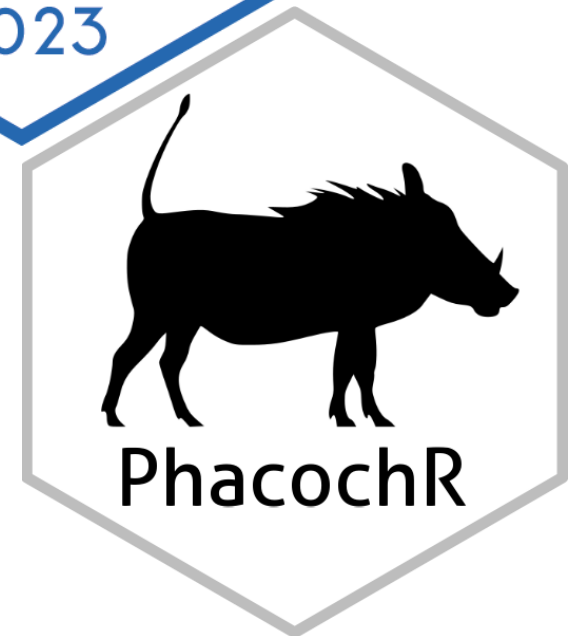


PHACOCHR

Outil de géocodage libre pour la Belgique

Joël Girès – Observatoire de la Santé et du Social
Hugo Périlleux – IGEAT-ULB



Observatorium
voor Gezondheid en Welzijn
Brussel



Observatoire
de la Santé et du Social
Bruxelles

ULB
IGEAT

1. INTRODUCTION : PHACOCHR, QU'EST CE QUE C'EST ?

Géocodeur

PhacochR est un outil qui produit des coordonnées X-Y à partir de listes d'adresses.

Libre

L'outil est entièrement libre :

- C'est un **package R** qui est constitué d'un code sous licence libre ;
- Il repose sur les **données publiques BeST Address** (compilation de Urbis, Icar et Crab) :
<https://opendata.bosa.be/index.fr.html>

Le code est disponible sur Github : <https://github.com/PhacochR/PhacochR>

La documentation est disponible sur un site dédié : <https://PhacochR.github.io/PhacochR/>

Rapide, léger et local

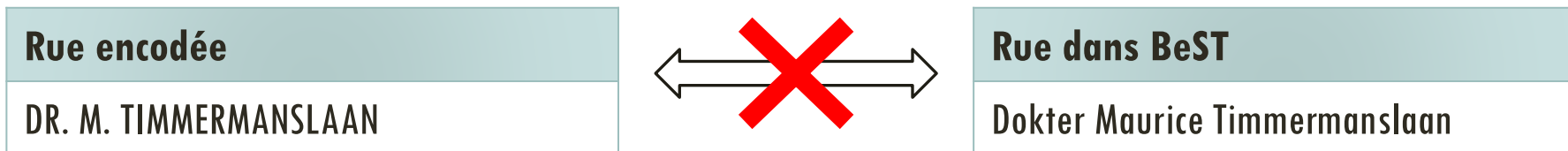
Alternative **facile, légère** et **fiable** par rapport à Nominatim (OSM) ou Google Maps.
Fonctionnement entièrement **local** : idéal pour traiter des adresses confidentielles.

2. LOGIQUE DE PHACOCHR

Données du problème : comment géocode-t-on ?

Pour géocoder, il faut joindre chaque adresse de la base de données à géocoder à la bonne adresse correspondante de **BeST Address**, comprenant les coordonnées.

Le problème : l'orthographe encodée ne correspond jamais exactement à l'orthographe « officielle » (abréviations, coquilles, manque le mot « rue », etc.)



=> Le but de PhacochR est spécifiquement de rendre cette jointure possible, en faisant une **jointure inexacte** (fautes permises) ! C'est ce que fait la fonction principale du package : `phaco_geocode()`

2. LOGIQUE DE PHACOCHR

Préambule 1 : les données ne sont pas intégrées

PhacochR ne contient pas directement les données (**BeST Address** et autres) nécessaires au géocodage : il faut lancer la fonction `phaco_setup_data()` après l'installation du package

=> la fonction télécharge les fichiers et les stocke de manière permanente dans un répertoire de travail déterminé grâce au package `rappdirs`. C'est après cette étape que `phaco_geocode()` peut fonctionner.

Plusieurs raisons à ce choix :

- Les données sont trop volumineuses (problème pour CRAN + pénible pour la mise à jour) ;
- Il donne à l'utilisateur/trice la possibilité de mettre à jour lui/elle-même les données (voir slide suivante).

2. LOGIQUE DE PHACOCHR

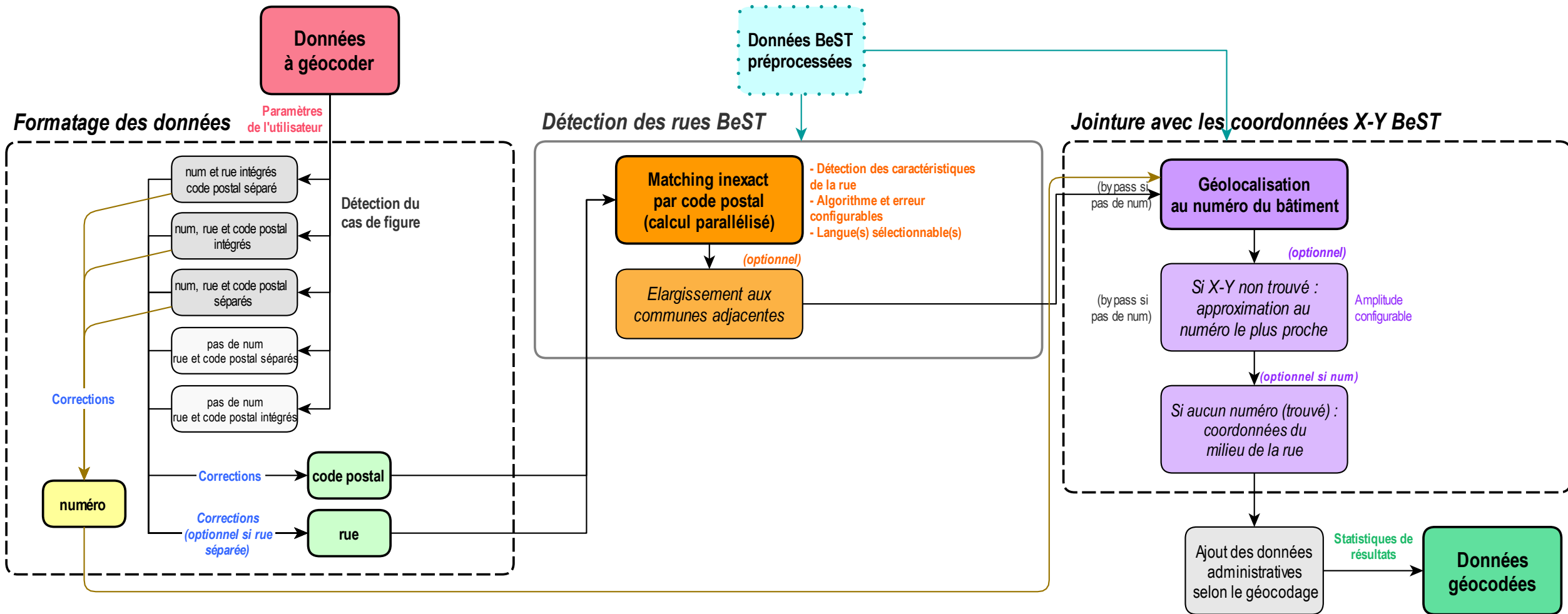
Préambule 2 : les données BeST sont reformatées

Les données **BeST Address** utilisées sont une transformation des données originales, dans le but d'augmenter la vitesse, la performance et les options du géocodage.

