Préparation du datastore et nettoyage des données

# La validation d’un point de route

L’objectif est de valider au maximum les informations sur les points de route qui sont obtenus par l’EDI. En effet ces informations ne sont pas fiables car souvent incomplètes voir erronées.

Pour ce faire nous devons trouver un processus de validation du point de route (livraison ou ramasse de marchandise).

Nous avons décidé de les valider avec :

* La Base d’Adresse Nationale ([adresse.data.gouv.fr](https://adresse.data.gouv.fr/))
* l’établissement, lorsque c’est une entreprise, en s’appuyant sur le site gouvernementale ([L’Annuaire des Entreprises françaises : les informations légales officielles de l’administration](https://annuaire-entreprises.data.gouv.fr/))
* La reconnaissance du passe du véhicule (connexion à l’informatique embarquée du véhicule)

# Validation à l’aide de la Base d’Adresse Nationale

Contexte :

Objectif : Utiliser la BAN comme référentiel pour identifier l’adresse d’un point de collecte ou livraison

## Importation et suivi de la Base d’adresse Nationale

### L’équipe :

* Développement des outils  : Alexander Dalziel
* Réalisation et analyse des importations : Alexander Dalziel, Stéphane Millot, El Moundir Faraoun
* Rédacteur : Stéphane Millot

### Le planning

### Les travaux réalisés

**Rappel :** Nous utilisons les fichiers CSV mis à disposition sur le site [Données nationales | adresse.data.gouv.fr](https://adresse.data.gouv.fr/donnees-nationales) pour les importer dans une table SQL avec l’intégralité des colonnes mise à disposition.

**Lors d’une mise à jour :**

* soit l’identifiant dans la BAN (id) est retrouvé et on met à jour la fiche à partir des informations (avec un éventuel réveil si la fiche était en sommeil)
* soit l’identifiant n’est pas trouvé et on insert la fiche
* et on met en sommeil les fiches non présentes dans le nouveau fichier mis à disposition

**Identification d’une nouvelle adresse annulant et remplaçant une adresse**

Certaines adresses mises en sommeil sont remplacées par d’autres avec un nouvel identifiant.

*Exemple :*

|  |  |  |  |
| --- | --- | --- | --- |
| Ancienne ID BAN | Nouvelle ID BAN | Ancienne Adresse | Nouvelle adresse |
| 89086\_3xnnc6\_00003 | 89086\_hf25o3\_00003 | 3 RUE DE LA CROIX SAINT PIERRE | 3 RUE DES MARNES |
| 89046\_h9b9hu\_00143 | 89046\_0060\_00143 | 143 RUELLE DU CHATEAU | 143 RUE DU CHATEAU |

Pour les relier :

Identification des adresses actives ayant les mêmes coordonnées GPS qu’une adresse en sommeil

### Suivi de la base d’adresse nationale

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date d’importation | Nb fiches actives | Nb fiches en sommeil | Nb fiches totales | Nb ligne dans le fichier CSV |
| 07/04/2023 | 68 749 | 0 | 68 749 | 210 022 |
| 23/04/2024 | 206 144 | 2 224 | 208 368 | 206 144 |
| 06/06/2024 | 205 787 | 11 169 | 216 956 | 205 787 |
| 27/06/2024 | 205 347 | 12 052 | 217 399 | 205 347 |
| 01/07/2024 | 205 133 | 12 391 | 217 524 | 205 133 |
| 03/07/2024 | 205 134 | 12 405 | 217 539 | 205 134 |
|  |  |  |  |  |

### Les problèmes rencontrés

Les adresses étant en cours de validation nous constatons les problèmes suivants :

* La BAN est incomplète car des bâtis n’ont pas d’adresse

Exemple : la rue gustave cotteau à Auxerre au 03/07/2024 n’est pas référencée



Les écuries du bas four 89310 Fresnes n’ont pas d’adresse



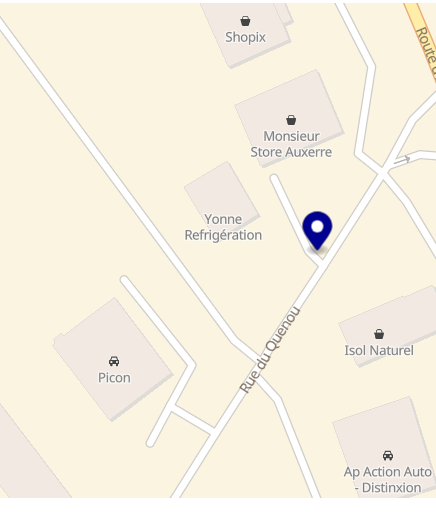
* La BAN n’est pas précise au moins sur les adresses non certifiées

