format SensorThings issues de différentes bases et hébergées sur différents serveurs au travers d'une interface cartographique.

Un test a été réalisé montrant la capacité d'interopérabilité des données de DoneSol, de l'IRD et du Cirad *via* l'API Sensorthings autour des données sur le carbone du sol en Guyane (projet ANR Data4C+, 2018-2021, coordination Cirad, https://www.data4c-plus-project.fr/). Un autre test d'interopérabilité des données DoneSol a été réalisé avec la base de données BDSolU (projet FGU III - Etablissement de fonds pédogéochimiques urbains, convention Ademe, 2018-2023, coordination BRGM et projet SUPRA - Sols Urbains et Projets d'Aménagement, de l'échantillonnage des sols à l'outil d'aide à la décision d'affectation des sols, 2018-2021, coordination Laboratoire Sols et Environnement de l'Université de Lorraine) (SUPRA, 2021). En effet, les deux bases de données sont complémentaires du point de vue des analyses réalisées, DoneSol pouvant contenir des analyses de substances potentiellement polluantes et BDSolU des analyses sur la qualité des sols agricoles ou sur leur capacité à stocker du carbone, mais aussi du point de vue géographique, DoneSol couvrant plutôt des espaces ruraux et BDSolU des espaces urbains. L'objectif est de mettre en place un langage commun aux deux bases de données afin de permettre leur interrogation simultanée. L'utilisation de SensorThings permet de décrire les données *via* le même modèle de données, avec des recodages en amont pour que les propriétés observées soient décrites de la même manière.

Le web sémantique, tel que défini par le W3C qui en développe les technologies, fournit une structure commune qui permet aux données d'être partagées et réutilisées malgré les frontières entre les applications, les entreprises et les communautés. La diffusion des données selon le web sémantique repose sur le modèle standard RDF (*Resource Description Framework*). Ce modèle se présente sous la forme de triplet composé d'un sujet, d'un prédicat et d'un objet : le sujet représente la ressource à décrire (*e.g.* un profil), l'objet est une autre ressource permettant de décrire le sujet (*e.g.* un horizon), le prédicat détermine en quoi l'objet décrit le sujet, à savoir la relation qui lie les deux ressources entre elles, ainsi que la direction de cette relation (*e.g.* contient). Chaque élément du triplet est relié à un identifiant unique, l'URI (Uniform Resource Identifier) qui permet à l'ensemble des triplets RDF de définir un graphe orienté et annoté dans lequel sujets, prédicats et objets sont définis, décrits et réutilisables. Pour permettre une description standardisée des données dans le web sémantique, une ontologie adaptée doit être développée pour modéliser les concepts associés aux données. La diffusion des données du GIS Sol dans le web sémantique est en cours. Il s'appuie technologiquement sur les outils développés dans le cadre d'AnaEE France, notamment fondé sur le pipeline d'annotation sémantique Coby qui, à partir d'une modélisation des données sous forme d'annotation sémantique, les publie dans un système de gestion de graphes de données (Blazegraph) et les stocke sous forme de triplets RDF dans un « triplestore ». La modélisation des données et l'ontologie à utiliser sont des points essentiels qui sont en cours d'élaboration, car il n'existe actuellement pas d'ontologie spécifique aux données sur les sols. Ces travaux s'appuient sur des standards technologiques favorisant l'interopérabilité, et fournissent une ou plusieurs implémentations sous forme de projets open-source, leurs buts étant la simplification du partage et de la réutilisation des données entre plusieurs applications et/ou systèmes d'information.

L'intérêt de ces technologies est de pouvoir diffuser les données avec leur description (métadonnées) permettant ainsi une totale interopérabilité. Un effort amont est nécessaire pour que ces métadonnées soient non seulement les plus précises et détaillées possibles, mais également pour qu'elles s'appuient sur des vocabulaires contrôlés et des ontologies dans le cadre du web sémantique.

