

PHACOCHR

Outil de géocodage libre pour la Belgique

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



PLAN DE LA PRÉSENTATION

1. Introduction

2. Logique

- Formatage des données
- Détection des rues
- Jointure avec les coordonnées BeST

3. Interface Shiny

4. Performances

5. Développements futurs

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 une **librairie R** qui est constituée d'un code sous licence libre ;
- Il repose sur les **données publiques** BeST Address (compilation de Urbis, Icar et Crab).

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

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

=> Démonstration sur R

1. INTRODUCTION : PHACOCHR, POURQUOI ?

Le géocodage est un processus complexe

Pour géocoder, on utilise souvent Google Map ou OpenStreetMap (via Nominatim).

Ces solutions présentent différents problèmes :

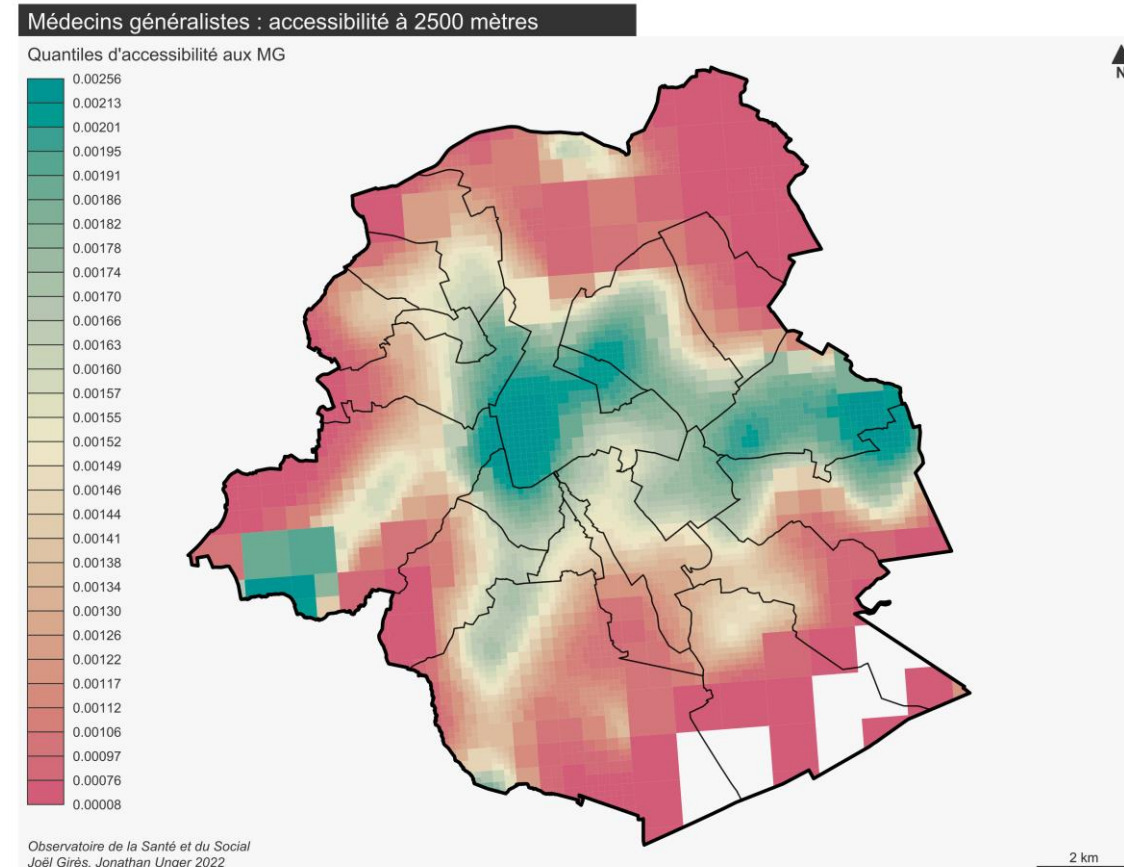
- **Google Map** est propriétaire, payant et utilise les données à des fins problématiques ;
- **Nominatim** est libre mais trouve peu d'adresses si elles sont mal orthographiées ;
- Ils constituent des dispositifs lourds et lents.

PhacochR comme alternative

Dans le cadre d'une recherche sur la pénurie de médecins généraliste, PhacochR a été développé. Il est construit pour être léger, portable et utilisable en local, ce qui permet **confidentialité** et **rapidité**.

Ce que PhacochR n'est pas

- But de recherche : peu performant pour trouver individuellement des adresses (usage prévu : géocodage en masse).
- Pas d'API ou webservice pour s'y connecter (mais outil en ligne).

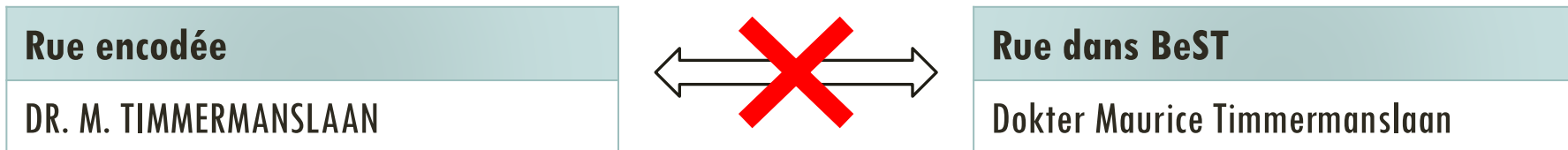


2. LOGIQUE DE PHACOCHR

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.