Exemple : Garage Roizot 12 rue du quenou 89380 Appoigny

Selon google :



Selon la BAN au 04/07/2024

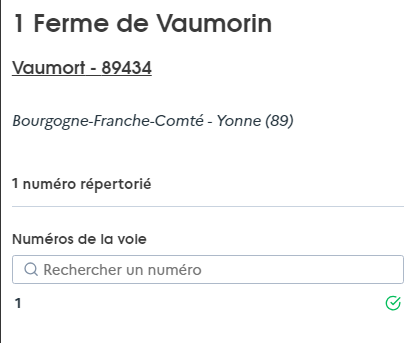


* La validation des adresses par les communes entraine des problèmes de mises à jour car :
* Une adresse peut disparaitre dans un certain laps de temps. Des adresses avec des bâtis (dans l'exemple plusieurs immeubles) sont supprimées sans nouvelles adresses sur une commune qui annonce 100% de réalisation de la BAN.

exemple : ID BAN : 89418\_zkk3od\_00015 à TONNERRE (89700) 15 rue des vals profondes qui n'existe plus Au moins 4 immeubles sont sans adresses



* Des noms de voies sont incorrects car le numéro a été mis dans le nom de la voie :



* Nous n’avons pas trouvé comment relier la nouvelle adresse avec l’ancienne car les id sont différentes.

L’identification est réalisée par l’ajout d’une comparaison sur les coordonnées gps, les informations textuelles (lieu-dit, …)

### Travaux à réaliser

Ajouter un champ dans la table BAN pour relier une adresse mise en sommeil avec une adresse active en se basant sur des stricts égalités sur l’adresse textuelle puis pour celle non trouvées sur l’adresse exacte sur les coordonnées GPS.

Ré entrainer le modèle pour qu’il apprenne à reconnaitre l’ancienne adresse et la nouvelle.

## Les destinataires et l’identification dans la Base d’Adresse Nationale

### L’équipe

* Développement des outils : Alexander Dalziel
* Réalisation et analyse des importations : Alexander Dalziel, Stéphane Millot, El Moundir Faraoun
* Rédacteur : Stéphane Millot

### Le planning

### L’algorithme

L’identification de la voie de la BAN à partir de l’adresse du destinataire nécessite :

* que l’adresse du destinataire soit suffisamment précise (présence du nom de la voie, du numéro, commune correcte,…)
* qu’elle existe dans la BAN

### La réaffectation des id ban en sommeil sur un id ban actif

Voir la requête d’Alex pour la décrire

### Les résultats

Les résultats n’étant pas satisfaisant nous souhaitons reprendre les travaux sur la détection d’entité nommé pour améliorer le modèle :

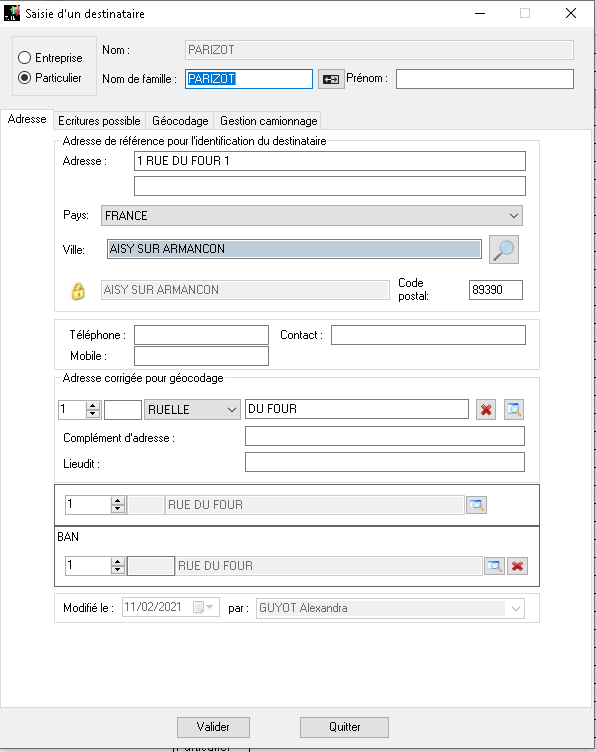
* Apporter plus de contexte au modèle
* Améliorer l’entrainement
* Identifier le meilleur prétendant ou l’impossibilité de le trouver