Si l'utilisateur désire mettre à jour les données BeST, le package dispose de la fonction `phaco_best_data_update()` : La fonction télécharge les dernières données BeST, les transforme, y intègre d'autres informations (issues de Statbel et Urbis) et sauvegarde le résultat dans le répertoire de travail de PhacochR pour être utilisées lors du géocodage.

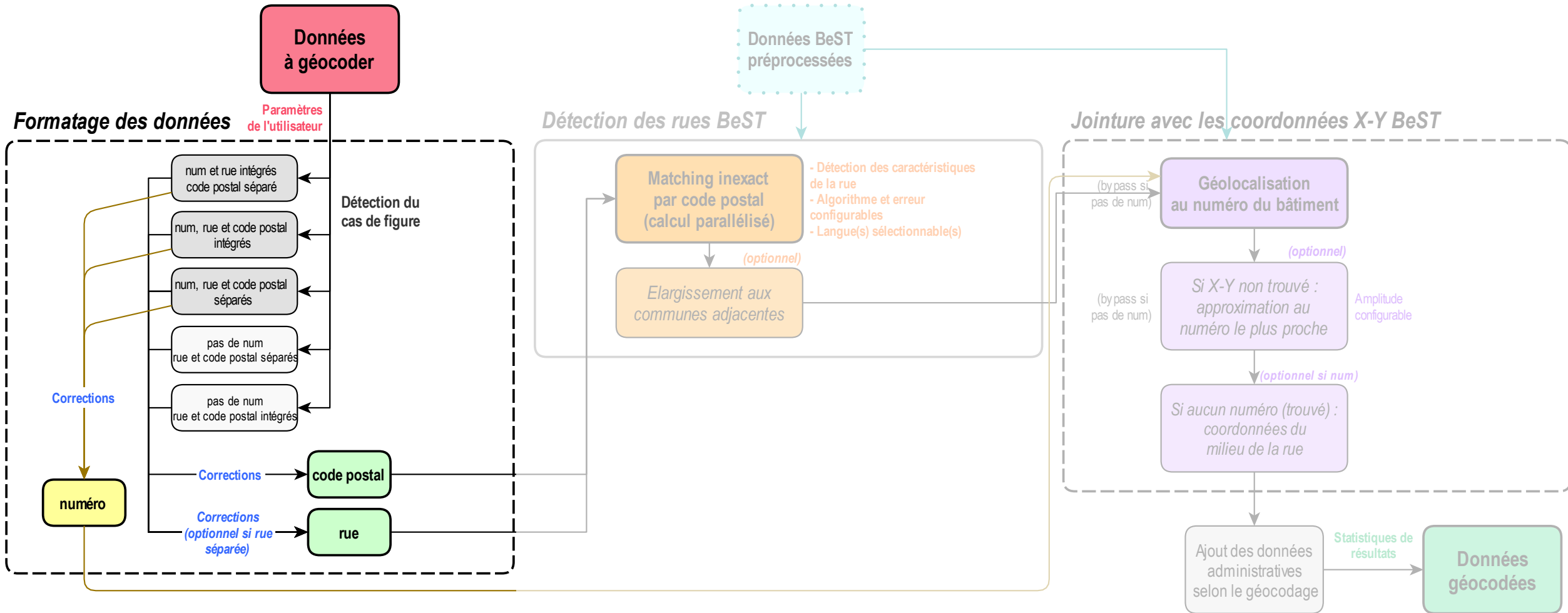
=> *Avantage* : l'utilisateur/trice peut réaliser lui/elle-même la mise à jour des données. Les données BeST sont mises à jour de manière hebdomadaire, et il aurait été trop contraignant de mettre le package à jour toutes les semaines pour suivre ce rythme !

2. LOGIQUE DE PHACOCHR



2. LOGIQUE DE PHACOCHR

1) FORMATAGE DES DONNÉES



2. LOGIQUE DE PHACOCHR

1) FORMATAGE DES DONNÉES

La première chose que fait `phaco_geocode()` est détecter les variables nécessaires (numéro, code postal) si besoin. Quelle que soit la manière dont il est encodé : **le code postal est nécessaire !**

num_rue_code_postal
15, rue notre-seigneur Bruxelles 1000
Boulevard du Triomphe, 153 Bte 7614 Ixelles 1050
Rue Royale, 344 bte 3.2 Schaerbeek 1030
Promenade de l'Alma, 49/312 BRUXELLES 1200
45 Rue des palmiers Woluwe-Saint-Pierre 1150
Rue Picard 68 Sint-Jans-Molenbeek 1080



rue_recoded	num_rue_clean	code_postal_to_geocode
rue notre-seigneur	15	1000
Boulevard du Triomphe	153	1050
Rue Royale	344	1030
Promenade de l'Alma	49	1200
Rue des palmiers	45	1150
Rue Picard	68	1080

2. LOGIQUE DE PHACOCHR

1) FORMATAGE DES DONNÉES

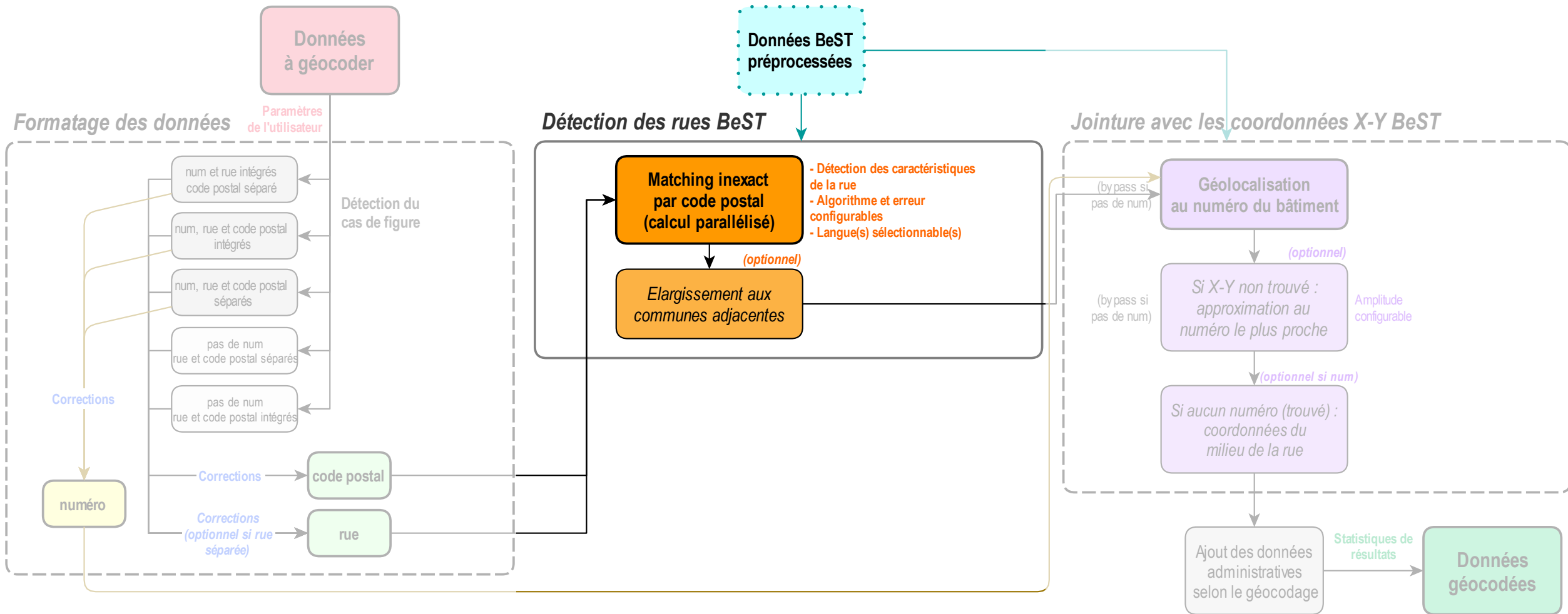
PhacochR nettoie et corrige les adresses. La nature du recodage est indiquée dans la colonne recode.