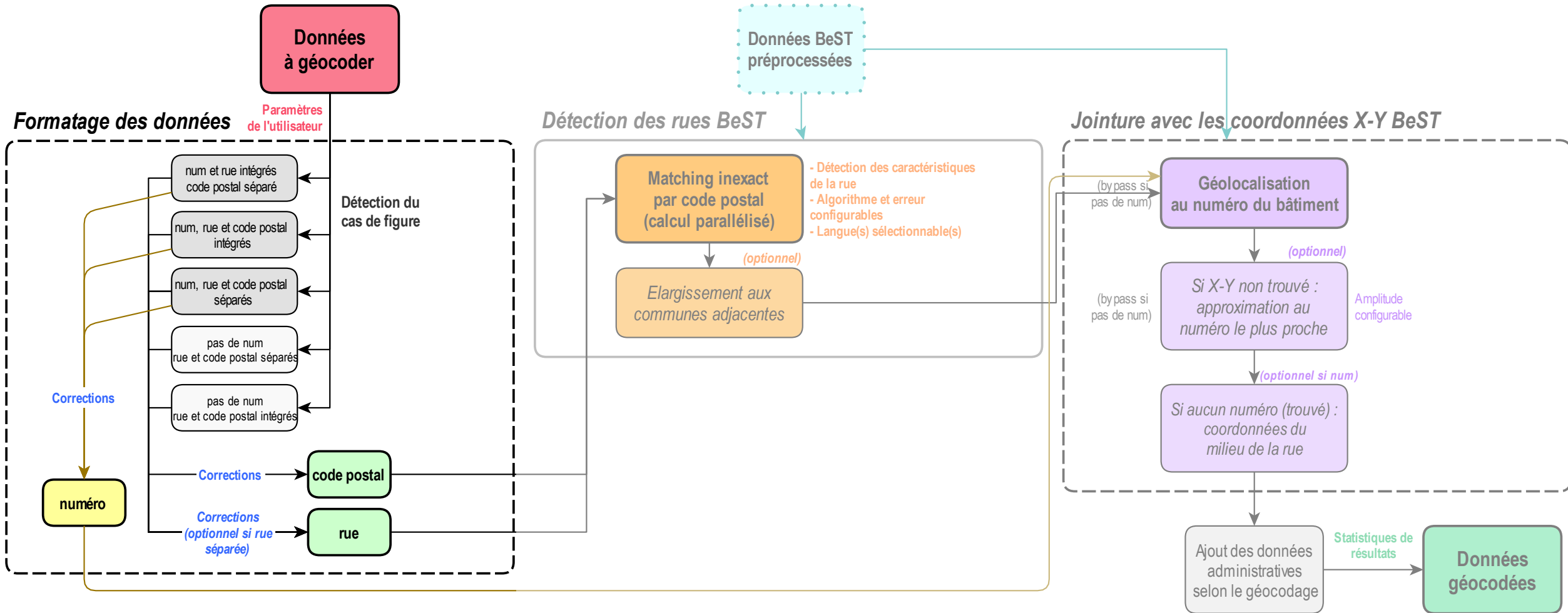
Le problème : l'orthographe encodée ne correspond jamais exactement à l'orthographe « officielle » (abréviations, coquilles, manque le mot « rue », etc.)



=> Le but de PhacochR est spécifiquement de rendre cette jointure possible, en faisant une **jointure inexacte** (fautes permises) !

2. LOGIQUE DE PHACOCHR

1) FORMATAGE DES DONNÉES



2. LOGIQUE DE PHACOCHR

1) FORMATAGE DES DONNÉES

La première chose que fait PhacochR est de nettoyer et corriger les adresses. La nature du recodage est indiquée dans la colonne **recode**.

rue
Rue Sous Lt. Catoire(D)
KON. ELISABETHPLEIN
Av. de Tervueren 116 BP 14
Torhoutsesteenweg 44/6.03
CHEE DE ST JOB
Kouterstraat(LOO)
de l'Ecureuil,
Burg. Gillonlaan



rue_recoded	recode
Rue Sous Lieutenant Catoire	parenthese ; Lieutenant
Koningin ELISABETHPLEIN	koning
Avenue de Tervueren	BP_CP ; num ; avenue
Torhoutsesteenweg	slash ; num
Chaussee DE Saint JOB	Saint ; chaussee
Kouterstraat	parenthese
Rue de l'Ecureuil	virgule ; Rue
Burgemeester Gillonlaan	Burgemeester

2. LOGIQUE DE PHACOCHR

1) FORMATAGE DES DONNÉES

PhacochR détecte les variables nécessaires (numéro, code postal) si besoin.

Quelle que soit la manière dont il est encodé : **le code postal est nécessaire !**

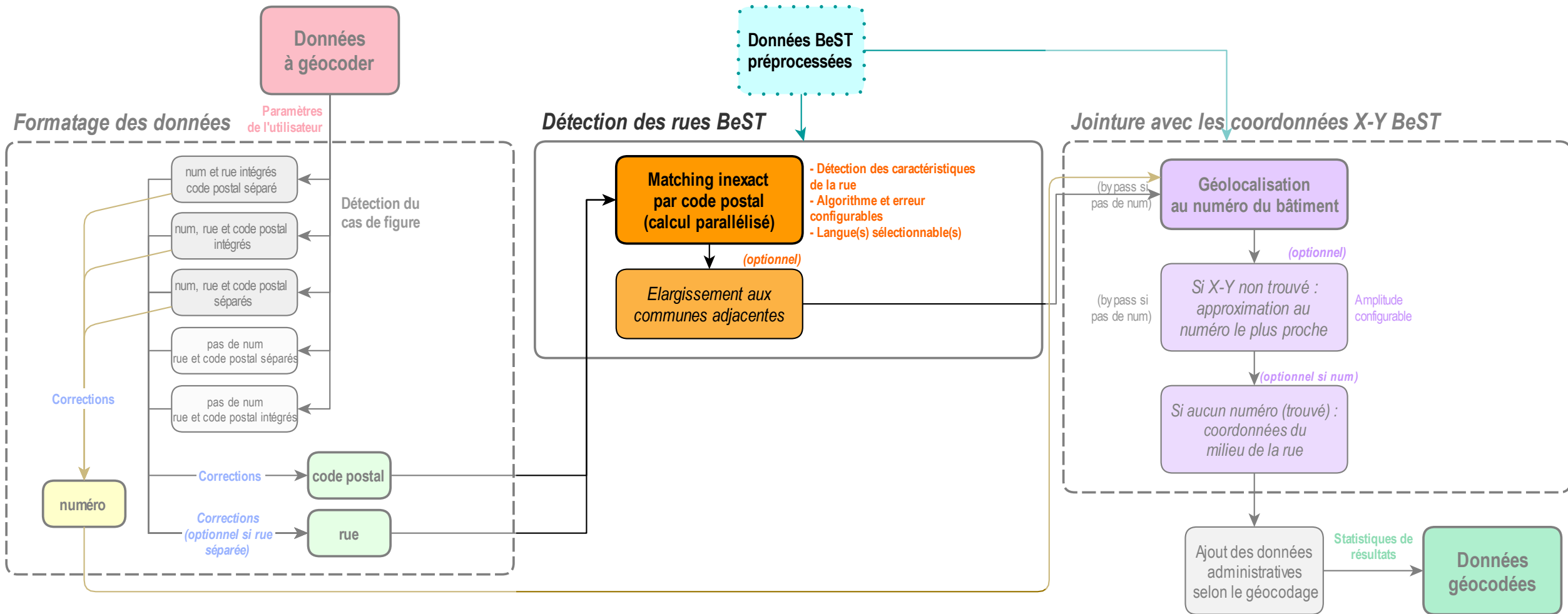
num_rue_code_postal
15, rue notre-seigneur Bruxelles 1000
Boulevard du Triomphe, 153 Bte 7614 Ixelles 1050
Rue Royale, 344 bte 3.2 Schaerbeek 1030
Promenade de l'Alma, 49/312 BRUXELLES 1200
45 Rue des palmiers Woluwe-Saint-Pierre 1150
Rue Picard 68 Sint-Jans-Molenbeek 1080



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

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 (processus multi-threadé pour la vitesse)

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

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 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

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

On peut indiquer à PhacochR de ne faire la détection que dans une seule langue.

