

Webinaire données

Systemes d'information géographiques

Expériences en DREAL Hauts-de-France

Plan

1. Notions (pas si) théoriques
2. Organisation pratique
3. Organisation de la DREAL Hauts-de-France avec une solution numérique transverse : SIGNE
4. Ouvrons le capot : cas pratiques



Notions théoriques



Un peu de vocabulaire pour commencer

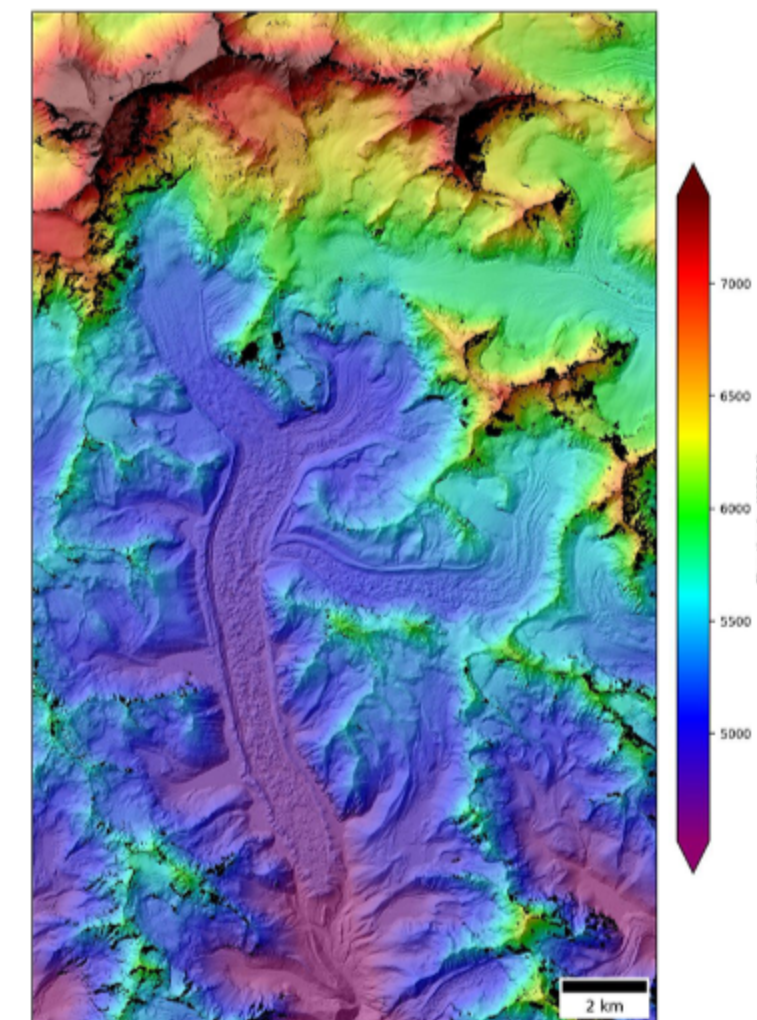
La géomatique, ce n'est pas (que) de la cartographie !

La géomatique regroupe l'ensemble des connaissances et des technologies nécessaires à la production, au traitement et à la diffusion des données numériques (souvent massives, 3D et en temps réel) décrivant le territoire, ses ressources ou tout autre objet ou phénomène ayant une position géographique.

Qu'est-ce que la géomatique? Ses racines sont "Géo", qui veut dire Terre, et "matique" vient de informatique, soit le traitement automatique de l'information géographique.

Méconnue du grand public, la géomatique est pourtant de plus en plus présente dans plusieurs de nos activités quotidiennes. Quatre-vingt pourcent des données ont une position sur la Terre : imaginez les possibilités!

Université de Laval, Faculté de foresterie, de géographie et de géomatique, en ligne 



Source : Eyes Roger & NSIDC, via Wikimedia
licence CC Attribution-2.0 generic 

Vous avez dit SIG ?

SIG = Système d'Information Géographique

Le SIG est un domaine relativement nouveau, qui est apparu dans les années 1970. Le SIG, utilisant l'informatique, n'était disponible et utilisé que par des entreprises et des universités qui avait un équipement informatique qui coutait cher. De nos jours, quelqu'un ayant un ordinateur individuel ou un ordinateur portable peut utiliser un logiciel SIG. Au fil du temps, les applications SIG sont aussi devenues plus faciles à utiliser — cela exigeait beaucoup de formation pour utiliser une application SIG, mais maintenant il est beaucoup plus facile de commencer à utiliser un SIG, même pour des amateurs et des utilisateurs occasionnels.

Gentle GIS Introduction (QGIS Project, v.3.28), en ligne 



Source : Daniele Masi, via Wikimedia licence CC
Attribution-Share alike 4.0 international 

On parle donc de SIG (système d'information géographique), de logiciel SIG, de SIG-iste (géomaticien en bon Français).

Les métiers traditionnels

le géomaticien



C'est lui qui traite les données ; il transforme la donnée et se charge de la visualisation. Il produit des couches de données (digitalisation, traitement SIG, etc.), des cartes statiques (images) ou dynamiques (PDF, web) et parfois des packages QGIS. En pratique, il doit aujourd'hui connaître les données métier et savoir où trouver la donnée pour être force de proposition.

l'administrateur des données locales



Il se charge de centraliser les données locales sur un serveur commun et d'acquérir les données de référence auprès des producteurs de données. Il est le garant de la qualité des données (nettoyage des géométries, etc.). C'est aussi lui qui assure la diffusion des données géographiques (en lien avec la géoplateforme régionale).

Les outils métiers

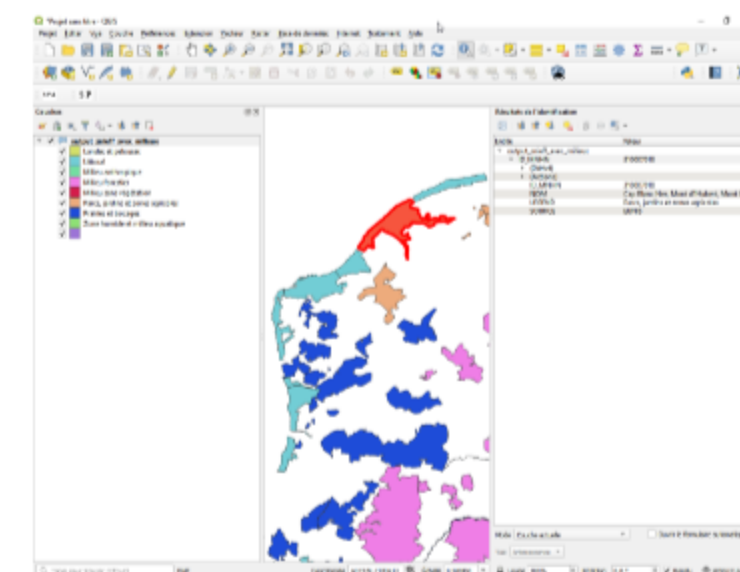
La géomatique utilise des outils métiers qui peuvent être très pointus, mais aussi très abordables (pour certains).