Ces détections et corrections sont réalisées à l'aide du package `stringr` et d'**expressions régulières** (regex).

rue		rue_recoded	recode
Rue Sous Lt. Catoire(D)		Rue Sous Lieutenant Catoire	parenthese ; Lieutenant
KON. ELISABETHPLEIN		Koningin ELISABETHPLEIN	koning
Av. de Tervueren 116 BP 14		Avenue de Tervueren	BP_CP ; num ; avenue
Torhoutsesteenweg 44/6.03		Torhoutsesteenweg	slash ; num
CHEE DE ST JOB		Chaussee DE Saint JOB	Saint ; chaussee
Kouterstraat(LOO)		Kouterstraat	parenthese
de l'Ecureuil,		Rue de l'Ecureuil	virgule ; Rue
Burg. Gillonlaan		Burgemeester Gillonlaan	Burgemeester

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES



2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

Imaginons que nous voulons détecter de quelle rue il s'agit lorsqu'on fournit à PhacochR la rue « De la ligne », numéro 57 au code postal « 1000 ».

PhacochR corrige d'abord la rue et la compare ensuite à toutes les rues avec le même code postal

rue	num	code postal
De la ligne	57	1000



rue_recoded	recode
Rue De la ligne	Rue



postal_id	street_FINAL_detected	langue_FINAL_detected
1000	Rue du Pavillon	FR
1000	Paviljoenstraat	NL
1000	Rue de l'Homme Chrétien	FR
1000	Kerstenmannekensstraat	NL
1000	Rue du Parlement	FR
1000	Parlementsstraat	NL
1000	Rue des Riches Claires	FR
1000	Rijckelarenstraat	NL
1000	Rue de l'Ommegang	FR
1000	Ommegangstraat	NL
1000	Rue de la Flèche	FR
1000	Pijlstraat	NL
1000	Rue Watteu	FR
1000	Watteustra	NL
1000	Rue d'Egmont	FR
1000	Egmontstraat	NL
1000	Rue du Faubourg	FR
1000	Voorstadsstraat	NL
1000	Rue Lesbroussart	FR
1000	Lesbroussartstraat	NL
1000	Rue Auguste Orts	FR
1000	Auguste Ortsstraat	NL
1000	Chemin des Oiseleurs	FR
1000	Vogelvangersweg	NL
1000	Impasse du Cheval	FR
1000	Paardgang	NL
1000	Rue du Chevreuil	FR
1000	Reebokstraat	NL
Etc...		

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

La détection des rues est réalisée à l'aide d'une boucle opérant par code postal. Nous utilisons le package `fuzzyjoin` pour réaliser le **matching inexact**. Le package `foreach` est utilisé pour paralléliser la boucle (voir capture d'écran) afin de gagner en vitesse, s'agissant du premier goulot d'étranglement de la fonction.

```
### ii) Boucle de jointure par commune -----  
  
# /\ NOTE : la cle de jointure est en minuscule (d'ou les str_to_lower() avant), car stringdist identifie la diff de case comme une diff !  
# /\ NOTE2 : la jointure cree les colonnes de postal_street, meme si 0 match ! Important pour la suite, notamment le if statement pour la creation de l'objet sf  
res <- tibble()  
res <- foreach (i = unique(data_to_geocode$code_postal_to_geocode),  
               .combine = 'bind_rows',  
               .packages=c("dplyr","fuzzyjoin")) %dopar% {  
  
  data_to_geocode_i <- data_to_geocode %>%  
    filter(code_postal_to_geocode == i)  
  
  postal_street_i <- postal_street %>%  
    filter(postal_id == i)  
  
  stringdist_left_join(data_to_geocode_i,  
                      postal_street_i,  
                      by = c("address_join" = "address_join_street"),  
                      method = method_stringdist,  
                      max_dist = error_max,  
                      distance_col = "dist_fuzzy",  
                      nthread= n.cores)  
  
}
```

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

PhacochR sélectionne ensuite toutes les rues qui ressemblent dans la limite d'erreur décidée (par défaut 4), et sélectionne la rue avec le moins d'erreurs (l'erreur max est configurable).

rue	num	code postal
De la ligne	57	1000



rue_recoded	recode
Rue De la ligne	Rue



street_FINAL_detected	dist_fuzzy
Rue de Ligne	3
Rue de la Cigogne	4
Rue de la Colline	4
Rue de la Reine	4
Rue de la Loi	4

Dans le cas d'un ex-aequo, PhacochR calcule une deuxième mesure d'erreur (Jaro-Winkler), et ne sélectionne que l'adresse la plus ressemblante (non illustré ici).

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

Afin de détecter les rues contenant des prénoms abrégés, nous avons recréé (lors de la transformation des données BeST) des doublons des rues contenant des prénoms avec leur équivalent avec prénom abrégé. Une rue contenant un prénom abrégé est indiquée dans la colonne `nom_propre_abv`.

rue	street_FINAL_detected	nom_propre_abv	dist_fuzzy
A DE COCKSTRAAT	Alfons De Cockstraat	1	0
	<i>A De Cockstraat</i>		
RUE C. BUYSSE	Rue Cyriel Buysse	1	1
	<i>Rue C Buysse</i>		
JB. VAN MONSSTRAAT	Jean-Baptiste Van Monsstraat	1	1
	<i>JB Van Monsstraat</i>		
Avenue F. Ferrer	Avenue Francisco Ferrer	1	1
	<i>Avenue F Ferrer</i>		
RUE LENOIR	Rue Ferdinand Lenoir	1	2
	<i>Rue F Lenoir</i>		

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

Certains codes postaux encodés sont erronés : dans ce cas, PhacochR ne trouve donc pas la rue, la comparaison étant effectuée par code postal.

Exemple d'adresse :

27 rue du moulin à 1030 Schaerbeek, qui se trouve en réalité à 1210 Saint-Josse.