## Analyse des adresses non identifié avec le BAN

### Adresse inexistante dans notre BAN importé

Exemple : 8 RUE DU TRIANON 89 car définit dans le BAN comme une cage d’escalier. Il faut donc ajouter les cages d’escalier dans le BAN

### Problème de reconnaissance dans le traitement automatique alors qu’individuellement cela fonctionne



### Adresse en dehors du référentiel

Nous avions traité uniquement le département 89

Il faudrait aussi importer le 10 et le 95

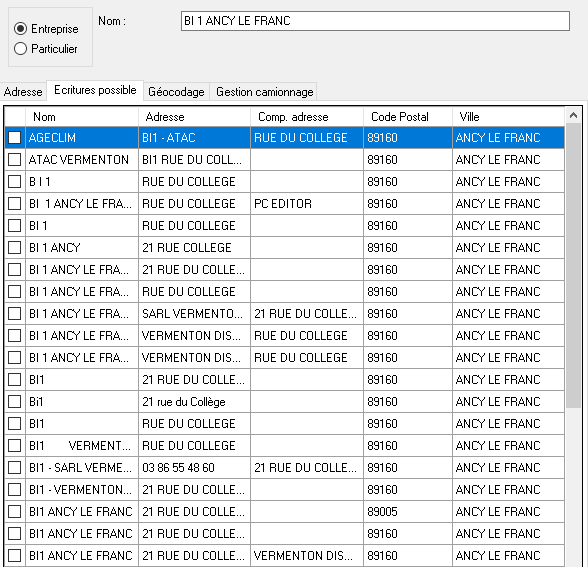
### Problème d’historisation de l’adresse

Exemple : BI1 ANCY LE FRANC RUE DU COLLEGE 89160 ANCY LE FRANC

Le système identifie le point de livraison lors de sa création (première livraison ou collecte). Si le numéro n’est pas mentionné sur cette première livraison ou collecte il ne le sera jamais même si on l’obtient sur les livraisons suivantes.

Ceci vient en partie de la construction du jeu de test qu’il faudrait améliorer pour qu’il corrige les adresses sans numéro.

