Scraping…

…or not scraping

1

Construire une liste de SIRET possibles

données RP et appel à l'API Sirene

2

Calculer la « pertinence » des SIRET

deux indicateurs : distance sur les libellés de voies + analyse des champs entre SIRENE et RP

3

Sélectionner les meilleurs SIRET

tri et sélection des 3 « meilleurs » SIRET

4

Affiner la sélection par géolocalisation

avec croisement des données RP et SIRUS, puis avec GoogleWays

Entre 20 et 50 % de concordance…

… sur un maximum de 4 000 BI

Deux approches complémentaires :

- API Sirene pour trouver des échos
- Géolocalisation pour affiner

Scraping ☹

API Sirene : performance « variable », codes NAF sans « . »

Majuscules versus minuscules avec / sans accents