5. CONCLUSION ET PERSPECTIVES

En plus de 20 ans, la connaissance sur les sols s'est considérablement améliorée grâce aux actions du GIS Sol. Les efforts ont été faits en matière d'acquisition mais aussi de gestion, d'analyse et de diffusion des données. Ils ont permis d'augmenter l'intérêt de multiples acteurs pour les données sur les sols. Des efforts subsistent pour répondre aux enjeux actuels. En matière de capitalisation, de nouveaux types de données sont à prendre en compte, comme les données de biodiversité ou celles issues des sciences participatives. L'intérêt pour les sols urbains est grandissant et demandera un effort important de capitalisation pour pouvoir répondre aux besoins des métropoles. Les premiers travaux de cartographie à modélisation statistique rencontrent un fort intérêt des utilisateurs, parce qu'à échelles fines, mais l'offre est encore trop faible. La mise à disposition des données est encore vue comme trop complexe et trop dispersée. Le développement et la maintenance des bases de données, et le développement de l'opérationnalité des workflows de traitement doivent se poursuivre pour augmenter l'offre de données traitées et améliorer la diffusion des données, pour les rendre plus FAIR, pour progresser dans une offre plus générique, moins attachée à chacun des programmes d'acquisition des données du GIS Sol. Cela nécessite des moyens conséquents dans un contexte d'évolution des technologies toujours plus rapide.

REMERCIEMENTS

Les auteurs adressent leurs remerciements à l'ensemble des partenaires du GIS Sol pour leur contribution à la collecte, la gestion, l'analyse et la diffusion des données du GIS Sol.

Ils tiennent également à remercier les relecteurs François Hissel et Hervé Squividant pour leur relecture attentive et leurs commentaires ayant permis d'améliorer significativement le manuscrit de cet article.

BIBLIOGRAPHIE

Ademe (2018). Méthodologie de détermination des valeurs de fonds dans les sols : Echelle d'un site/d'un territoire. Groupe de travail sur les valeurs de fonds. 107 p. https://librairie.ademe.fr/sols-pollues/32-guide-pour-la-determination-des-valeurs-de-fonds-dans-les-sols-echelles-d-un-territoire-d-un-site.html

Augusto L., Pousse N., Legout A., Seynave I., Jabiol B., Levillain J. (2018). INSENSE : Indicateurs de SENSibilité des Ecosystèmes forestiers soumis à une récolte accrue de biomasse. Rapport de recherche. ADEME. 262 p. [hal-03023040]

Arrouays D., Stengel P., Feix I., Lesaffre B., Morard V., Bardy M., Bispo A., Laroche B., Caquet T., Juille F., Rabut M., Soussana J.-F., Voltz M., Gascuel-Odoux C. (2022). Le GIS Sol, sa genèse et son évolution au cours des vingt dernières années, Étude et Gestion des Sols, 29, 365-379. https://hal.inrae.fr/hal-03815433.

Baize D., Saby N.P.A., Bispo A., Feix I. (2006). Analyses totales et pseudo-totales d'éléments en traces dans les sols - Principaux résultats et enseignements d'une collecte nationale. Étude et Gestion des Sols, 13, 3, pp. 181-200. https://hal.inrae.fr/hal-02655076.

Beaudou A., Le Martret H. (2004). MIRURAM/VALSOL : un système d'information et une base de données pour représenter les sols tropicaux et leurs environnements. Étude et Gestion des Sols, 11, 3, pp. 271-284.

Belbèze S., Rohmer J., Négrel P., Guyonnet D. (2023). Defining urban soil geochemical backgrounds: A review for application to the French context. Journal of Geochemical Exploration, 254, 107298. https://doi.org/10.1016/j.gexplo.2023.107298

Bertrand R., Falipou P., Legros J.-P. (1979) Notice pour l'entrée des descriptions et analyses de sols en banque de données. STIPA 1979. INRA - IRAT, Montpellier. 119 p.