Cas de figure : géocodage à Bruxelles (où les rues sont systématiquement traduites dans les 2 langues) dans une base de données dont on sait que les adresses sont *exclusivement* encodées en FR.

postal_id	street_FINAL_detected	langue_FINAL_detected
1000	Rue du Pavillon	FR
1000	Paviljoenstraat	NL
1000	Rue de l'Homme Chrétien	FR
1000	Kerstenmannekenstraat	NL
1000	Rue du Parlement	FR
1000	Parlementsstraat	NL
1000	Rue des Riches Claires	FR
1000	Rijkeklarenstraat	NL
1000	Rue de l'Ommegang	FR
1000	Ommegangstraat	NL
1000	Rue de la Flèche	FR
1000	Pijlstraat	NL
1000	Rue Watteeu	FR
1000	Watteestraat	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

On peut indiquer à PhacochR de ne faire la détection que dans une seule langue.

Cas de figure : géocodage à Bruxelles (où les rues sont systématiquement traduites dans les 2 langues) dans une base de données dont on sait que les adresses sont *exclusivement* encodées en FR.

=> On peut alors choisir de ne rechercher que parmi les rues en FR, permettant un gain de vitesse et un moindre taux d'erreur (paramètre "lang_encoded").

postal_id	street_FINAL_detected	langue_FINAL_detected
1000	Rue du Pavillon	FR
1000	Paviljoenstraat	NL
1000	Rue de l'Homme Chrétien	FR
1000	Kerstenmannkensstraat	NL
1000	Rue du Parlement	FR
1000	Parlementsstraat	NL
1000	Rue des Riches Claires	FR
1000	Rijkeklarenstraat	NL
1000	Rue de l'Ommegang	FR
1000	Ommegangstraat	NL
1000	Rue de la Flèche	FR
1000	Pijlstraat	NL
1000	Rue Watteeu	FR
1000	Watteeustraat	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

Autre exemple d'adresse (en néerlandais) :
« Romeniestraat », numéro 6 au code postal « 8400 ».

Note : PhacochR sait géocoder des adresses en FR, NL et DE, et corriger des adresses en FR et NL.

rue	num	code postal
Romeniestraat	6	8400



postal_id	street_FINAL_detected	langue_FINAL_detected
8400	Fuutstraat	NL
8400	Smientstraat	NL
8400	Bronstraat	NL
8400	Broederlijkheidstraat	NL
8400	Veerbootstraat	NL
8400	Brugstraat	NL
8400	Brugsesteenweg	NL
8400	Brouwerijstraat	NL
8400	Buskruitstraat	NL
8400	Buitenpad	NL
8400	Brusselstraat	NL
8400	Camelialaan	NL
8400	Caelfstraat	NL
8400	Werkzaamheidstraat	NL
8400	Chaletstraat	NL
8400	Westdiepstraat	NL
8400	Cederdreef	NL
8400	Westhinderstraat	NL
8400	Cardynplein	NL
8400	Canadaplein	NL
8400	Cirkelstraat	NL
8400	Chrysantenstraat	NL
8400	Christinastraat	NL
8400	Werktuigkundigenstraat	NL
8400	Westlaan	NL
8400	Wezellaan	NL
8400	Rietgansstraat	NL
8400	Kolgansstraat	NL
8400	Waddenlaan	NL
8400	Wilgenlaan	NL
8400	Zoutziedersstraat	NL
Etc...		

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

Autre exemple d'adresse (en néerlandais) :

« Romeniestraat », numéro 6 au code postal « 8400 ».

Dans ce cas-ci, on voit un ex-aequo dans la plus petite erreur entre « Roemeniëstraat » et « Romestraat ».

rue	num	code postal
Romeniestraat	6	8400



street_FINAL_detected	dist_fuzzy
Reestraat	4
Roemeniëstraat	3
Romestraat	3

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

Autre exemple d'adresse (en néerlandais) :
« Romeniestraat », numéro 6 au code postal « 8400 ».

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.

rue	num	code postal
Romeniestraat	6	8400



street_FINAL_detected	dist_fuzzy	Jaro-Winkler
Reestraat	4	
Roemeniëstraat	3	0.070329670
Romestraat	3	0.046153846

2. LOGIQUE DE PHACOCHR

2) DÉTECTION DES RUES

Afin de détecter les rues contenant des prénoms abrégés, nous avons recréé 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