- outils sous licence : ArcGis, MapInfo...
- écosystème libre : QGIS (et toute la suite : SAGA, GRASS,...), Google Earth...
- des langages de programmation : Python, R, javascript
- du web (avec une pléthore de librairies javascript dédiées à la cartographie en ligne)

Pour aller plus loin, on trouve des benchmarks **en ligne**  , pour l'essentiel en anglais GB (comme souvent dans le domaine du libre).

Aujourd'hui, la plupart des géomaticiens du pôle ministériel travaillent certainement sur des outils libres.

Chaque outil a sa spécificité ; on choisit l'outil en fonction de l'usage (cartographie, digitalisation, modélisation de terrain ou de processus) et du format des données (vecteur ou raster).



Copie d'écran QGIS

Encore une notion théorique, celle de projection

Chaque couche de donnée géographique est établie dans un **système de coordonnées**.

La projection officielle en France métropolitaine est le **Lambert 93**, mais de nombreuses autres projections existent (par exemple, ce n'est pas la référence pour les DOM/TOM). Plus d'une trentaine sont référencés [sur wikipedia](#) pour la France métropolitaine. Plus de documentation technique sur les projections françaises sur [ce document](#) de l'IGN.



Source : Sting, via Wikimedia licence CC
Attribution-Share alike 2.5 Generic

Toutes les projections ne s'expriment pas en mètres (incidences sur les arrondis, les calculs de distance, d'aire...) et la plupart des algorithmes spatiaux nécessitent une projection uniforme.

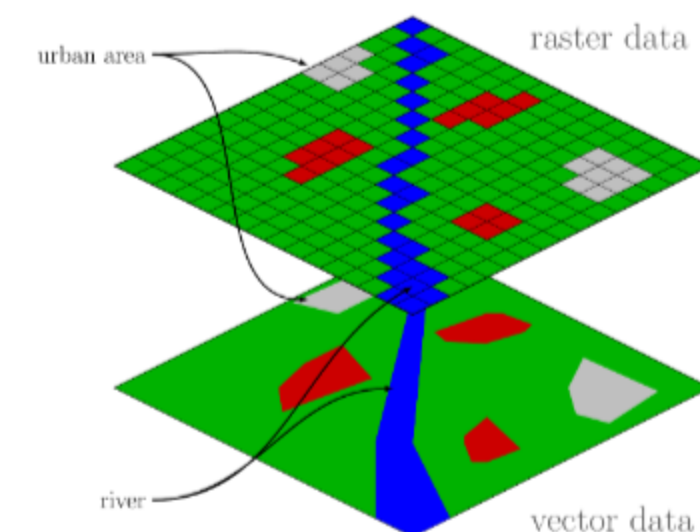
Dans le jargon, on utilise souvent le code "EPSG" (*European Petroleum Survey Group*) pour identifier un système de projection, mais tous les systèmes de projection n'ont pas un code EPSG.

Vecteur ou raster ?

Ce sont deux concepts différents de représentation d'une même donnée en deux dimensions :

- Raster = matrice de pixels, nombre variable de "bandes"
- Vecteur : des géométries point, lignes ou polygones (et des données tabulaires associées)

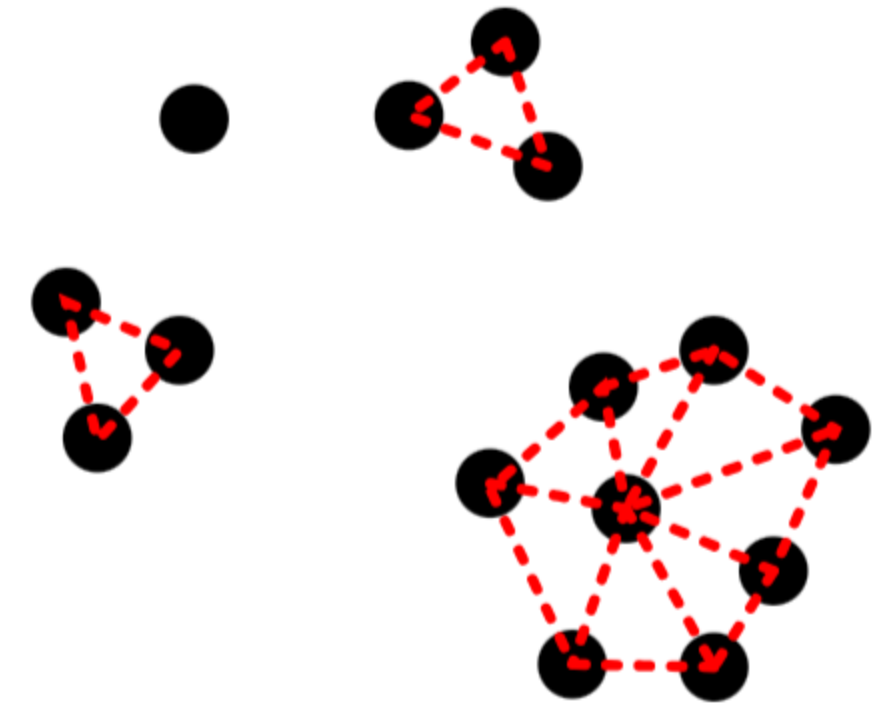
*Nota : ce n'est pas un sujet limité à la cartographie (on a exactement le même sujet en dessin). Des ressources supplémentaires **en ligne** *



Source : Wegmann, via Wikimedia licence CC
Attribution-Share alike 3.0 

Que peut faire une équipe SIG d'un format vecteur ?

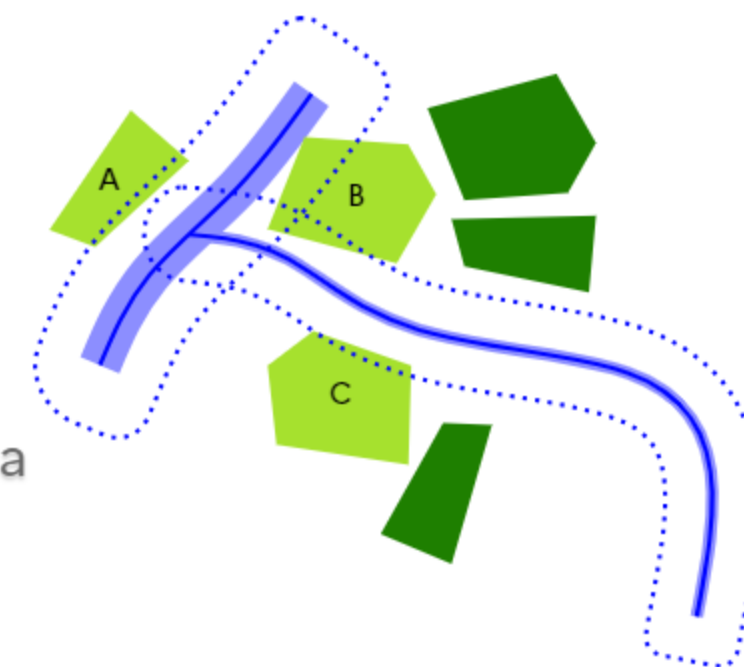
- Opérations géométriques unitaires : calcul d'aire, de longueur, ou de distances, calcul de centroïde, triangulation de Delaunay, enveloppe concaves ou convexes, rotations, transformations affines
- Opérations géométriques binaires (liste extensive des **commandes geopandas**  , en anglais) : *contains*, *crosses*, *disjoint*, *geom_equals*, *geom_almost_equals*, *geom_equals_exact*, *intersects*, *overlaps*, *touches*, *within*, *covers*, *covered_by*
- Opérations spatiales ensemblistes : intersection, différence, différence symétrique, union, ...
- Opérations de bases de données : union (sur la base d'un champ), explosion
- Intelligence artificielle : clustering, etc.