Brossard M., Fujisaki K., Jolivet C., Dupuits-Bonnin E., Jameux M., Jalabert S., Toulemone Le Ny E., Becquer T., Blavet D., Beaudou A., Boulonne L., Desjardins T., Le Martret H., Ratié C. (2023). Le GIS Sol dans les départements et régions d'Outre-mer français. Etude et Gestion des Sols, 30, pp. 145-168. https://hal.inrae.fr/hal-03962807

Brunet J.-F., Branchu P., Eychene C., Belbèze S., Guyonnet D. (2023). L'offre du GIS Sol aux politiques d'aménagement urbain, Étude et Gestion des Sols, 30, pp. 195-206, https://hal.science/hal-04056942/

Chapuis A., Toutain B., Chenu J.-P., Laroche B. (2013). Sivercoh, une application web pour vérifier la cohérence des bases de données pédologiques. Séminaire National IGCS « Inventaire, Gestion et Conservation des Sols », Rennes, France, 11-13/12/2013. https://hal.inrae.fr/hal-03753031

Chen S., Arrouays D., Mulder V.L., Poggio L., Minasny B., Roudier P., Libohova Z., Lagacherie P., Shi Z., Hannam J., Meersmans J., Richer-de-Forges A.C., Walter C. (2022). Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma, 409, 115567. https://doi.org/10.1016/j.geoderma.2021.115567

De Gruijter J., Brus D.J., Bierkens M.F., Knotters M. (2006). Sampling for natural resource monitoring. Springer Science & Business Media. 334 p. https://doi.org/10.1007/3-540-33161-1.

Duchaufour P. (1991). Pédologie. Sol, végétation, environnement, troisième édition. Paris, Masson, 289 p.

Fouquet (de) C. (2021). Vibrisse de Tukey et limites de quantification - Notes Mines Paris-Tech établie dans le cadre du projet « Etablissement de fonds pédogéochimique urbain (FGU) - 3e convention 2019-2023). Ademe, 8 p.

Gaultier J.-P., Legros J.-P., Bornand M., King D., Favrot J.-C., Hardy R. (1993). L'organisation et la gestion des données pédologiques spatialisées : le projet DONESOL. Revue de géomatique, 3, 3, pp. 235-253. https://hal.science/hal-02703961v1

Girot G., Millet F., Schnebelen N., Cousin I., Toutain B., Bardy M. (2017). Applicasol, un outil de partage des applications thématiques sur les sols - Etude et Gestion des Sols, 24, pp. 33-43 https://hal.inrae.fr/INFOSOL/hal-02619353

Grolleau E., Bargeot L., Chafchafi A., Hardy R., Doux J., Beaudou A., Le Martret H., Lacassin J.-C., Fort J.-L., Falipou P., Arrouays D. (2004). Le système d'information national sur les sols : DONESOL et les outils associés. Etude et gestion des sols, 11, 3, pp. 255-269.

Helsel D. (2005). Nondetects And Data Analysis: Statistics for Censored Environmental Data. John Wiley & Sons.250 p.

Jabiol B., Brethes A., Ponge J.-F., Toutain F., Brun J.-J. (1995). L'humus sous toutes ses formes, première édition. ENGREF, 64 p.

Le Martret H., Beaudou A., Blanca Y., Brossard M., Le Rouget Zuritta B. (2008). Interface web cartographique de visualisation des données pédologiques. Le Monde des Cartes : Revue du CFC, 197, pp. 31-36. fdi:010044925

Lehmann J., Bossio D.A., Kögel-Knabner I., Rillig M. C. (2020). The concept and future prospects of soil health. Nat Rev Earth Environ, 1, pp. 544–553. https://doi.org/10.1038/s43017-020-0080-8

Loiseau T., Richer-de-Forges A.C., Roudier P., Ducommun C., Chen S., Lagacherie P., Arrouays D. (2020). Essais de représentation cartographique de l'incertitude pour les utilisateurs de cartographies des sols par modélisation statistique. Etude et Gestion des Sols, 27, pp. 257-275. https://hal.inrae.fr/INFOSOL/hal-03141430

MEEM (2018). Méthodologie nationale de gestion des sites et sols pollués. Ministère de l'Environnement, de l'Energie et de la Mer. http://ssp-infoterre.brgm.fr/methodologie-nationale-gestion-sites-sols-pollues

MEFSIN (2022). Fonds pour le recyclage des friches. Ministère de l'Economie, des Finances et de la Souveraineté industrielle et numérique. https://www.economie.gouv.fr/plan-de-relance/mesures/fonds-recyclage-friches