Élargissement aux communes adjacentes

Adresse recherchée

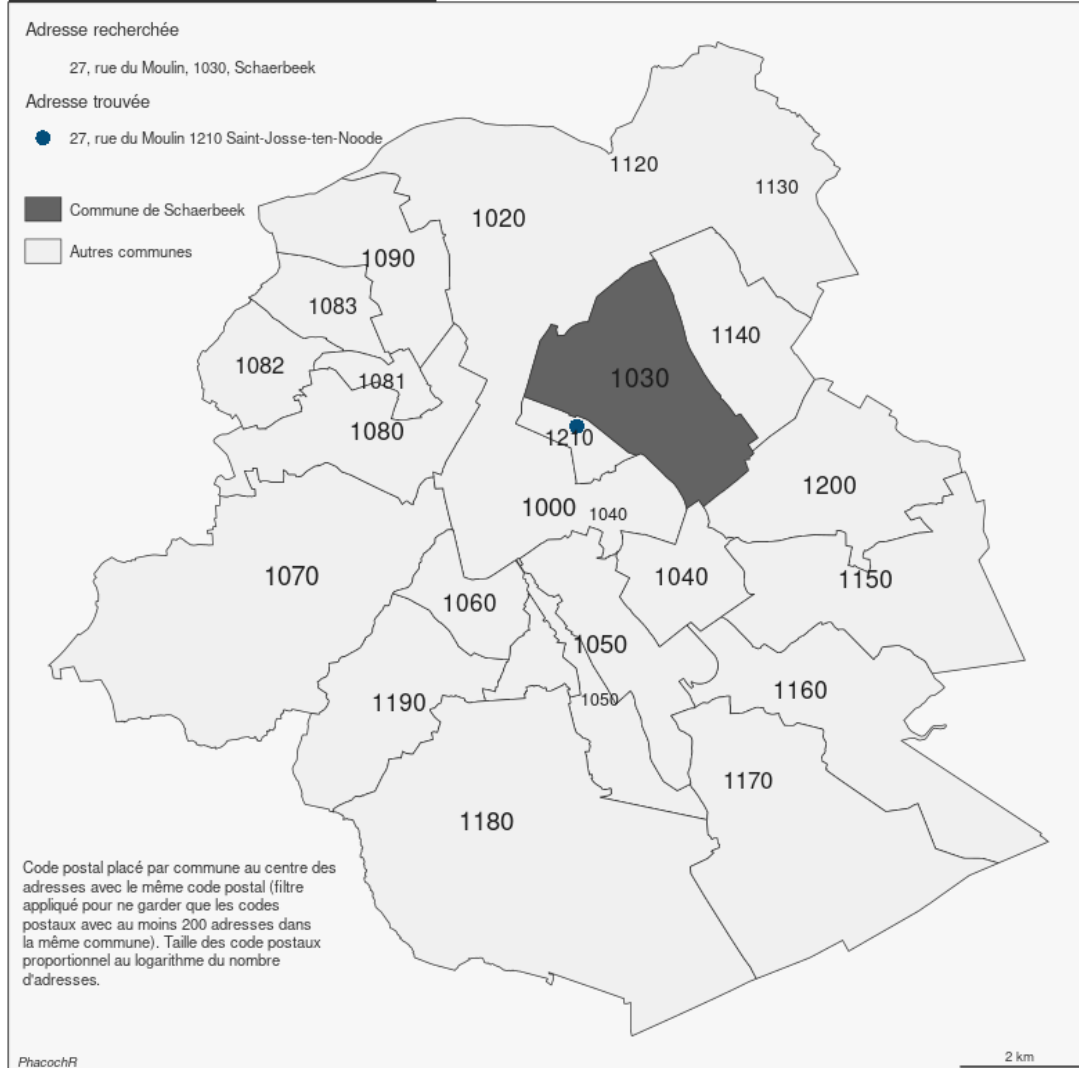
27, rue du Moulin, 1030, Schaerbeek

Adresse trouvée

● 27, rue du Moulin 1210 Saint-Josse-ten-Noode

■ Commune de Schaerbeek

□ Autres communes



2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

Élargissement aux communes adjacentes

Dans ce cas, PhacochR élargit sa recherche à la commune contenant le code postal et aux communes adjacentes (*optionnel*).

=> Il trouve alors le 27 rue du moulin à 1210 Saint-Josse.