En consultant les alias on voit pourtant que le numéro est le 21

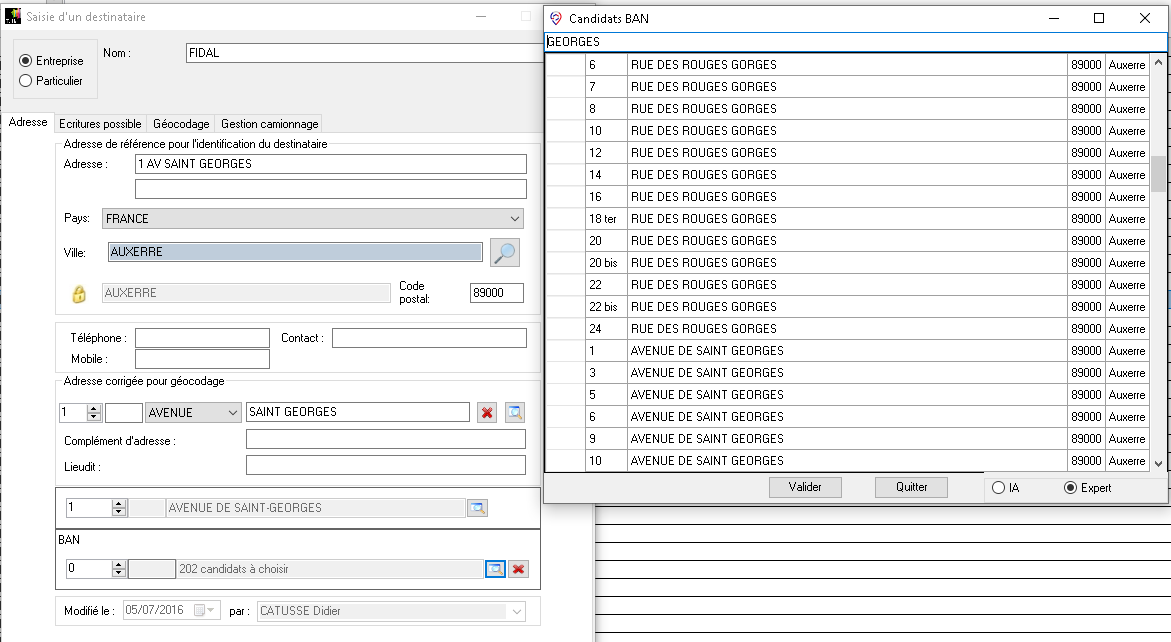


### Problème avec les entités nommées

4 PETITE RUE 89160 ANCY LE LIBRE identifié en 4 RUE PETITE donc non identifiable en automatique avec la BAN .

Il faudrait entrainer le modèle sur les inversions d’entités

### Problème de déterminant



# Validation à l’aide des SIRET entreprise

## Données entreprise destinataire

Il faudrait pouvoir consolider les données relatives aux entreprises destinataire. L’une des principales donnée à récupérer est le SIRET de l’établissement destinataire. Cela permettra de s’avoir si le destinataire est toujours en activité ou non, permettant de le mettre en sommeil dans la base de données.

Pour récupérer les informations il est envisageable d’utiliser l’API du gouvernement : <https://recherche-entreprises.api.gouv.fr/>

(Documentation : [ReDoc (api.gouv.fr)](https://recherche-entreprises.api.gouv.fr/docs/))

Depuis l’api 2 type de recherche :

* Recherche textuelle : Cet endpoint permet de récupérer les unités légales et les établissements correspondantes à la recherche textuelle effectuée sur la dénomination, l'adresse, les dirigeants et les élus.
* Recherche géographique : Cet endpoint prend en paramètre une latitude (:lat) et une longitude (:long) et renvoie les unités légales et leurs établissements autour de ces coordonnées. Vous pouvez également préciser un paramètre radius en km(défaut: 5 km).

Il est donc possible de croiser les données des deux requêtes pour être sur d’identifier le bon établissement.

### Exemple de réponse



Faire une table SIRET

* Cle
* Cledestinataire
* SIRET
* CodeAPE (activite\_principale)
* DateCreationEtablissement (date\_creation)
* DateFermetureEtablissement (date\_fermeture)
* GPSX
* GPXY
* Numerovoie
* Typevoie
* Libellevoie
* Codepostal
* Ville

Table SIRENEDIRIGEANTS

* cle
* cleSIREN
* nom
* prenom
* qualité
* typedirigeant

# Affectation des établissements :

Possibilité de rechercher l’établissement

* par similarité orthographique (Par défaut)
* avec un contient

Possibilité de rechercher sur

* Le nom de l’entreprise
* L’adresse
* Le nom du dirigeant

Empêcher de pouvoir affecter un SIRET établissement s’il est déjà utilisé par un autre destinataire. Dans ce cas afficher les deux fiches et proposer de les fusionner en gardant l’établissement sur la fiche qu’on garde en ayant la possibilité d’annuler

# Labélisation des erreurs

Pour entrainer l’IA dans l’identification et la correction des erreurs depuis les requêtes EDI (Nom destinataire + Adresse), nous devons labéliser manuellement les types d’erreur que nous rencontrons.

Pour cela nous pouvons envisager une table TypeErreur (cle, typeerreur en tinyint) contenant :

|  |  |
| --- | --- |
| 1 | Adresse imprécise |
| 2 | Création |
| 3 | Test |

Ainsi qu’une table Erreur (cle, cleType, sstypeerreur en tinyint, erreur varchar(max)) contenant :

|  |  |  |
| --- | --- | --- |
| sstypeerreur | Libelle | Cletype |
| 1 | Zone (ZI ZA ZAC …) | 1 |
| 2 | Mauvais numéro | 1 |
| 3 | Mauvais type de voie | 1 |
| 4 | mauvais nom de voie | 1 |
| 5 | Commune regroupement de commune | 1 |
| 6 | Commune approximative | 1 |
| 7 | Commune inattendue | 1 |
| 8 | Numéro inexistant dans le BAN | 1 |
| 9 | Nom d'un dirigeant | 1 |
| 10 | Nom d'un salarié | 1 |
| 11 | Utilisation du sigle sncf … | 1 |
| 12 | Adresse mal ordonée | 1 |
| 13 | Adresse différente | 1 |
| 14 | Création | 2 |
| 15 | Entreprise dans une entreprise (entrepôt logistique, …) | 1 |
| 16 | Intervention d'une entreprise à l'adresse de son client | 1 |
| 17 | Test EDI | 3 |