Exemple de cas d'usage

Trouver les ilots culturels à moins de 5 mètres d'un cours d'eau "BCAE"

1. récupérer les jeux de données BCAE et PAC
2. joindre la couche des cours d'eau BCAE à une donnée de référence (BD Topage, tronçons hydrographiques) par jointure relationnelle et/ou spatiale
3. s'assurer que la couche obtenue des cours d'eau BCAE est dans un système de projection exprimé en mètres
4. donner une épaisseur au cours d'eau (à partir des classes de largeur définies par la BD Topage) en effectuant un tampon pour obtenir des "polygones hydrographiques"
5. effectuer un tampon des "polygones hydrographiques" (5 mètres)
6. s'assurer de l'unicité du système de projection entre les polygones et les ilots cultureux
7. effectuer une jointure spatiale entre ilots cultureux et le tampon des polygones

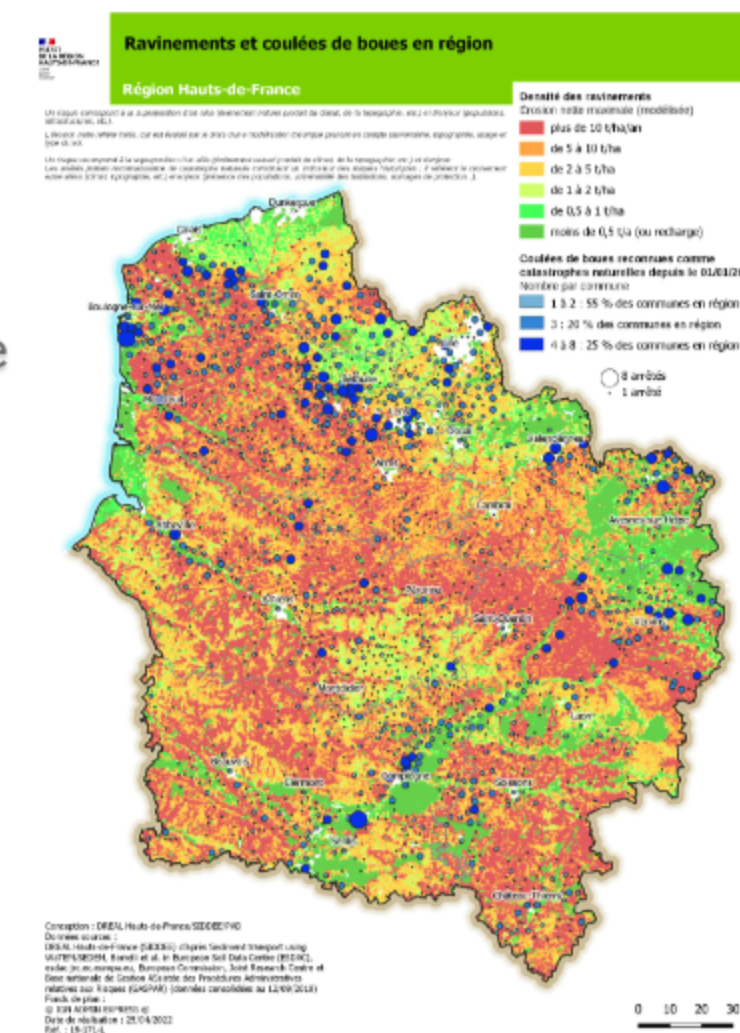


Sélection de parcelles par requête spatiale et tampon

Autre exemple 🌀 avec un projet personnel exploratoire impliquant un peu d'IA : analyse du front d'expansion du frelon asiatique en France.

Que peut faire une équipe SIG d'un format raster ?

- Fond de plan (3 ou 4 "bandes" : Red Green Blue Alpha) et ombrages (modèle numérique de terrain)
- Données numériques servant de couches d'entrée pour une modélisation physique :
 - Hydrographie : calcul de bassins versants à partir d'un "modèle numérique de terrain"
 - Pluviométrie : calcul d'une lame d'eau sur une superficie à partir d'images radar
 - Feux de forêt : modélisation du risque d'incendie (couches d'occupation du sol en herbacées, en forêt, modèle numérique de terrain, vitesse du vent...)
- Image satellitaire et usage émergent en intelligence artificielle : détection de zones inondées, pilotage de la politique de l'eau (humidité des sols : agriculture, aléa argiles...), droit des sols (ex. **détection des piscines**), occupation et usage des sols, base de données nationale des bâtiments **par le CSTB** ...

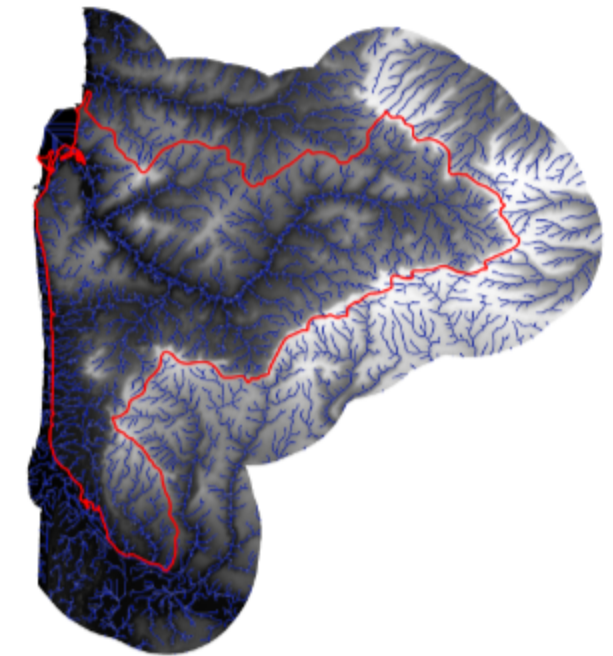


Source : DREAL Hauts-de-France, utilisation
d'une couche raster comme fonds de plan
atypique pour illustrer le risque de coulées de
boues

Exemple de cas d'usage

Utiliser les modèles numériques de terrain pour identifier une erreur dans le découpage des masses d'eau

1. récupérer les modèles de terrain de l'IGN
2. récupération du découpage (vecteur) du bassin existant et lui ajouter un tampon de plusieurs kilomètres
3. masque du MNT autour du tampon du bassin existant
4. post-traitement du MNT (traitement des "plats", et points bas)
5. calcul des directions d'écoulement
6. calcul de l'accumulation (valeur numérique = nombre de pixels en amont de chaque pixel)
7. repérage des thalwegs sur le raster (par exemple filtre des pixels sur la base d'un percentile 99 des valeurs d'accumulation)
8. vectorisation des thalwegs (polylignes) et repérage des exutoires
9. recalcul des bassins versant à partir d'exutoires judicieusement choisis



Identification des thalwegs sur le bassin versant
de la Liane

Vecteur : attention aux géométries

Une couche vecteur pose la question de la validité des géométries. Voir par exemple le guide [accompagnement à la détection et la correction des erreurs de géométrie dans les données d'urbanisme](#) .