Élargissement aux communes adjacentes

Adresse recherchée

27, rue du Moulin, 1030, Schaerbeek

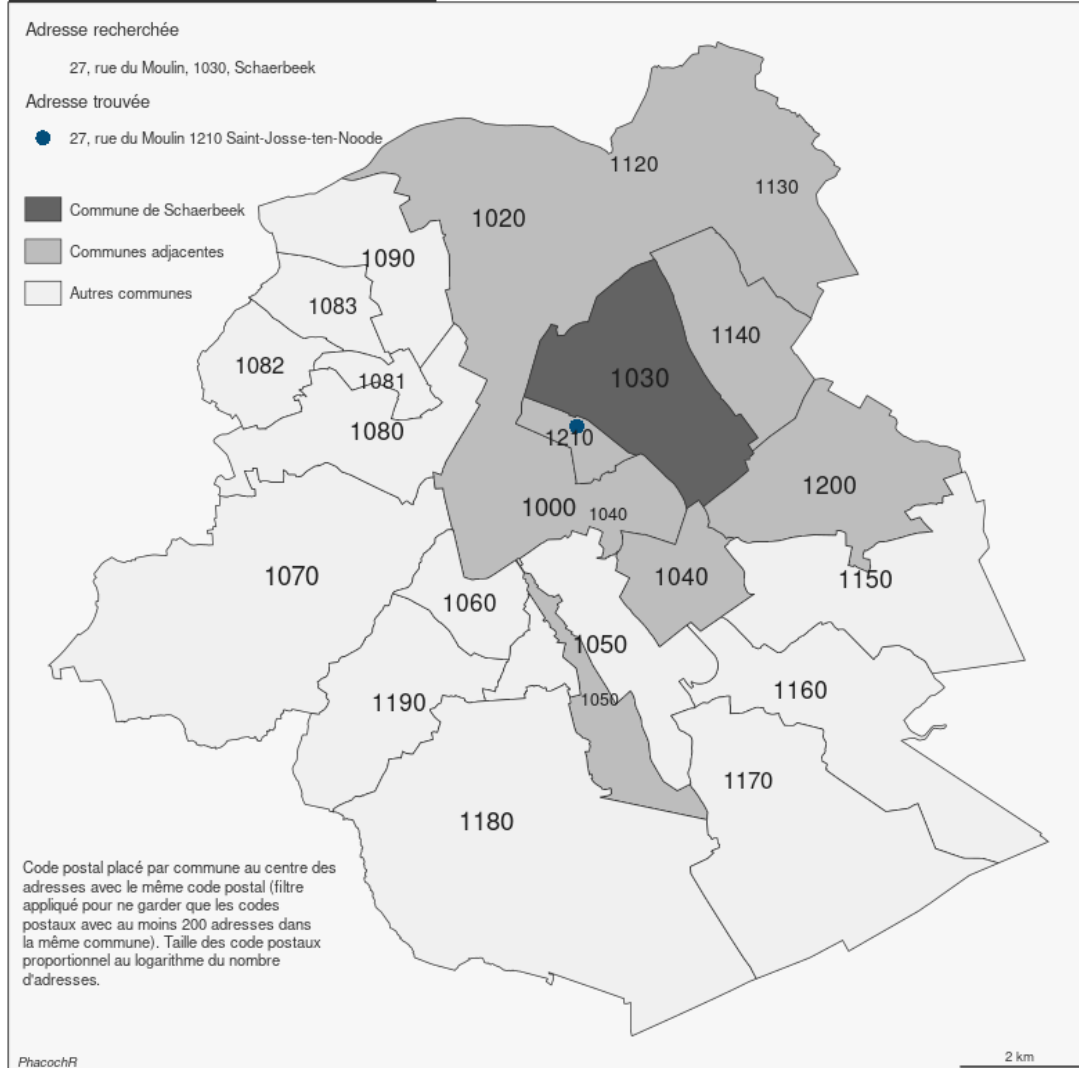
Adresse trouvée

● 27, rue du Moulin 1210 Saint-Josse-ten-Noode

■ Commune de Schaerbeek

■ Communes adjacentes

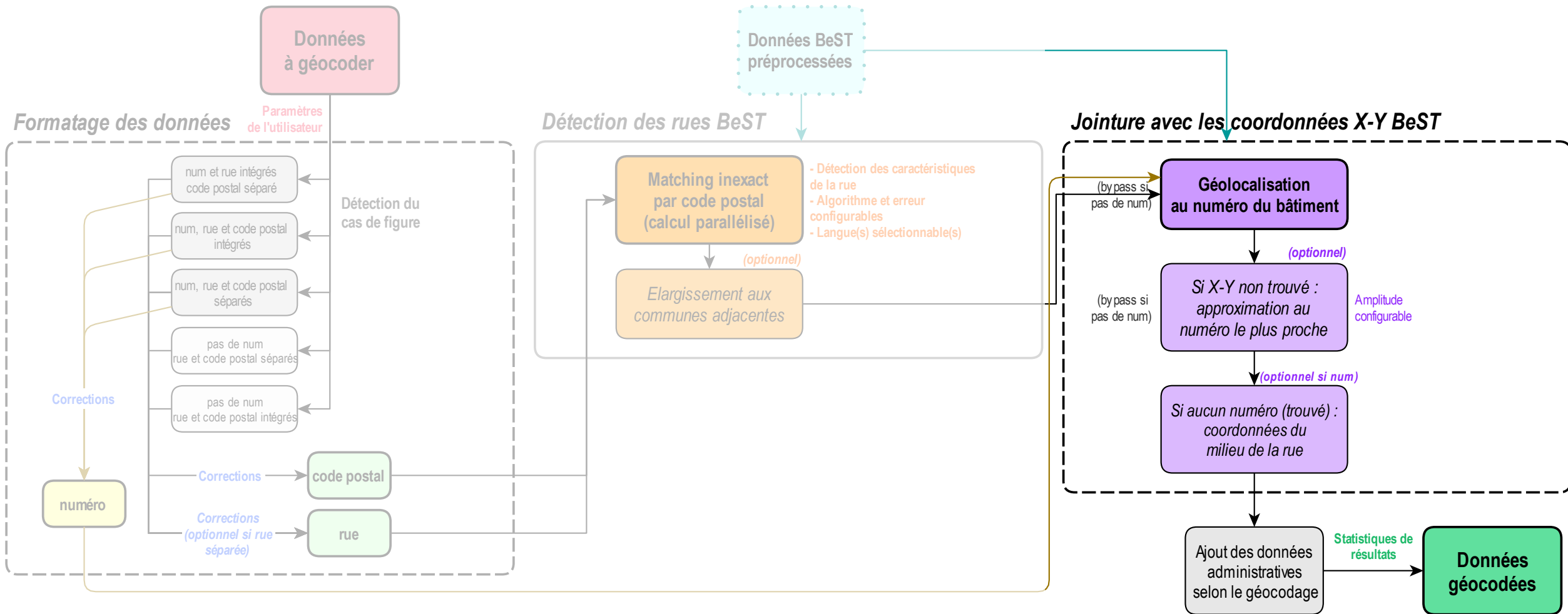
□ Autres communes



Code postal placé par commune au centre des adresses avec le même code postal (filtre appliqué pour ne garder que les codes postaux avec au moins 200 adresses dans la même commune). Taille des code postaux proportionnel au logarithme du nombre d'adresses.

2. LOGIQUE DE PHACOCHR

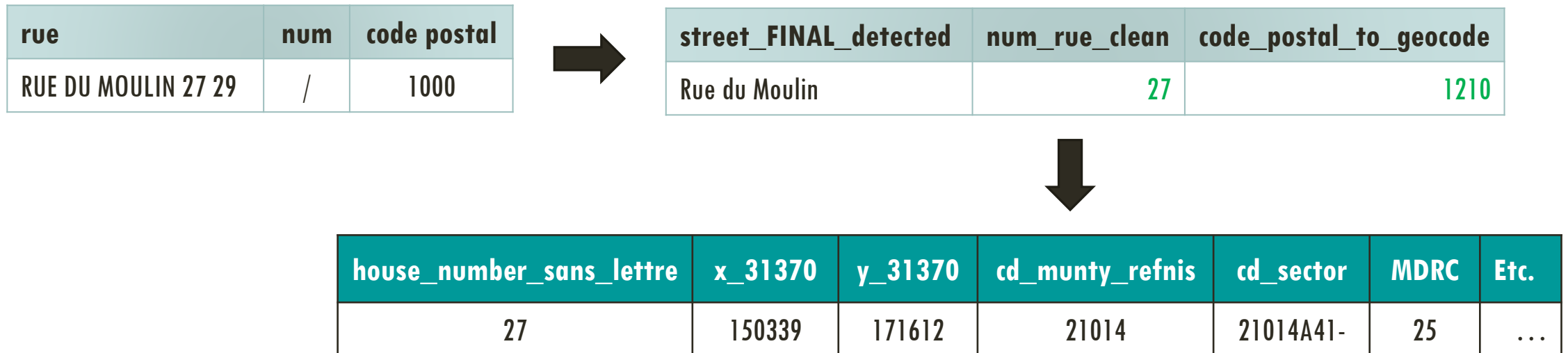
3) JOINTURE AVEC LES COORDONNÉES BEST



2. LOGIQUE DE PHACOCHR

3) JOINTURE AVEC LES COORDONNÉES BEST

Une fois les rues trouvées, il est désormais possible de réaliser une **jointure exacte** avec les données BeST géolocalisées au niveau du numéro. Des informations administratives (Statbel, Urbis) sont également jointes aux coordonnées X-Y.



















































2. LOGIQUE DE PHACOCHR

3) JOINTURE AVEC LES COORDONNÉES BEST

La jointure exacte nécessite le chargement des données BeST. Elles sont volumineuses : elles comprennent +/- 4.200.000 lignes. Cette étape est le deuxième goulot d'étranglement de la fonction.

