# Sirenisation ADM

**Version 1** :

Web scraping du site societe.com

L’analyse des résultats du script V0 montre qu’il peut y avoir plusieurs defendeur/demandeur et que dans ce cas, ils sont séparés par des point-virgules.

* Patch pour prendre en compte la multiplicité des demandeurs/defendeurs

Certains fichiers xml ne peuvent pas être parsés par ElementTree. La raison est que les balises legalmetrics ont été ajouté en dehors de la balise root « document » -> la norme XML n’est donc plus respectée.

* Lecture des fichiers en textes simples et détection des balises par expressions régulières

**Version 2** :

Utilisation d’Elasticsearch comme moteur de recherche

La constitution de la base de données prend environ 3 heures. Mais, les requêtes sont ensuite très rapides.

Exemples d’anomalies :

* SOCIETE SODIMAT
* COMMUNAUTE DE COMMUNES DE L'ILE DE NOIRMOUTIERS
* ﻿CENTRE HOSPITALIER SUD GIRONDE
* SOCIETE LUC SUR MER LOISIRS

**Version 3** (non concluant):

La réindexation de la base siren prend environ 20min.

Personnalisation des paramètres de la métrique de similarité BM25 :

* k1 : 1,2 reste inchangé
* b : est réduit à 0.1 afin de réduire l’impact de la longueur du champ

La gestion de l’apostrophe n’est pas claire dans Elasticsearch

**Version 3 :**

* Détection des noms des départements dans le nom de l’entreprise -> ajout d’un critère de recherche sur le préfixe (2 ou 3 caractères) du code postal
* Détection du nom des communes dans le nom de l’entreprise -> ajout d’un critère de recherche sur LIBCOM\_pretty
* Détection des entreprises connues (complété pour les organismes publiques) -> ajout d’un critère de recherche en « match\_phrase » sur la raison\_sociale

**Version 4** :

* Sirenisation « manuelle » des « company\_name\_normalized » les plus courants (les 70 premiers, ce qui représente 15000 résultats sur 51000)
* Suppression des caractères « \* » et « ‘’ »
* Suppression de la commune « Ville » dans df\_commune (car trop de confusion)

**Version 4bis** :

* Export des résultats dans les fichiers XML (ajout des balises)