Et pour lier les erreurs aux livraisons une table LivraisonErreur (cle, cleLivraison, cleErreur)

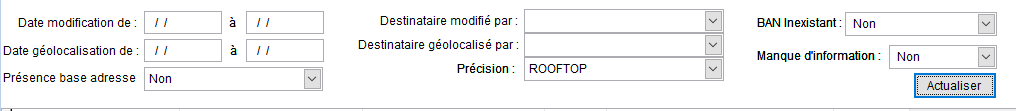
# Annexe 1 : Travaux sur la correction des destinataires

## Affectation des adresses BAN

### A partir des adresses reconnues en ROOFTOP par Google affecté l’adresse BAN

Au 31/07/2024 il reste 1067 fiches

Fichier / Géolocalisation / Recherche modification destinataire



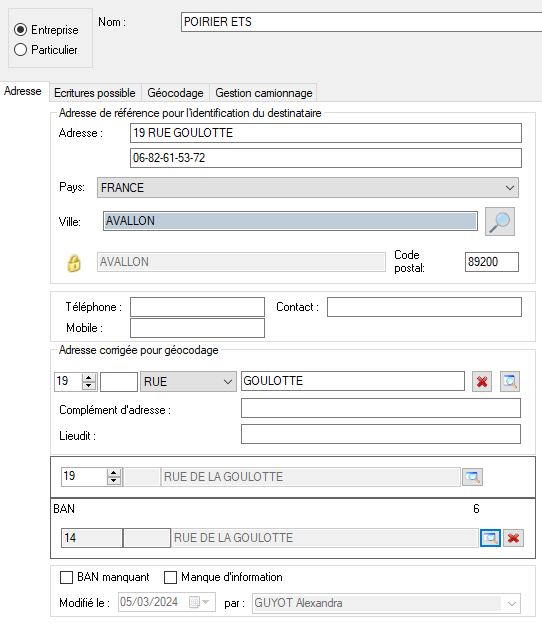
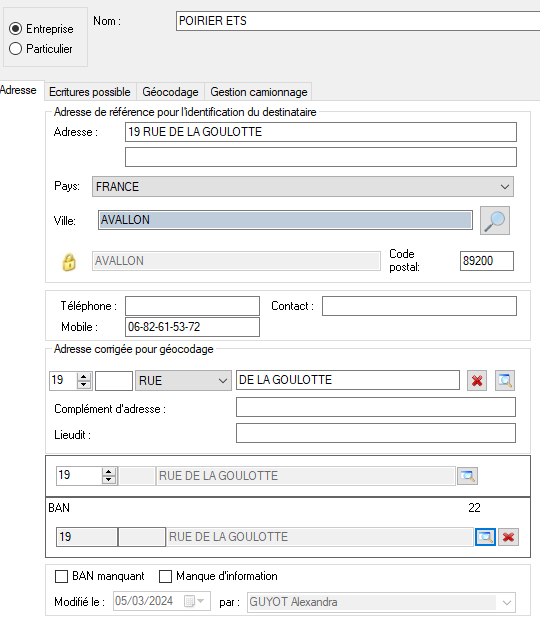
3 cas de figures :