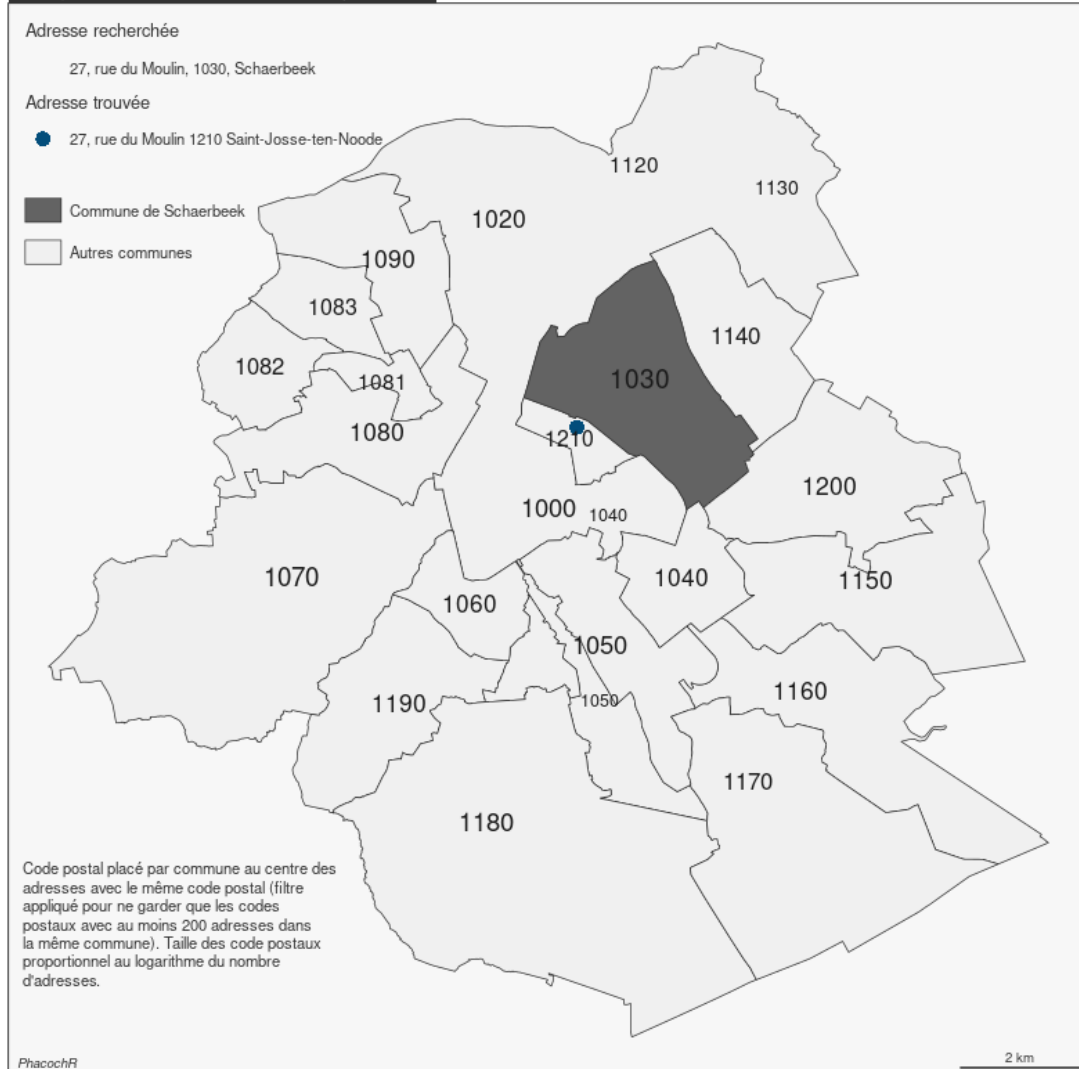
27, rue du Moulin, 1030, Schaerbeek

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

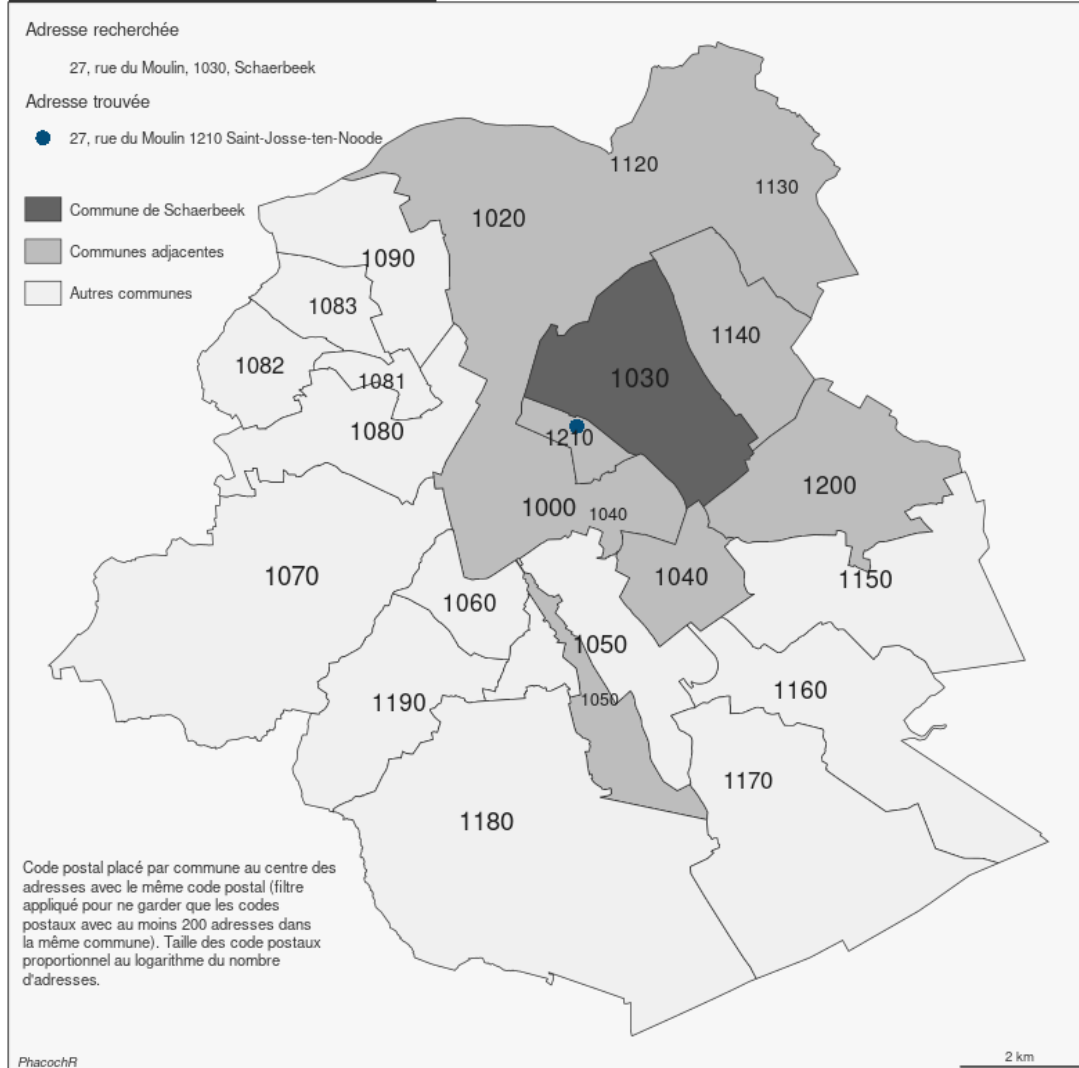
Adresse trouvée

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

■ Commune de Schaerbeek

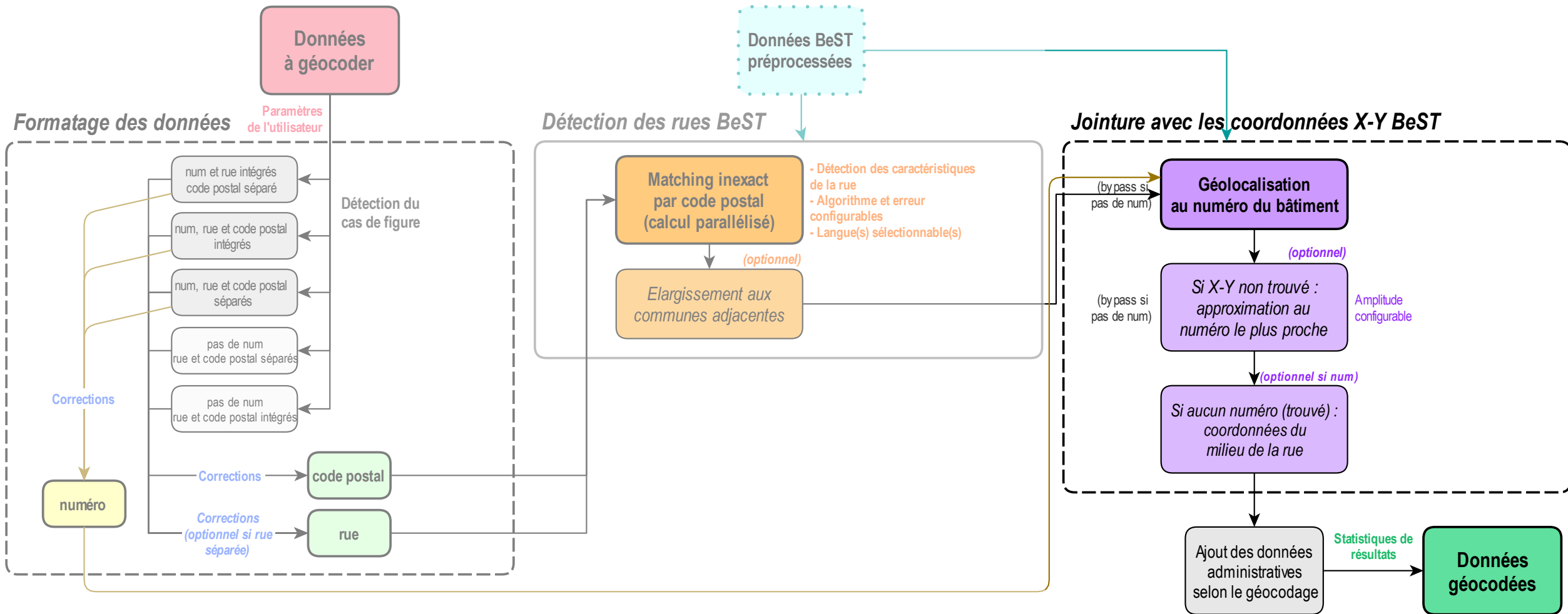
■ Communes adjacentes

□ Autres communes



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.

rue	num	code postal
