Journées de méthodologie statistique de l'Insee 2022

# Identification de collectivités territoriales en utilisant un moteur de recherche indexé à partir d'RMéS

*Théo LEROY (*)*

*(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale*

theo.leroy@insee.fr

**Mots-clés.** (6 maximum) : Code officiel géographique, RDF, moteur de recherche, ElasticSearch

**Domaines.** Codification automatique, Open data

## Résumé

Le besoin de coder des libellés géographiques dans le Code Officiel Géographique (COG) revient dans de nombreuses production de l'Insee (le recensement de la population, le traitement des déclarations sociales nominatives ou la constitution de Fidéli). Dans ces cas, la codification est assurée par l'outil mutualisé Sicore. Cependant, chaque année, la mise en oeuvre est particulièrement lourde car elle demande une actualisation en partie manuelle des bases de connaissances Sicore à partir du COG. De plus, Sicore ne répond pas au besoin grandissant en matière d'autocomplétion car ne permet de coder que sur un libellé complet.

Dans ce contexte, cet article présente une alternative aux environnements Sicore pour la géographie (communes et pays) reposant sur le moteur de recherche ElasticSearch. Les moteurs de recherche permettent de retrouver des documents selon un critère de similarité défini en fonction du besoin. Ils se prêtent particulièrement à ces circonstances où les objectifs sont différents selon la fonctionnalité désirée. En effet, il est souhaitable, par exemple, de plutôt valoriser la concordance sur les premiers caractères dans le cas de l'autocomplétion contre tous dans le cas d'un libellé complet.

Les données du COG ont été indexées dans le moteur de recherche directement à partir du graphe RDF du COG. Un raccordement automatique entre le système de codage (le moteur de recherche) et la base de vérité (le COG) existe et assure une fraîcheur dans les données mobilisées par le système de codage. Néanmoins, l'index, produit exclusivement à partir du COG, peut être perçu comme insuffisant. En pratique, on dispose parfois de codes postaux en plus du libellé de la commune sur lequel il serait possible de s'appuyer. Par ailleurs, on trouve régulièrement des libellés de lieux-dits absents du COG à la place d'un vrai nom de commune dans la source à coder. L'index a donc été enrichi en ajoutant des attributs provenant d'autres sources directement en interrogeant leurs services web (Laposte pour les codes postaux et OpenStreetMap via l'API

Overpass pour les lieux-dits ou les communes frontalières).

Pour coder en s'appuyant sur l'index précédemment décrit, une méthode de recherche par relâchement de contraintes a été mise en oeuvre. Une requête d'abord très stricte sur le libellé et le code département ou postal est effectuée. Si l'entité géographique n'a pas pu être identifiée, plusieurs passes avec des critères de similitude entre l'input à coder et les documents contenu dans l'index de plus en plus souples sont exécutées. Cette approche donne des performances (taux de libellés codés automatiquement et taux de bien codés) légèrement meilleures que Sicore Commune sur une base de données de communes annotée à la main extraite des déclarations sociales nominatives.

# Bibliographie

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval, Cambridge University Press, 2009.