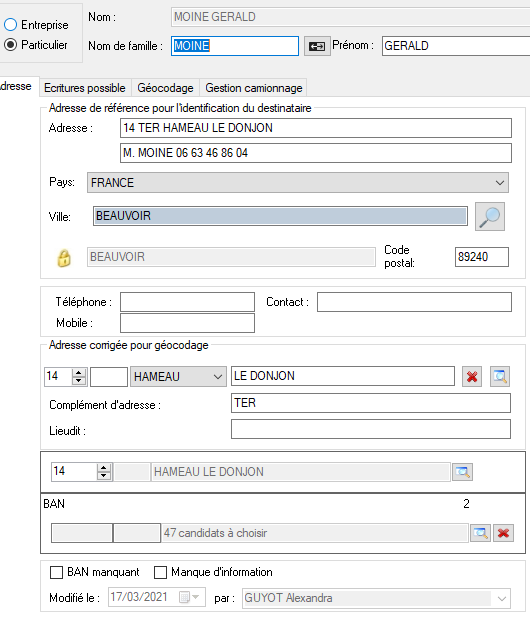
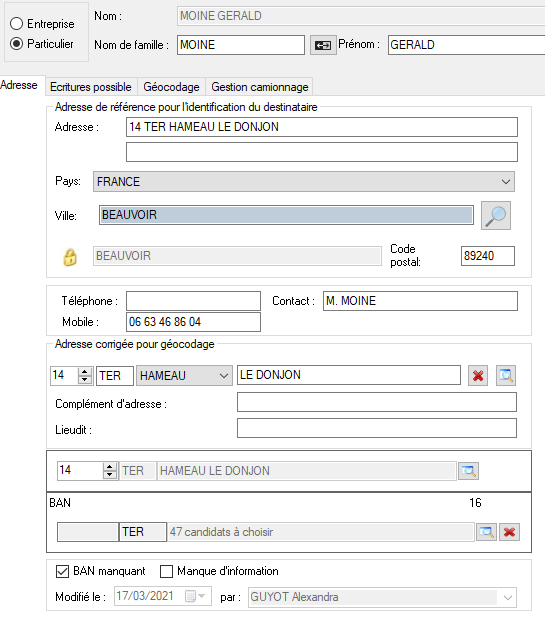
* On retrouve l’adresse dans le BAN
* L’adresse n’existe pas dans le BAN
* Il manque des informations (pas de N° dans la voie par exemple)

Dans tous les cas il faut :

* Nettoyer l’adresse pour que ce soit celle du BAN
* Mettre le téléphone fixe dans Téléphone
* Mettre le téléphone mobile dans Mobile
* Mettre le contact dans contact

Avant Après

L’affectation automatique n’a pas fonctionné car :

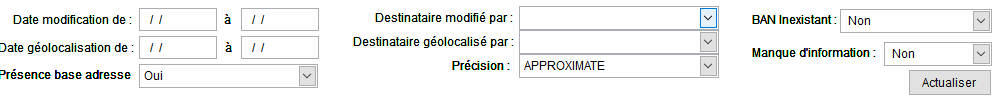
* L’adresse est trop loin selon damerau levenstien (19 RUE GOULOTTE et 19 RUE DE LA GOULOTTE)
* L’adresse n’est pas dans la bonne commune (CHARNY / DICY)

### Affectation des bonnes coordonnées GPS

#### Adresse BAN Ok géocodage APROXIMATE

Au 31/07/2024 il reste 1245 fiches

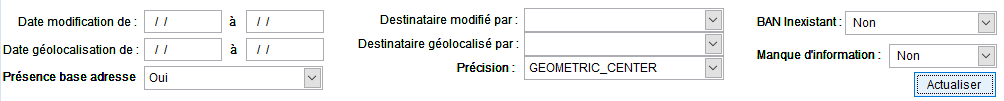
Fichier / Géolocalisation / Recherche modification destinataire



#### Adresse BAN Ok géocodage GEOMETRIC CENTER

Au 31/07/2024 il reste 1076 fiches

Fichier / Géolocalisation / Recherche modification destinataire



# Annexe 2 : Journal synthétique de l’avancé de l’affectation d’une adresse avec une adresse de la BAN

## Le 21/04/2024

### Résultats

* 45483 destinataires identifiés et considéré comme actif (destinataire is null)
* 21142 destinataires non identifiés avec le BAN (soit 46,48%)
* 24341 destinataires identifiés avec le BAN (soit 53,52%)

## Le 21/05/2024

### Résultats

* 46768 destinataires identifiés et considéré comme actif (destinataire is null)
* 13353 destinataires non identifiés avec le BAN (soit 28,55%)
* 33415 destinataires identifiés avec le BAN (soit 71,45%)

## Le 21/06/2024

### Améliorations apportées

Application des correctifs suivant :

* Prise en compte de la problématique des sous localités de commune regroupée (exemple charny orée de puisaye)
* Prise en compte des compléments de numéros

### Résultats

* 46768 destinataires identifiés et considéré comme actif (destinataire is null)
* 10721 destinataires non identifiés avec le BAN (soit 22,93%)
* 36047 destinataires identifiés avec le BAN (soit 77,07%)

## Le 31/12/2024

### Résultats

* 47522 destinataires identifiés et considéré comme actif (destinataire is null)
* 10721 destinataires non identifiés avec le BAN (soit 15,3%)
* 40250 destinataires identifiés avec le BAN (soit 84,7%)