Absence de géométrie

Polygones non fermés

Surfaces nulles

Noeuds trop proches ou en doublons

(apprécié à l'aune de la précision des
valeurs numériques...)

Arcs pendants



Polygones en papillon



Contour de polygone se refermant sur
lui-même



Contours de polygone se touchant



Des outils existent pour aider les équipes de géomaticiens à corriger ces géométries. De manière assez suprenante, ces opérations de nettoyage doivent aussi être conduites sur les jeux de données de référence...

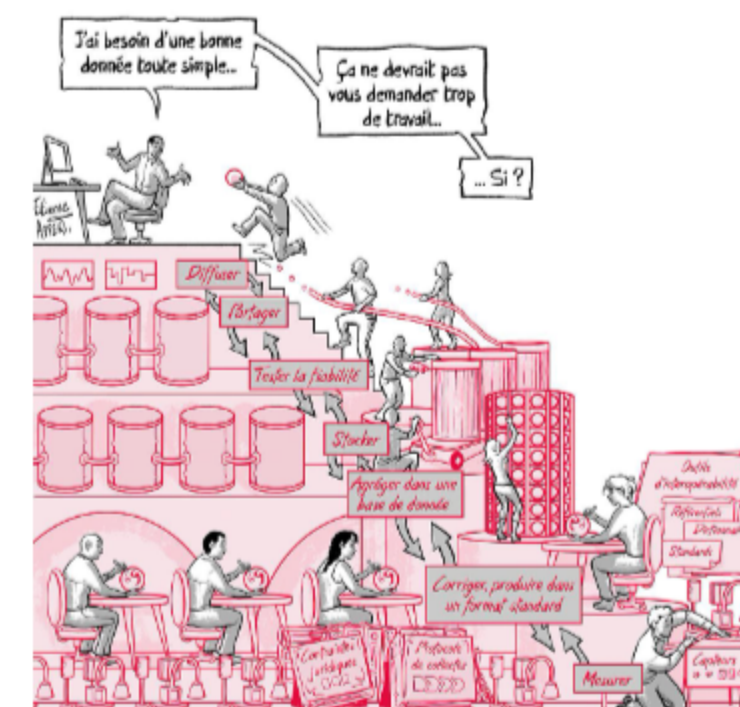
*Nota : **certaines** vont avoir des effets de bord : faut-il préserver la topologie ou la géométrie apparente ?*

Un dernier rappel théorique

L'équipe SIG est rarement le producteur initial de la donnée, même s'il en produit.
Pour de bons résultats, il faut donc disposer de données **de qualité** 📎.

En vrac, parmi les sujets de la qualité qui reviennent fréquemment dans un service SIG de service déconcentré :

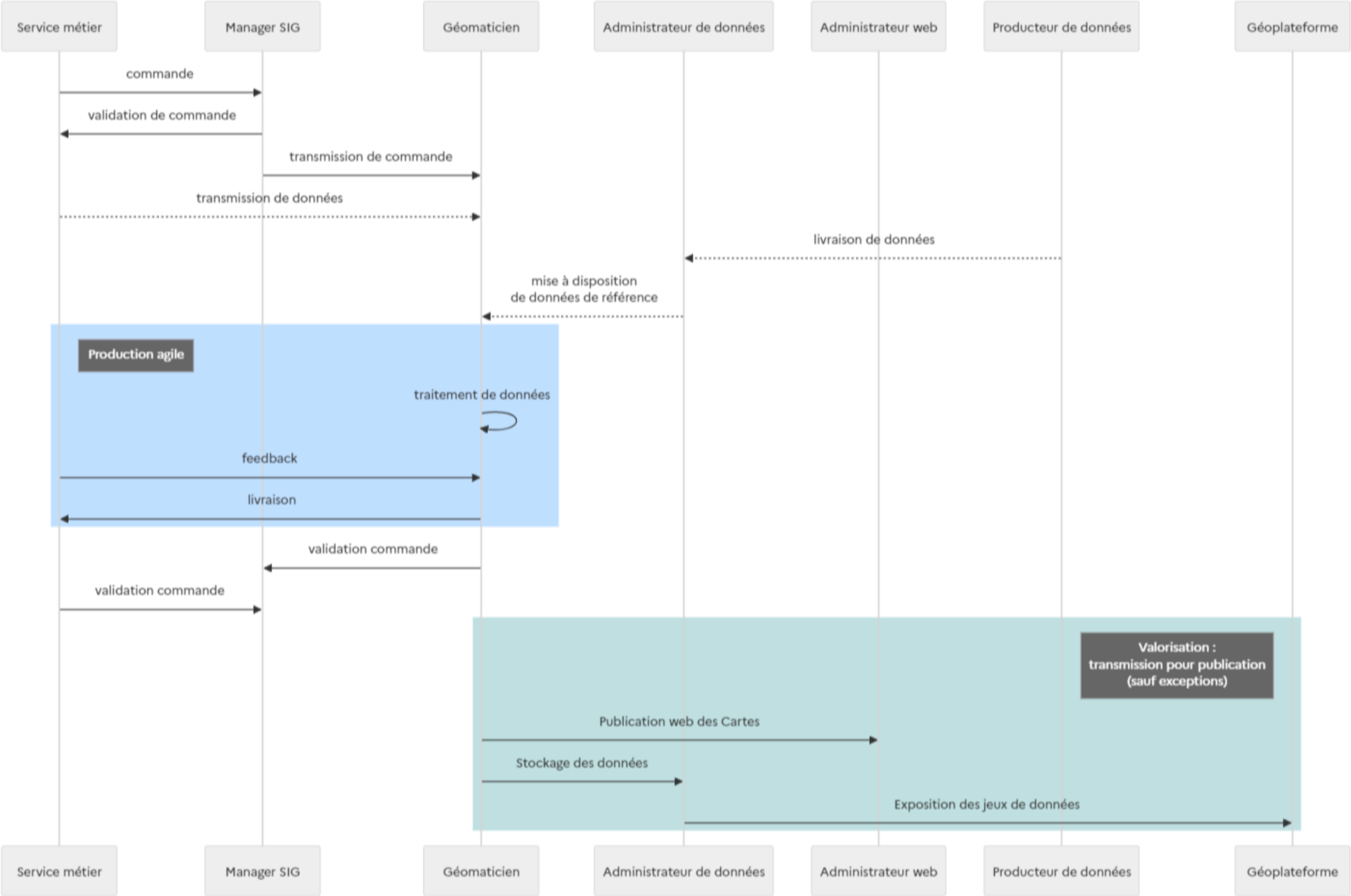
- la validité de la géométrie,
- la projection géographique,
- le millésimage des codes commune (derniers développements en Hauts-de-France, un package python 📎 en test pour re-millésimer un jeu de données multi-millésimé)
- le typage des champs (texte ou numérique : "02001" ≠ 2001 !),
- l'encodage des fichiers (UTF8, Windows-1252, y compris dans les jeux de données de référence...)
- ...