Messant A., Lehmann S., Moulin J., Lagacherie P., Jalabert S., Noraz A., Lemercier B., Chafchafi A., Mure J.-P., Laroche B., Sauter J. (2021). Diffusion des Référentiels Régionaux Pédologiques sous la forme d'une carte des sols dominants (France métropolitaine hors-Corse) accessible sur le Géoportail. Etude et Gestion des Sols, 28, pp. 57-69 https://hal.inrae.fr/INFOSOL/hal-03121194

MTECT (2022). Sites et sols pollué. L'article 173 de la loi ALUR. Ministère de la Transition écologique et de la Cohésion des territoires. https://www.ecologie.gouv.fr/sites-et-sols-pollues#scroll-nav__5

OGC (2022). SensorThings – Introduction. http://opengeospatial.github.io/e-learning/sta/text/main.html

Rennes S., Le Bas C., Le Bideau S., Hissel F. (2023). Prendre en compte le statut juridique des données dans les programmes du GIS Sol. Etude et Gestion des Sols, 30, pp. 253-262. https://hal.inrae.fr/hal-04018933

Richer-de-Forges A.C., Laroche B., Lehmann S., Chenu J.-P., Boukir H., Perrier C. (2018). Les offres de formations à DoneSol. 14e Journées d'Etude des Sols : « Le sol au cœur des enjeux sociétaux », Rouen, France, 9-12 juillet 2018. https://hal.inrae.fr/hal-02738427

Román Dobarco M., Bourennane H., Arrouays D., Saby N.P.A., Cousin I., Martin M.P. (2019). Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma, 344, pp. 14–30. https://doi.org/10.1016/j.geoderma.2019.02.036

Saby N.P.A., Lemercier B., Arrouays D., Leménager S., Louis B., Millet F., Paroissien J.-B., Schellenberger E., Squividant H., Swiderski C., Toutain B., Walter C., Bardy M. (2014). Le programme Base de Données des

Analyses de Terre (BDAT) : Bilan de 20 ans de collecte de résultats d'analyses. Etude et Gestion des Sols, 21, pp. 141-150. https://hal.science/hal-01209243

Saby N.P.A., Marchant B.P., Arrouays D., Jolivet C. (2018). Spatial predictions of total and exchangeable trace elements content in France. https://doi.org/10.15454/VN9F6H, Recherche Data Gouv, V3

Sauvaget B., de Fouquet C., Le Guern C., Renard D., Roussel H. (2022). Geostatistical filtering to map a 3D anthropogenic pedo-geochemical background for excavated soil reuse. Journal of Geochemical exploration, 240, 107031, https://doi.org/10.1016/j.gexplo.2022.107031

Supra (2021). Sols urbains : les caractériser pour aider la décision de leur affection lors des projets d'aménagement. https://expertises.ademe.fr/content/supra-sols-urbains-projets-damenagement

Toutain B. (2013). Evolution du système d'information Donesol. Séminaire IGCS (Inventaire Gestion et Conservation des Sols), Rennes, France, 11-13 décembre 2013. 14 p. https://hal.inrae.fr/hal-02805037

Tukey J.W. (1977). Exploratory Data Analysis. Addison-Wesley Publishing Company. 688 p.

Vie Publique (2015). Loi du 7 août 2015 portant nouvelle organisation territoriale de la République. https://www.vie-publique.fr/loi/20721-loi-notre-loi-du-7-aout-2015-nouvelle-organisation-territoriale-de-la

Vinatier J.-M., Chafchafi A., Bargeot L, Toutain B., Laroche B., Arrouays D., Squividant H. (2013). Websol : une plateforme Internet de diffusion des données pédologiques. Etude et Gestion des Sols, 20, 3, pp. 7-18. https://hal.science/hal-01209115

Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L. B., Bourne P.E., Bouwman J., Brookes A.J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C.T., Finkers R., Gonzalez-Beltran A., Gray A.J.G., Groth P., Goble C., Grethe J.S., Heringa J., Hoen P.A.C.'t, Hooft R., Kuhn T., Kok R., Kok J., Lusher S.J., Martone M.E., Mons A., Packer A.L., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S.-A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M.A., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 1, 160018. https://doi.org/10.1038/sdata.2016.18