RUE DU MOULIN 27 29	/	1000



street_FINAL_detected	num_rue_clean	code_postal_to_geocode
Rue du Moulin	27	1210



house_number_sans_lettre	x_31370	y_31370	cd_munty_refnis	cd_sector	MDRC	Etc.
27	150339	171612	21014	21014A41-	25	...

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

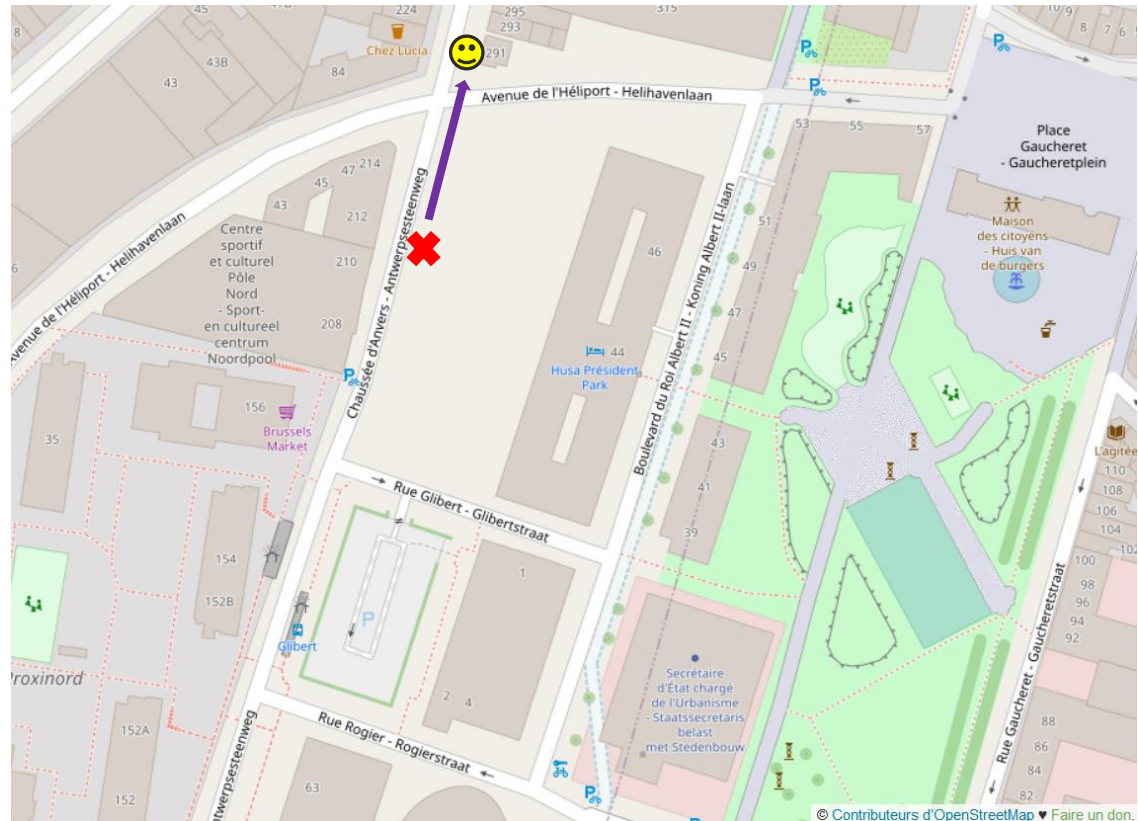
Il approxime au numéro le plus proche
(maximum configurable)

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*).