Source : Principes généraux de qualité des données, dessins d'Étienne Appert 📎

Organisation pratique

Organisation classique d'une équipe SIG (prises de commande)

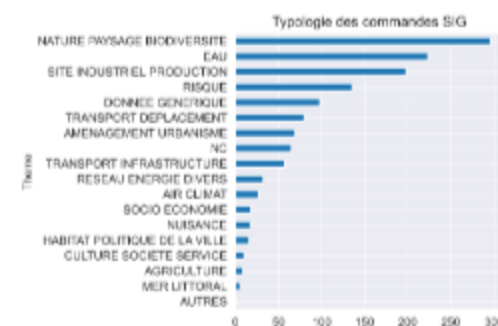


Quelles commandes pour le SIG ?



Part des commanditaires SIG

Une offre de service très tournée vers la DREAL



Typologie des commandes

Une production dominée par trois thématiques : milieu naturel, paysage et risques



Nuage de mots sur l'objet des commandes

Et surtout, une production dominée par les spécificités locales



Nuage de mots sur l'objet des commandes (épuré)

En gommant les spécificités locales, on voit bien réapparaître la transversalité des missions

Évolution des commandes

Combinaison de facteurs :

- autoformation des services métier durant les confinements ;
- montée en puissance de SIGNE ;
- développement de cartographies dynamiques mises à jour en continu (ex. **carte dynamique éolien** 📍 sur GéoIDE).

Une évolution des commandes qui ne reflète pas exactement l'évolution du plan de charge (mutation des commandes).

Globalement une baisse des commandes qui permet tout de même de trouver le temps d'accompagner la mutation du métier (autoformation python, javascript).



Évolution des commandes SIG

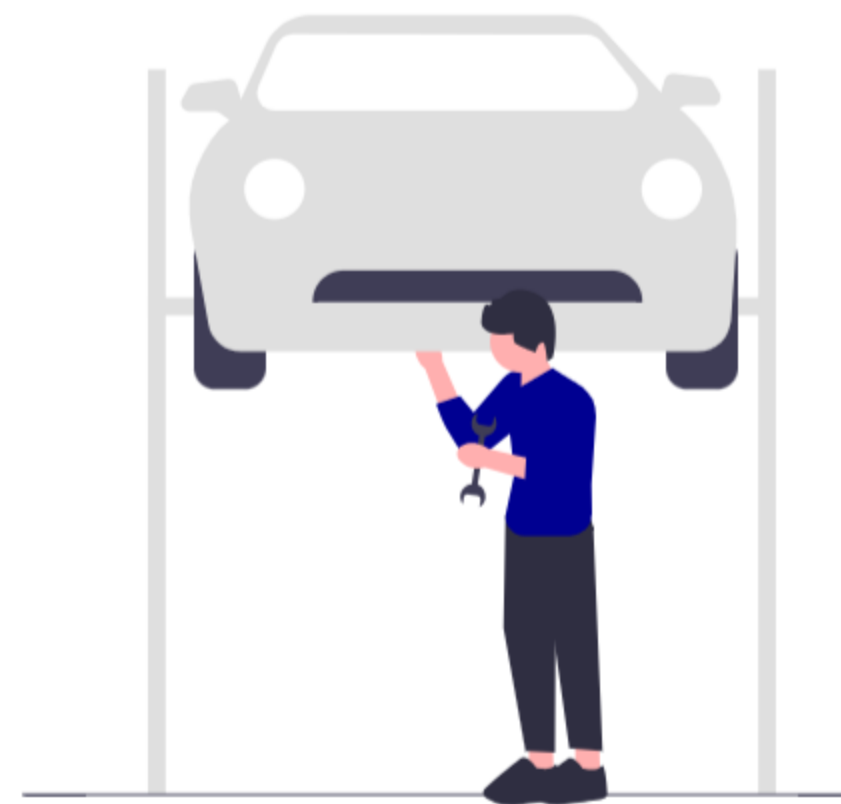
SIGNE

On change d'interlocuteur !



Ouvrons le capot de SIGNE

Suivi de la chaîne de la donnée autour d'un jeu de données particulier : GUN & Infosols

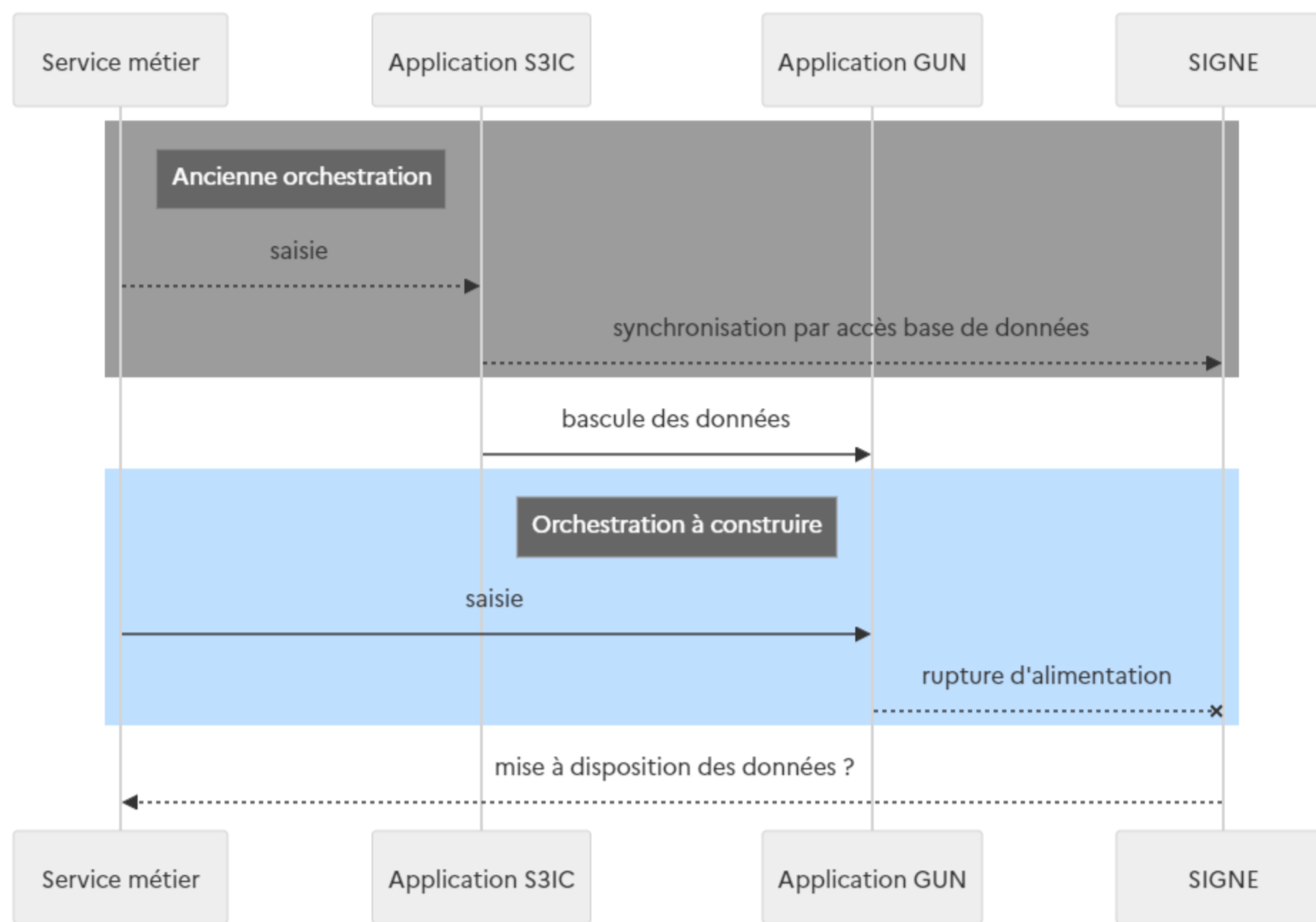


1. Écoute utilisateurs et définition du problème

Scope : métier des inspecteurs ICPE (Installations classées pour la protection de l'environnement)