Pour diminuer le temps de chargement, nous avons simplifié les données et scindé les fichiers selon les 43 arrondissements belges : seuls les arrondissements qui comprennent les adresses à géocoder sont chargés et réunis à la volée dans R.

 belgium_street_abv_PREPROCESSED.csv	 data_arrond_PREPROCESSED_35.csv	 data_arrond_PREPROCESSED_53.csv	 data_arrond_PREPROCESSED_81.csv
 data_arrond_PREPROCESSED_11.csv	 data_arrond_PREPROCESSED_36.csv	 data_arrond_PREPROCESSED_55.csv	 data_arrond_PREPROCESSED_82.csv
 data_arrond_PREPROCESSED_12.csv	 data_arrond_PREPROCESSED_37.csv	 data_arrond_PREPROCESSED_56.csv	 data_arrond_PREPROCESSED_83.csv
 data_arrond_PREPROCESSED_13.csv	 data_arrond_PREPROCESSED_38.csv	 data_arrond_PREPROCESSED_57.csv	 data_arrond_PREPROCESSED_84.csv
 data_arrond_PREPROCESSED_21.csv	 data_arrond_PREPROCESSED_41.csv	 data_arrond_PREPROCESSED_58.csv	 data_arrond_PREPROCESSED_85.csv
 data_arrond_PREPROCESSED_23.csv	 data_arrond_PREPROCESSED_42.csv	 data_arrond_PREPROCESSED_61.csv	 data_arrond_PREPROCESSED_91.csv
 data_arrond_PREPROCESSED_24.csv	 data_arrond_PREPROCESSED_43.csv	 data_arrond_PREPROCESSED_62.csv	 data_arrond_PREPROCESSED_92.csv
 data_arrond_PREPROCESSED_25.csv	 data_arrond_PREPROCESSED_44.csv	 data_arrond_PREPROCESSED_63.csv	 data_arrond_PREPROCESSED_93.csv
 data_arrond_PREPROCESSED_31.csv	 data_arrond_PREPROCESSED_45.csv	 data_arrond_PREPROCESSED_64.csv	 table_commune_adjacentes.csv
 data_arrond_PREPROCESSED_32.csv	 data_arrond_PREPROCESSED_46.csv	 data_arrond_PREPROCESSED_71.csv	 table_INS_recod_code_postal.csv
 data_arrond_PREPROCESSED_33.csv	 data_arrond_PREPROCESSED_51.csv	 data_arrond_PREPROCESSED_72.csv	 table_postal_arrond.csv
 data_arrond_PREPROCESSED_34.csv	 data_arrond_PREPROCESSED_52.csv	 data_arrond_PREPROCESSED_73.csv	 table_postal_com_name.csv

2. LOGIQUE DE PHACOCHR

3) JOINTURE AVEC LES COORDONNÉES BEST

Cependant, il arrive que PhacochR ne trouve pas les coordonnées X-Y du numéro dans BeST.

2 réponses à ce problème :

A. Approximation du numéro

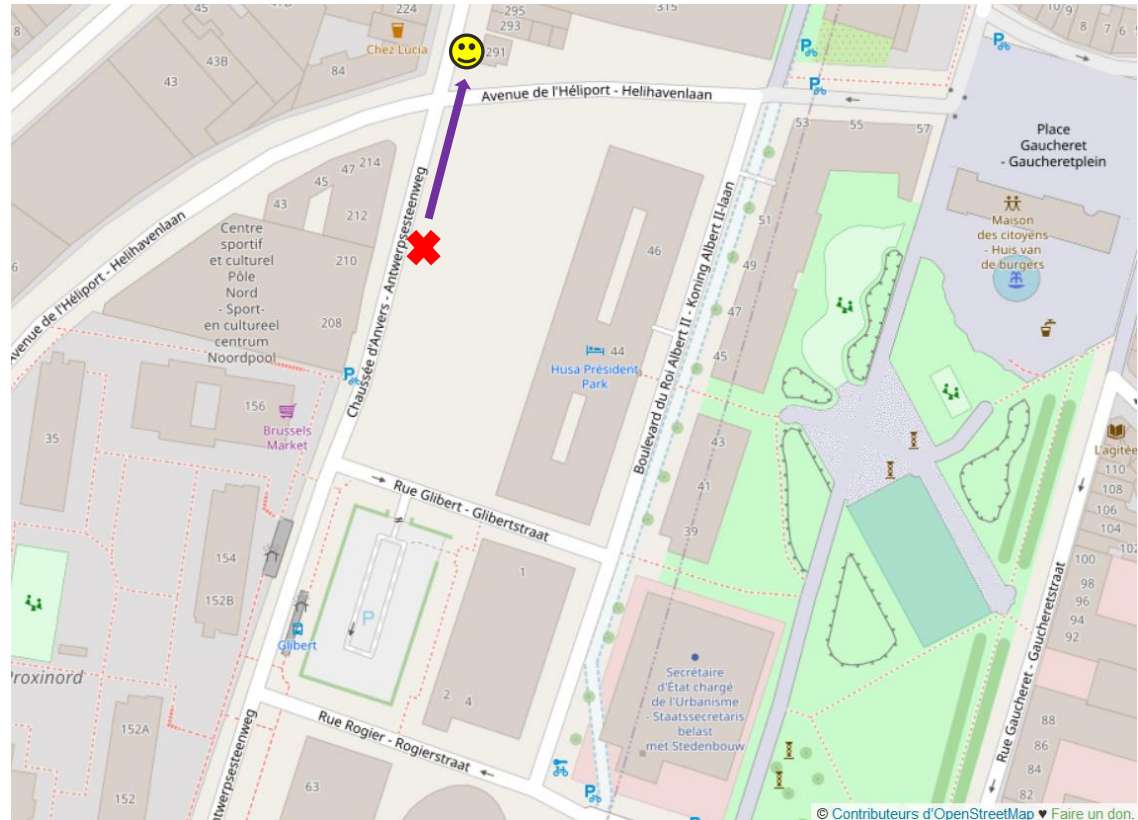
Il approxime au numéro le plus proche (*maximum configurable*) de préférence du même côté de la rue.

Plusieurs objectifs :

- Faire face aux erreurs d'encodage
- Faire face au manque des données wallonnes dans BeST
- Trouver des adresses qui n'existent plus

=> Exemple de la friterie

« J. Vandernot » au 223 Chaussée d'Anvers, 1000.



2. LOGIQUE DE PHACOCHR

3) JOINTURE AVEC LES COORDONNÉES BEST

B. Milieu de la rue

Si PhacochR ne trouve pas les coordonnées au niveau du bâtiment, il peut indiquer les coordonnées du milieu de la rue (*optionnel*). Il s'agit de l'une des informations ajoutée pendant la tranformation des données BeST.