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

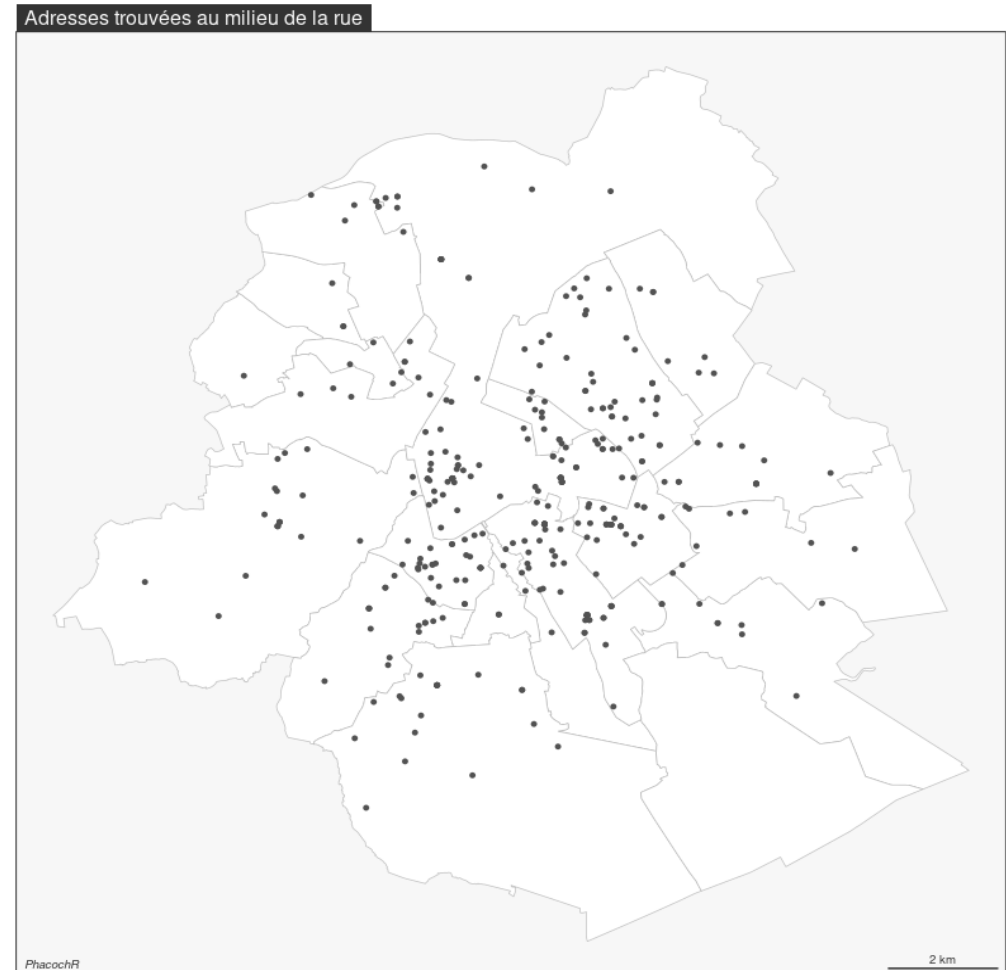
2. LOGIQUE DE PHACOCHR

3) JOINTURE AVEC LES COORDONNÉES BEST

B. Milieu de la rue

Cas 1 : Si pas de numéro introduit dans les données à géocoder

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



2. LOGIQUE DE PHACOCHR

3) JOINTURE AVEC LES COORDONNÉES BEST

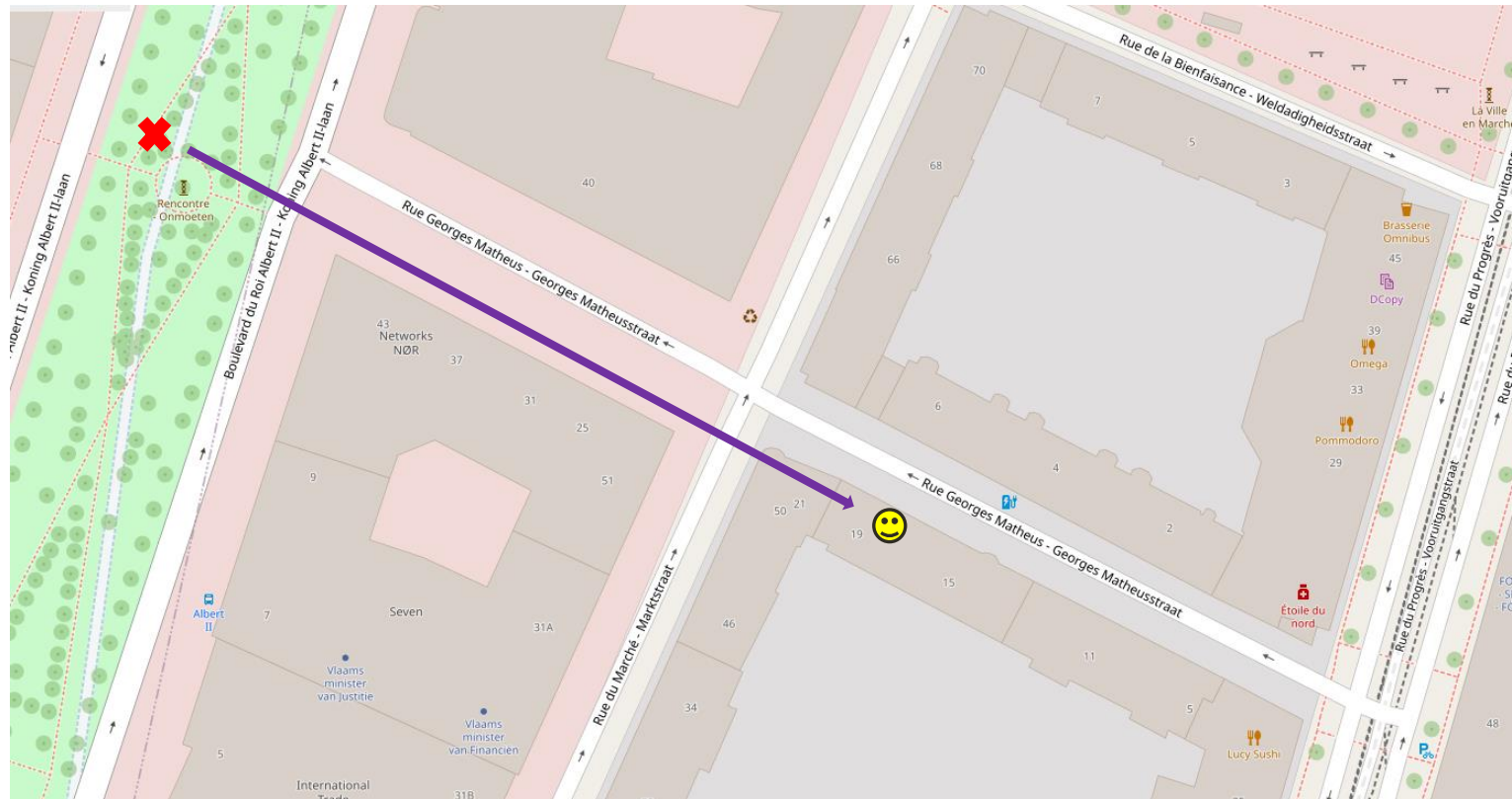
B. Milieu de la rue

Cas 2 : si aucun numéro trouvé dans la limite de l'approximation

=> Exemple de la friterie « Friture Henri » au 211 rue Georges Matheus 1000 Bruxelles.