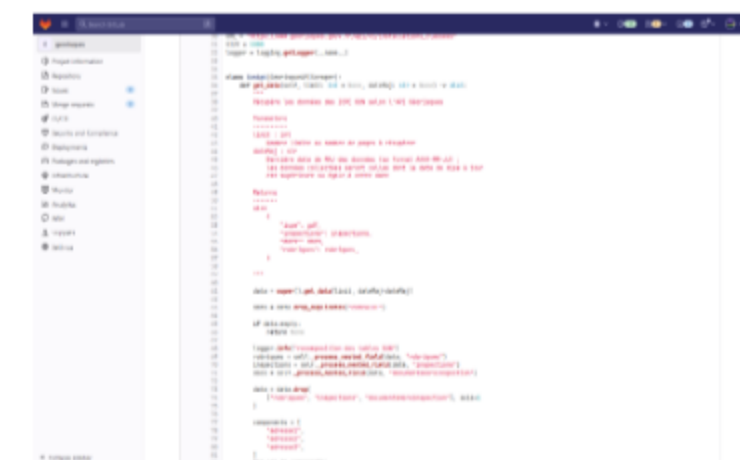
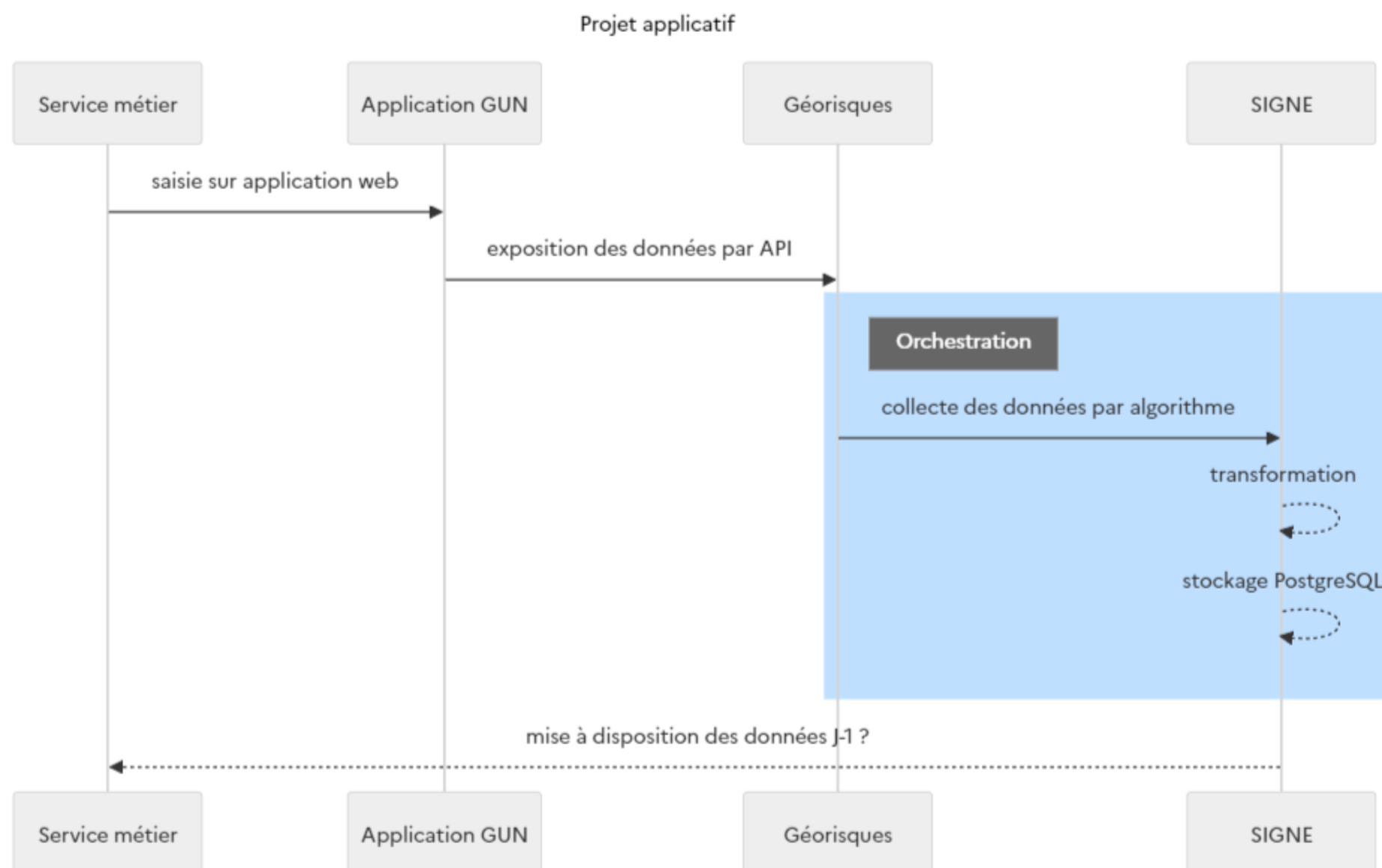
Anciennement, synchronisation de SIGNE à S3IC (ancêtre de GUN) avec remontée des données ICPE sur SIGNE, digitalisation poussée (polygone) directement sur SIGNE. Aujourd'hui saisie de données dans les applications métier GUN (vie administrative des installations classées) et Infosols (sites et sols pollués).

Problème : quid de la mise à jour des données GUN & Infosols sur SIGNE pour les missions régaliennes ?



Prototypage

Ecriture d'un algorithme qui collecte les données de l'API GUN de Géorisques .



Extrait du script de collecte GUN sur repository
Gitlab .

Aparté # 1 : qu'est-ce qu'un algorithme ?

La CNIL définit un algorithme comme "une suite finie et non ambiguë d'instructions permettant d'aboutir à un résultat à partir de données fournies en entrée."

ETALAB, les algorithmes publics : enjeux et obligations [en ligne](#) 

- Concerne les algorithme papiers **et** numériques !
- Principe de redevabilité vis à vis des algorithmes :
 - signalement,
 - description du fonctionnement,
 - justification du recours aux algorithmes,
 - explicabilité des résultats,
 - accessibilité au code source et à la documentation,
 - voies de recours.