street_FINAL_detected	code_postal_to_geocode
Avenue Mutsaard	1020



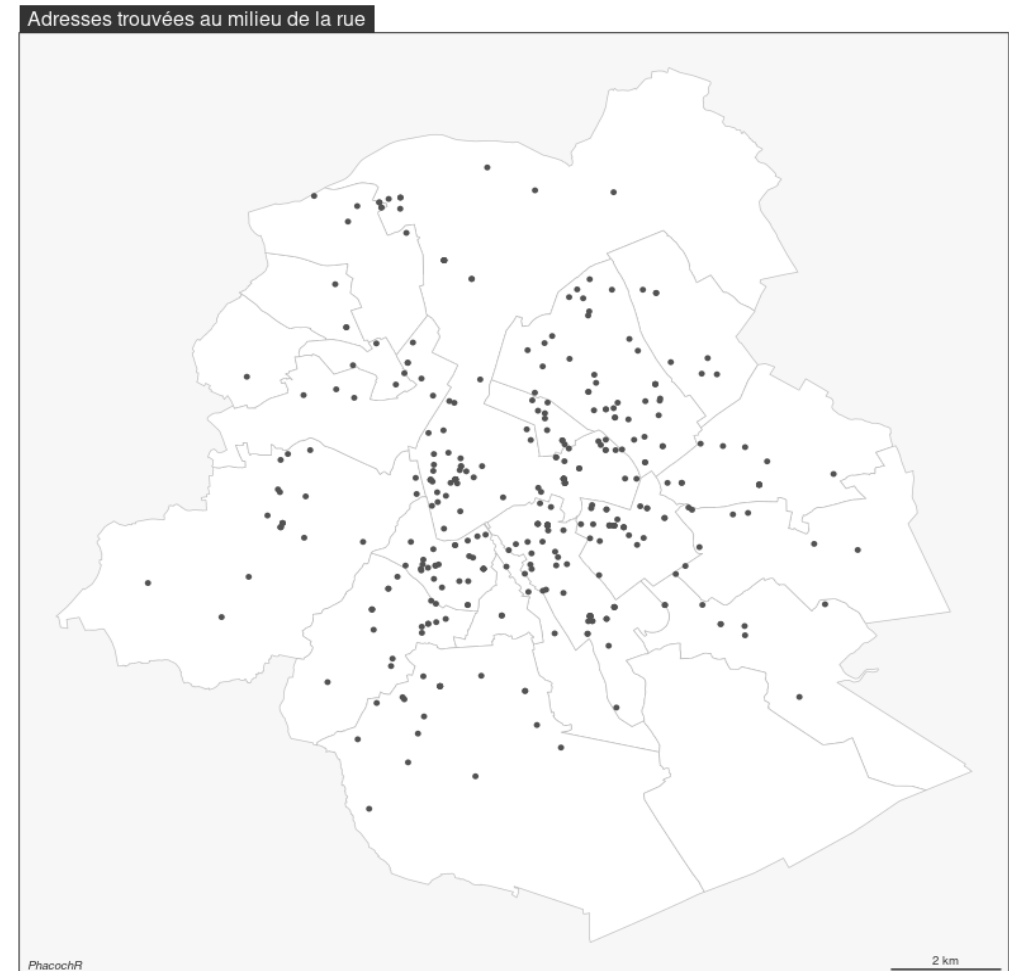
mid_num	mid_x_31370	mid_y_31370	mid_cd_sector
38	149108	176270	21004E233

2. LOGIQUE DE PHACOCHR

3) JOINTURE AVEC LES COORDONNÉES BEST

B. Milieu de la rue

Voici un exemple de localisation d'adresses ne possédant pas de numéro : adresses de co-living récoltées sur internet (**Charlotte Casier, 2023**)



3. PERFORMANCES

1) FIABILITÉ

Nous avons comparé les résultats de 11 géocodeurs sur un échantillon aléatoire de 22.000 adresses écrites à la main, d'opérateurs économiques en Belgique (y compris copropriétés).

=> En l'absence de coordonnée réelle, on fait l'hypothèse que le point médian tend vers la vraie coordonnée.

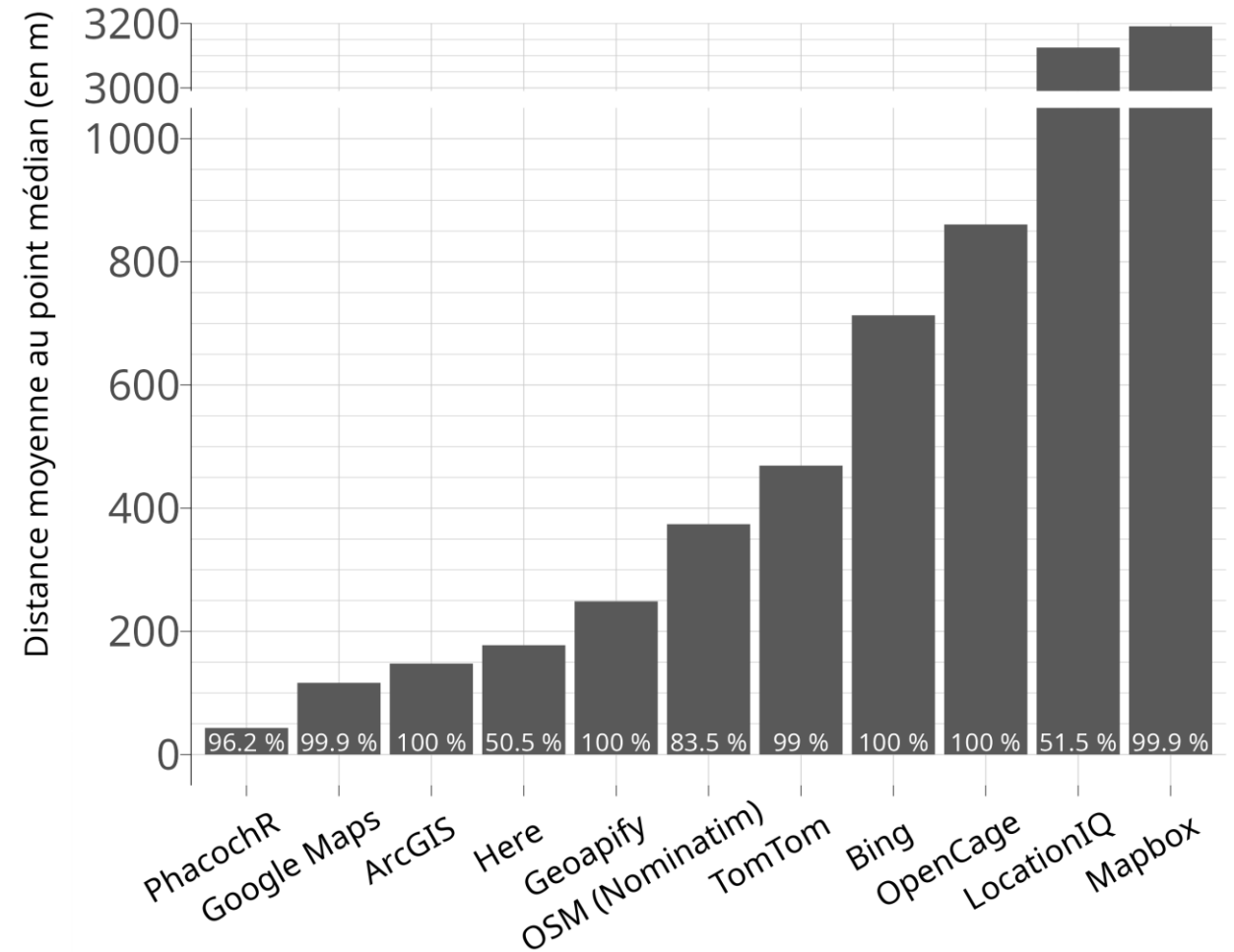
PhacochR est ainsi en moyenne le plus proche de ce point médian (43m), suivi par Google Maps (116m).

PhacochR trouve 96,2% des adresses.