Numéro max = 43 (trop loin de la limite d'approximation par défaut)

→ **milieu de la rue : 19**



3. INTERFACE SHINY


PhacochR a une application web que l'on peut utiliser à cette adresse :

https://PhacochR.github.io/PhacochR/articles/shiny_app.html

phacochr 0.9.0.4

[Reference](#)

[Géocoder](#)

 github

Importation des données

Géocodage

Cartes

Export

Type de fichier à géocoder

☒ .csv ☐ .xlsx

Fichier à géocoder (max 300 MB)

Browse...

No file selected

☒ Entête (header)

Separateur de champs

☐ , ☒ ; ☐ | ☐ Tab ☐ ~

Guillement pour les champs texte

☒ Pas de guillemet

☐ Doubles guillemets (" ")

☐ Simples guillemets (' ')

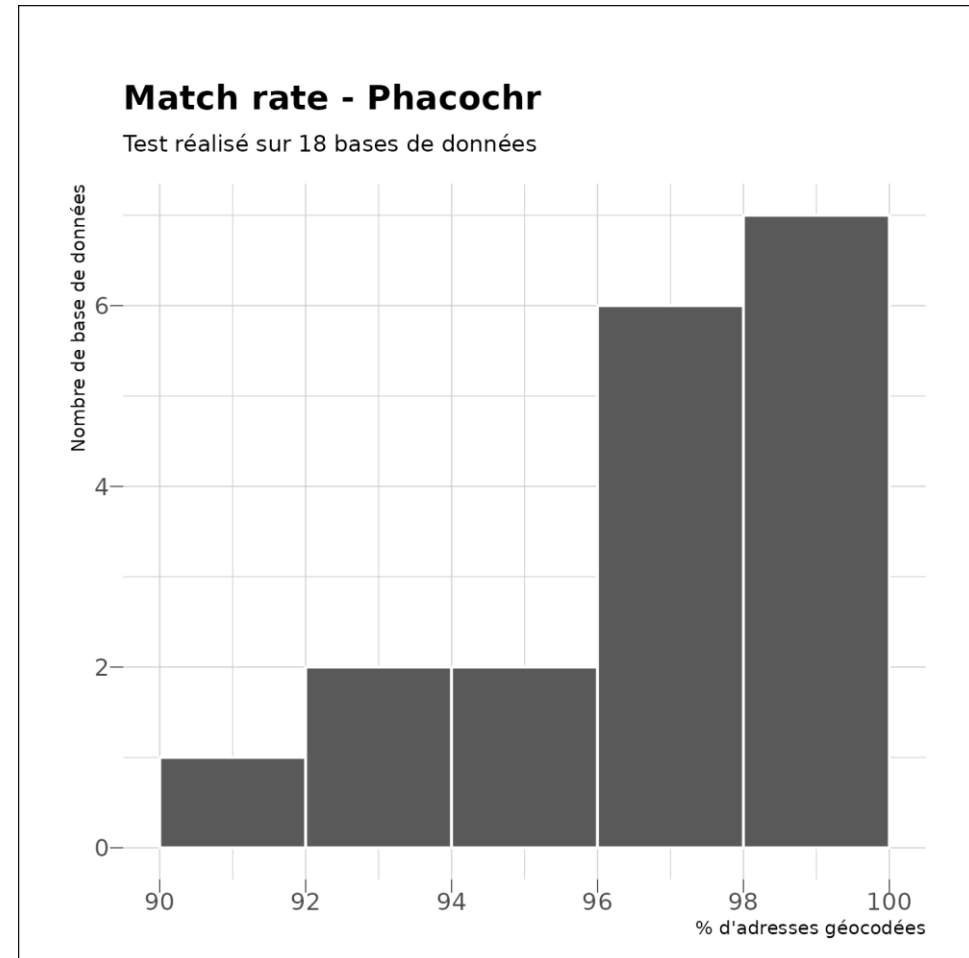
Aperçu des données à géocoder

nom	rue	num	code_postal
Observatoire de la Santé et du Social	rue Beilliard	71	1040
ULB	avenue Antoine Depage	30	1000

4. PERFORMANCES

1) CAPACITÉ À TROUVER

PhacochR possède une bonne capacité à trouver les adresses. Sur un ensemble de 18 bases de données réelles, la médiane du pourcentage d'adresses trouvées est de **97%**.



4. PERFORMANCES

2) TEMPS DE CALCUL

PhacochR est rapide pour du géocodage en batch.