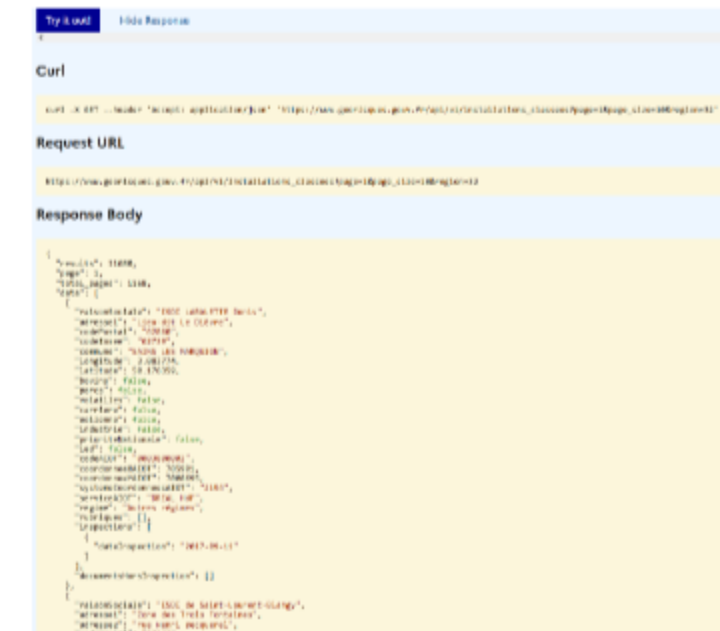
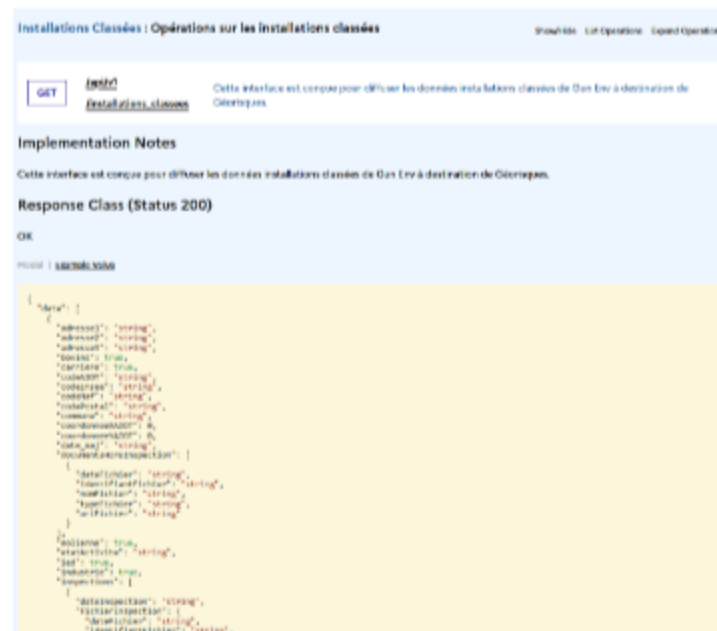
"With great power comes great responsibility: keeping public sector algorithms accountable"



Source : reproduit avec l'autorisation de FIX, <http://www.fix-dessinateur.com>

Aparté #2 : qu'est-ce qu'une API ?

API = Application programming interface (interface de programmation d'application)



Les API permettent de se connecter à des systèmes d'information tiers. Elles sont fondamentales dans l'accès à la donnée pour le partage applicatif. Elles sont largement citées dans les feuilles de routes ministérielles sur la donnée (retrouvez celle du MTECT [ici](#)).

Pour un inventaire, rendez-vous sur la plateforme api.gouv.fr .

Pour mémoire, les API disponibles sur Géorisques sont (à date) :

- AZI : Opérations sur les Atlas des Zones Inondables (AZI)
- CATNAT : Opérations sur les catastrophes naturelles
- Cavites : Opérations sur les Cavités Souterraines (Cavites)
- DICRIM : Opérations sur les Documents d'Information Communal sur les Risques Majeurs (DICRIM)
- Etats documents PPR : Opérations sur les états des documents PPR
- Familles risque PPR : Opérations sur les familles de risque PPR
- Installations Classées : Opérations sur les installations classées
- MVT : Opérations sur les Mouvements de terrains (MVT)
- PAPI : Opérations sur les Programmes d'Actions de Prévention des Inondations (PAPI)
- PCS : Opérations sur les Plans Communaux de Sauvegarde (PCS)
- PPR : Opérations sur les documents PPR
- Radon : Opérations sur le risque radon
- Rapport PDF : Opération pour la génération du rapport PDF
- Retrait gonflement des argiles : Opérations sur les retrait de gonflement des argiles
- Risques : Opérations sur le Détail des risques
- SIS : Opérations sur les données SIS
- TIM : Opérations sur les Transmissions d'Informations au Maire (TIM)
- TRI : Opérations sur les Territoires à Risques importants d'Inondation (TRI)
- Zonage Sismique : Opérations sur le risque sismique

3. Retour à nos moutons : retour utilisateurs du premier prototype de récupération de données

Limites techniques de l'API

- l'API GUN peut s'interroger par code commune, mais nous n'avons pas le millésime du jeu de données ;
- on constate d'ailleurs des mélanges de codes commune et de codes arrondissement municipaux ;
- **mais** l'API utilise le code commune comme paramètre optionnel, on peut donc collecter la totalité des données nationales (127260 résultats sur une API limitée à une page de 1000 entités : 128 pages à collecter)
- une évolution de l'API GUN permet en plus d'interroger par date de mise à jour = grosse avancée sur le plan écoconception !

Limites métier de l'API

- 😞 toutes les ICPE ne sont pas présentes (critères liés au régime et à l'inspection) ;
- 😞 géolocalisation dégradée au centroïde communal pour les sites présentant des risques pour la sécurité publique.

Feedback service métier

- ⚠ « on saisit des polygones, pourquoi on récupère des points ? »

Nota : une erreur détectée (aujourd'hui résolue) sur les codes communes (Fretin et Lesquin apparaissent tous les deux sous le code commune 59256). Ce bugue est aujourd'hui résolu, mais symptomatique du problème de millésimage des codes commune. **Un code commune n'est pas suffisant pour identifier une commune, il faut également disposer de la date de validité.**

4. Poursuite de l'écoute utilisateurs

- GUN autorise la saisie d'une géométrie "point" mais comme la section INFOSOLS permet la saisie de polygones, on observe en pratique une double saisie
- mais pas d'export possible d'INFOSOLS (exposition de données SIS, SUP, ex-BASOL, ex-CASIAS sur géorisques : données morcelées, incomplètes)
- le choix est fait de tester un **scraping** d'Infosols avec des droits ad hoc.