Comparaison de géocodeurs

Test réalisé sur 22000 adresses issue de la Banque-Carrefour des Entreprises (Belgique):

- Distance au point médian
- Match rate (% des adresses trouvées)



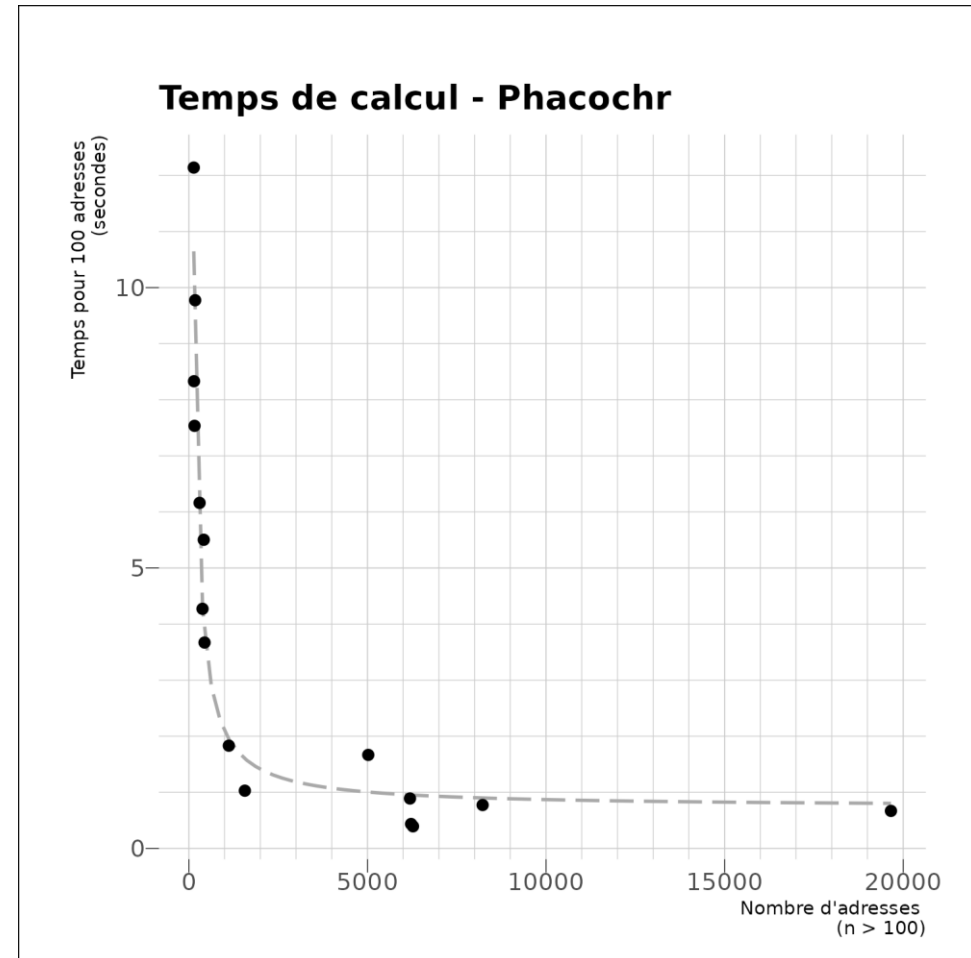
3. PERFORMANCES

2) TEMPS DE CALCUL

PhacochR est rapide pour du géocodage en batch.

Relation $1/x$ entre le temps de calcul et le nombre d'adresses à trouver :

- Charger les données → lent pour peu d'adresses (minimum $\sim 15s$)
- Rapide pour beaucoup d'adresses (à partir d'environ 1000)

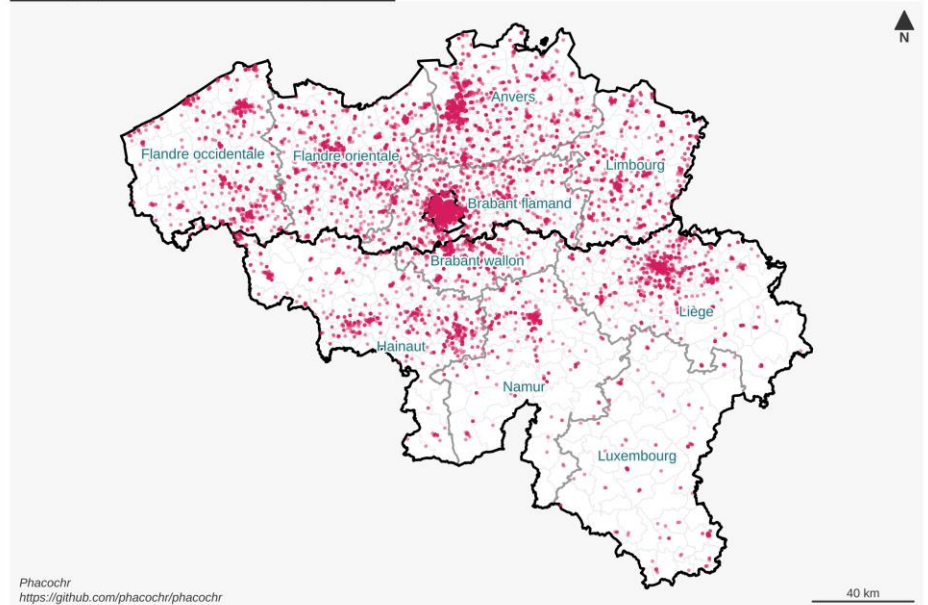


4. CARTOGRAPHIE

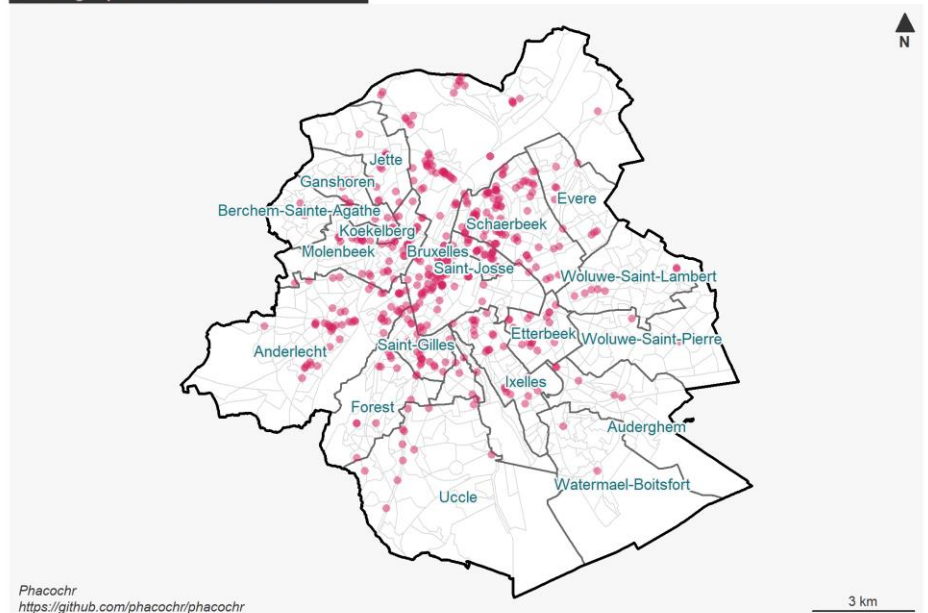
PhacochR intègre également la fonction `phaco_map_s()` qui permet de cartographier directement les résultats du géocodage.

La fonction repose sur le package `mapsf`, simple et léger. Les données vectorielles avec les entités géographiques belges sont contenues dans les données du package.

Cartographie : Dentistes en Belgique



Cartographie : Snacks à Bruxelles



5. QUESTIONNEMENTS

PhacochR est notre premier package !

Nous avons tâtonné, et certaines questions restent ouvertes :

1. **L'implémentation des données séparées est-elle adéquate ?** Quelles solutions ont été apportées par d'autres développeurs face à cette question ?
2. **L'optimisation du chargement des données est-elle possible ?** La logique suivie implique de mauvaises performances pour le géocodage d'un petit nombre d'adresses.
3. **Les dépendances du package sont-elles problématiques ?** Par facilité de codage et pour une meilleure lisibilité du code, nous avons utilisé les packages du **tidyverse**. Quelle implication dans la durée des dépendances ?