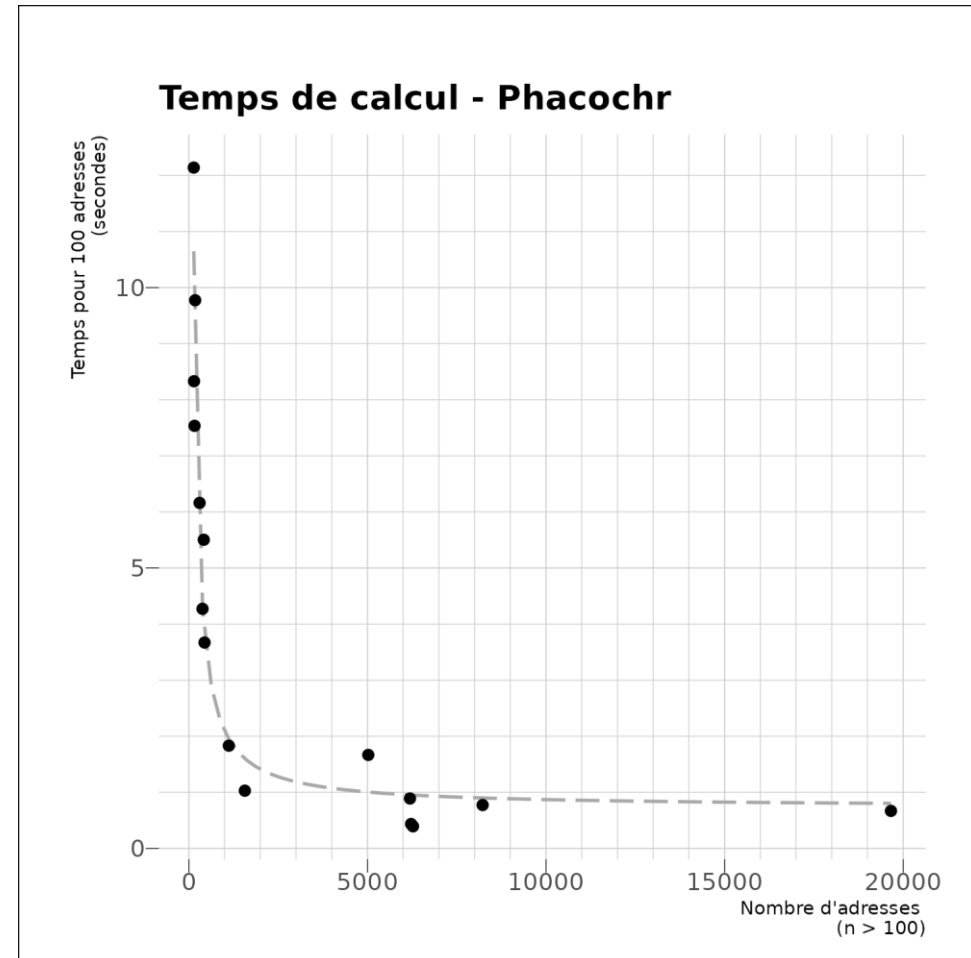
Relation $1/x$ entre le temps de calcul et le nombre d'adresses à trouver :

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

Test sur 20000 adresses :

I. Nominatim (local) $\sim 1h40$

II. PhacochR $\sim 1m30$

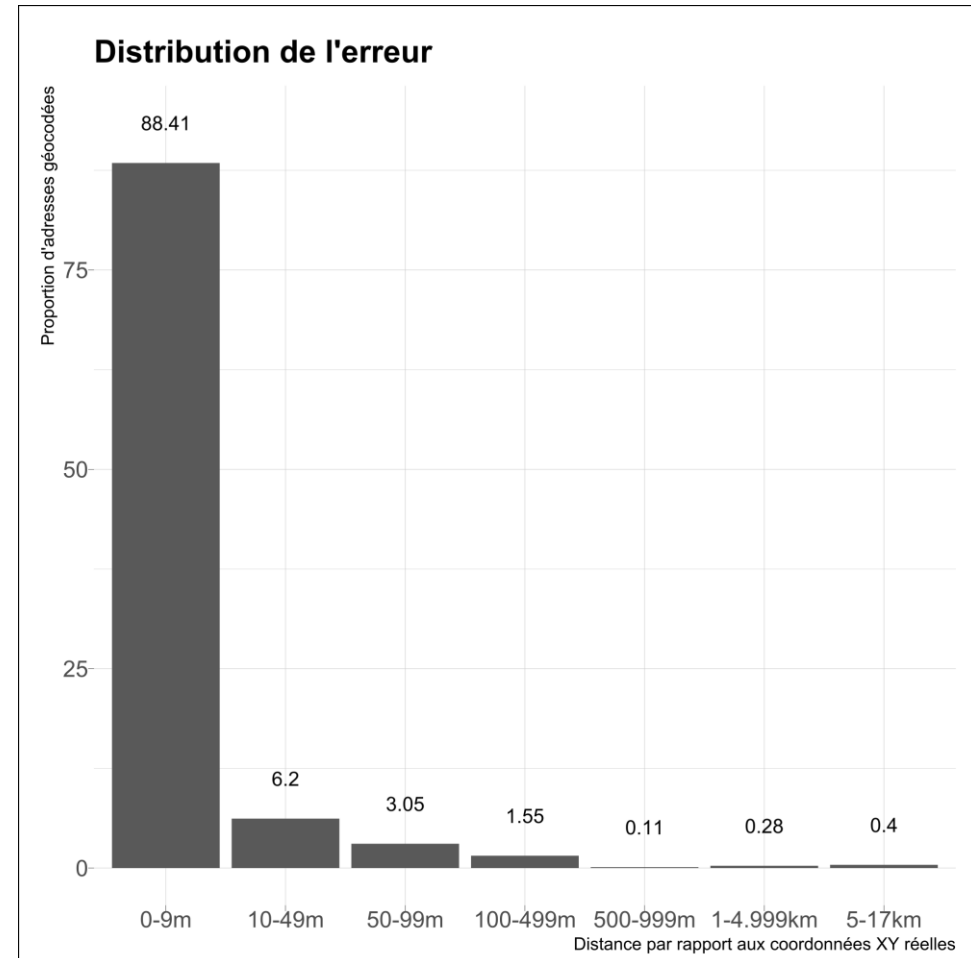


4. PERFORMANCES

3) FIABILITÉ

Test sur 2 bases de données dont on connaît les coordonnées.

97,6% des adresses sont localisées à **moins de 100m** de leurs coordonnées “réelles”, montrant un degré de fiabilité tout à fait satisfaisant.



4. PERFORMANCES

4) ESPACE DISQUE NÉCESSAIRE

La vitesse de PhacochR tient notamment à sa légèreté en comparaison d'autres géocodeurs.

I. **PhacochR** ~ 280 Mo

- BeST ~ 152 Mo
237Mo à télécharger, 1 Go une fois extrait, puis reformatées pour être les plus légères possibles
- Statbel ~ 127Mo

II. **Nominatim** (en local) ~ 13 Go

5. DÉVELOPPEMENTS FUTURS

Besoin d'utilisateurs

Il nous serait utile que de nouveaux utilisateurs testent le programme sur d'autres données pour faire remonter des erreurs inattendues, impensés, demandes de nouvelles fonctions...

Amélioration déjà prévues

Nous avons déjà quelques projets d'amélioration :

- Améliorer les corrections orthographiques (nettoyer le code et affiner la fonction) ;
- Harmonisation et simplification des noms des colonnes et arguments des fonctions ;
- Créer une interface Shiny en local ;
- Traduire la documentation en NL ou EN ;
- *Eventuellement (si demande) : optimiser la vitesse de traitement pour géocoder des adresses seules ?*

Code libre

Le code est par ailleurs libre pour que n'importe qui s'en saisisse, le copie, l'améliore... !