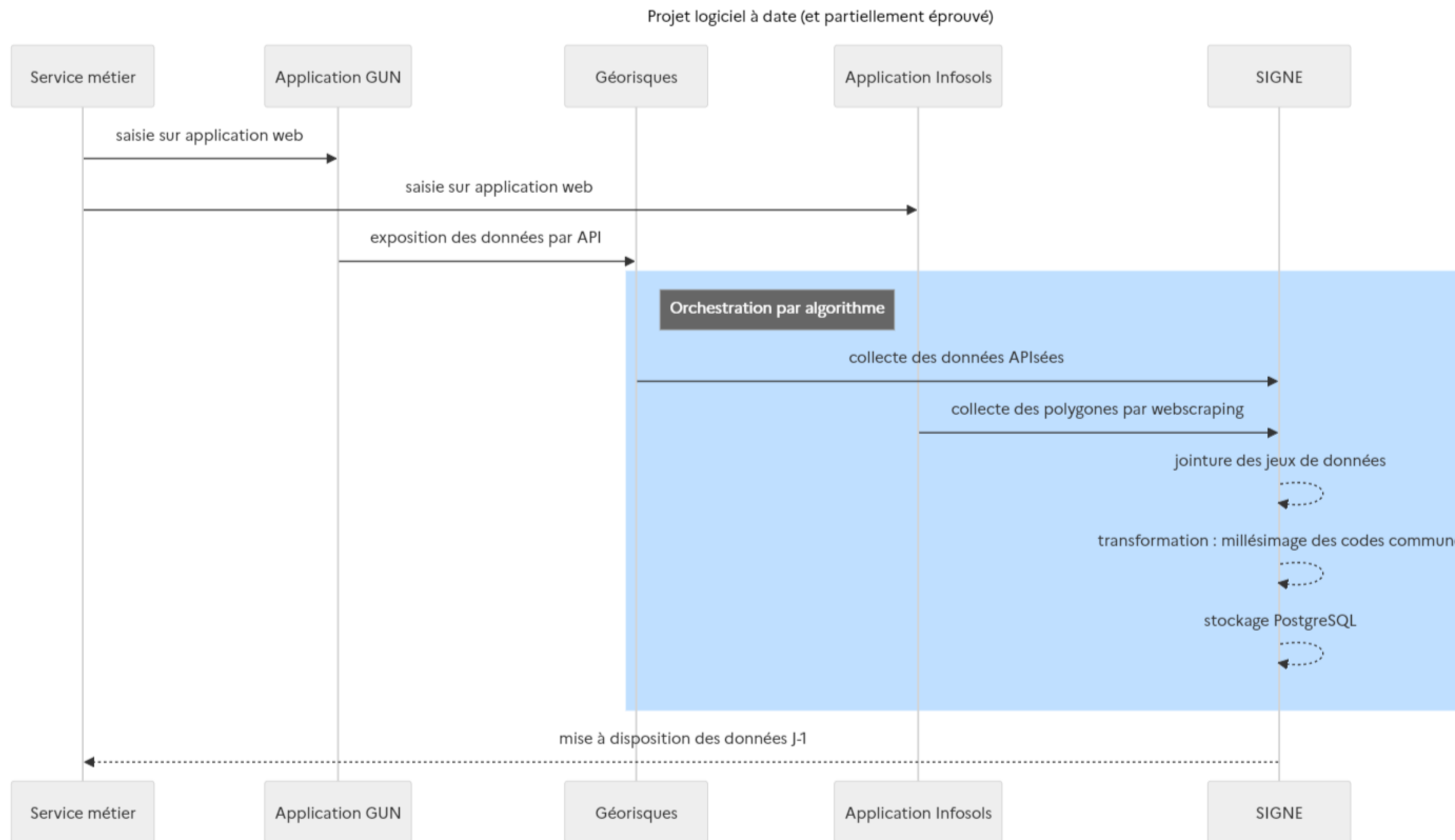
The screenshot shows the 'LISTE DES ÉTABLISSEMENTS' page in the Infosols application. On the left, there is a 'FILTRES LA LISTE' sidebar with sections for Identification, Localisation, and Fiche. The main area displays a table of establishments with columns for ID, Nom de l'établissement, Commune (RDE), Commune(s) (ex-BASOL), Date à jour, Objectif, État de la fiche, Coût, and Actions. The table lists several establishments, including 'SMP - Site Sier-ave', 'RAHBY ENVIRONNEMENT', 'GREA SANTERLANT', 'CHEZ FRED', 'SECA DU PRODIGE', 'EARL DE LA BROSSE', 'SYNDICAT MIXTE DE RAHASSE DES DEBRUS MENAGERES FLANDRE - NORD', 'LA MINOISERIE (SARL)', 'SOURFLET', and 'TECHNOL'. At the bottom right, there is a pagination bar showing '6/177 résultats' and 'Résultats par page 10'.

Copie d'écran d'Infosols

Aparté #3 : c'est quoi du scraping, et pourquoi il faut l'éviter

- le **web scraping**  est une technique qui consiste à simuler l'accès à un site web en utilisant un robot  qui est chargé de collecter certaines données
- ce n'est pas un piratage (sauf exception) ; ici, le profil est administré classiquement et l'authentification transite par Cerbère ;
- il s'appuie souvent sur des "API internes" qui sont aujourd'hui au coeur d'un site internet moderne ;
- mais le scraping doit être utilisé **avec modération** :
 - les serveurs ne sont pas nécessairement dimensionnés pour cela : même si un scraping correctement conduit est "léger" (ciblé sur les seules requêtes nous intéressant là ou un utilisateur humain déclenche plusieurs dizaines de requêtes), il se déplace très vite sur un site internet (facilement plusieurs centaines de requête par seconde), ce qui peut faire tomber un serveur (cf. **attaque par déni de service** )
 - c'est souvent moins bien qu'une extraction ciblée par API (utilise souvent plus de requêtes, ce qui entraîne des délais et de la consommation d'énergie supplémentaire)
 - et surtout la maintenance est une gageure à long terme (à chaque évolution du site web, risque de devoir tout reprendre, contrairement à une API qui se doit d'être versionnée).

5. Second prototype



6. Conclusion et perspectives

- faisabilité validée et intérêt confirmé des services métiers (repérage d'anomalies dans les données extraites)
- conduite du projet "en lignes de commandes" sans ETL
- création d'un package python (publié sur le **gitlab ministériel** ) pour la collecte des données Infosols
- création d'un package python en cours pour collecter les données de l'API GUN **repo gitlab** 
- orchestration à venir (script python, des pistes avec AirFlow)
- une restitution à organiser en dehors des SIG : des données dépourvues de géométrie, des erreurs de géolocalisation en dehors de l'emprise de la carte dynamique : SIGNE ne peut pas restituer ces anomalies !



Aparté #4 : c'est quoi un package ?

- le **packaging** d'un code, c'est une démarche initiée pour partager son code et permettre sa réutilisation par d'autres utilisateurs ;
- cela permet de partager son code sur un *repo* gitlab et de permettre à tout utilisateur (développeur ou simple utilisateur) de signaler ses bugs sur un support partagé, voire même de proposer ses améliorations ;
- la démarche est "code-agnostique" : on la retrouve en python et en R (pour ne citer que ceux-ci) ; ce processus est par exemple au coeur de la **démarche PROPRE**  initiée par la DREAL Pays-de-la-Loire.



Copie d'écran de PyPI

Nota : en python (le langage utilisé ici), on peut produire des packages python stricts, mais aussi spécifiques à QGis (ou même des exécutables windows, mais c'est une autre histoire).

Réflexions sur la dimension RH de ce "pipeline"

- des compétences de gestion de projet : méthode agile, écoute utilisateur
- des compétences technique (data) : API, webscraping, reverse-engineering, packaging
- des compétences géomatiques basiques : projections, gestion de couches SIG
- des compétences métier : compréhension des processus métier
- des **compétences devops** à développer : orchestration
- un projet à échelle humaine : travail en équipe séquentiel, 2 personnes

Mais pour l'essentiel : des agents data-couteaux-suisses....!



Merci pour votre attention !

Des questions ?



[illegible]