Independent Address Identification Search Engine for National Statistical Institute Using ElasticSearch

Raya Berova
Insee
raya.berova@insee.fr

2024-10-10

Introduction

In order for survey interviewers to reach individuals, it is essential to accurately identify and geolocate their addresses. Address data, used in a wide range of statistical processes—from census to surveys—is often difficult to process due to inconsistencies, variations in input, and the volume of records.

In matters of address search, many rely on established services like Google Maps or OpenStreetMap. However, these platforms often pose limitations in terms of data control and reliability. Creating a custom address identification search engine provides complete control over the data, addressing concerns about data source transparency and monthly data updates.

A solution employing ElasticSearch (ES), a powerful software used to create and configure search engines, is here proposed to build an independent process for identifying address data for the National Statistical Institute (NSI). Moreover, ES enables text-based address search and supports the storage of geometric objects, considering the spatial aspect of addresses. This approach optimizes both processing time and accuracy by employing a two-step strategy: an initial strict search for precise address identification, followed by a flexible matching phase for addresses not identified in the first step, which accounts for spelling errors and variations in the input.

Bibliography