IDENTIFICATION FLOUE : REPÉRER DANS SIRENE L'EMPLOYEUR DÉCLARÉ AU RECENSEMENT. L'APPORT DU HACKATHON « LES CHAMPS DE SIRENE »

Yves-Laurent BÉNICHOU (*), Frédéric COMTE (**), Julie DJIRIGUIAN (*), Benjamin SAKAROVITCH (*), Éric SIGAUD (**)

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

info-hackahton@insee.fr

Mots-clés : identification, recensement, Sirene, hackathon, record linkage

Résumé

Rapprocher un élément déclaré lors d'une enquête dans une table de référence présente de nombreuses applications depuis les méthodes de record linkage jusqu'à celles de codification automatique. Ainsi les personnes recensées sont invitées à indiquer où elles travaillent en fournissant 3 informations sur leur établissement employeur : une raison sociale, une adresse et une activité. Retrouver à partir de ces trois champs l'établissement correspondant dans le répertoire Sirene est un processus coûteux. En effet, dans seulement moins de la moitié des cas, l'outil automatique actuel (MCA) parvient à l'identification. Pour tous les autres cas, des gestionnaires reprennent manuellement ce travail de recherche dans Sirene.

C'est dans l'objectif d'augmenter la part d'identification automatique qu'un hackathon a été organisé à l'INSEE en janvier 2018 rassemblant des membres de l'Institut, des services statistiques ministériels (SSM) et de différents partenaires. Plus de soixante personnes ont donc pendant deux jours exploré de nombreuses pistes pour identifier automatiquement l'employeur déclaré. L'objectif de cette communication sera de présenter rapidement le principe de l'événement et des différents outils utilisés et d'offrir une synthèse des pistes couvertes par les participants.

À partir de ces travaux plusieurs étapes se distinguent dans le processus d'identification. Au-delà du nettoyage qui peut être spécifique en fonction des secteurs d'activité, une première étape consiste à enrichir les données. Pour ce faire des méthodes de webscraping permettent d'utiliser l'enrichissement fait sur le répertoire Sirene par des sites tels que societe.com ou de bénéficier de la souplesse de moteurs de recherche comme celui des pagesjaunes.fr ou Qwant. De plus, on ajoute aux données, à partir des champs d'adressage, des attributs géographiques à travers des outils de géolocalisation.

Une seconde étape consiste à effectuer des requêtes dans le répertoire Sirene à partir des données nettoyées et enrichies. Des moteurs de recherche reposant sur l'indexation permettent cela, comme SolR actuellement utilisé par la nouvelle API Sirene ou ElasticSearch, ceux-ci présentent l'avantage d'être finement paramétrables. Des moteurs alternatifs fulltext ont également pu être testés.

Ensuite une étape de scoring vise à attribuer un niveau de confiance dans l'établissement trouvé. Pour cela différents critères sont combinés à la fois en terme de proximité des termes employés ou de distance géographique par exemple.

En terme d'évaluation des résultats, on utilise une opération Recap Qualité, qui a fait passer un double codage manuel et un arbitrage pour un échantillon de bulletins individuels dans le but

justement de mesurer la qualité du codage automatique et de la reprise manuelle, ce qui permet de comparer les performances des différents prototypes proposés par les participants.