Hackathon
« Les Champs de Sirene »

Alors je cherche et je trouverai
ce Siret qui me tente tant

Journées de préparation

Les lundi 27 novembre et 4 décembre

Gaël de Peretti (Département des méthodes statistiques)

Direct dans le vif du sujet

cas	Raison sociale collectée	Adresse collectée	Établissement repris	Activité déclarée
1	MAISON DE RETRAITE	59375 MARCHIENNES	MAISON DE RETRAITE EMILE DUBOIS 2 RUE D ORCHIES 59 MARCHIENNES	87 MAISON DE RETRAITE
2	IME	59192 EMERCHICOURT	PAPILLONS BLANCS ARRONDISSEMENT DE DOUAI 7 RUE DE L EGALITE 59 EMERCHICOURT	87 IME
3	DEMECO	62249 COURCELLES-LES-LENS	SARL MERY - ARTDEM TRANSPORTS LOGISTIQUE MERY RUE DE L ABBE POPIELUSZKO 62 COURCELLES LES LENS	49 DEMENAGEMENTS
4	DENISART JEAN PHILIPPE	59387 MARQUETTE-EN-OSTREVANT	GENERATION ESPACE VERT - G.E.V 2 RUE PASTEUR ENTREE 5 59 MARQUETTE EN OSTREVANT	81 ESPACES VERTS
5	TEREOS	RUE D ERES 59122 CAMBRAI	TEREOS FRANCE RUE D'ERRE 59 ESCAUDOEUVRES	10 AGROALIMENTAIRE
6	YMERIS	RTE WAHAGNIES 59462 PHALEMPIN	IMERYS TC - IMERYS TOITURE ROUTE DE WAHAGNIES 59 PHALEMPIN	41 BATIMENT

Comment ça se passe aujourd’hui, enfin hier

- Les moteurs d’identification utilisés par l’enquête Emploi et les EAR ne trouvent rien
- l’API Sirene trouve 1 et 2
- Glouglou trouve tout le monde

Les échos glouglou pour identifier

- Cas 3 : Société SARL, Mery Artdem, agence de déménagement à Courcelles les Lens, et l'adresse du site commence par www.demeco.fr/[…]
- Cas 4 : erreur orthographe sur nom (Denizart), chef entreprise Génération Espace Vert qui fait écho à l'APE déclarée
- Cas 5 : erreur sur la ville (Cambrai vs Escaudoeuvres), erreur sur le libelle de la rue (Eres vs Erre), mais nom bon et APE bonne (agroalimentaire pour Fabrication de sucre)
- Cas 6 : faute d'orthographe sur le nom, APE incorrecte (mais pas complètement éloignée, Construction de bâtiment vs Fabrication d'autres produits minéraux non métalliques, sauf dans la NAF, industrie manufacturière vs Construction :o) mais bonne adresse

Quelles erreurs possibles

- orthographe sur la raison sociale ou la commune ou l’adresse ;
- raison sociale différente de celle inscrite dans le répertoire ;
- erreur sur l’adresse ou la commune ;
- erreur sur le libellé d’activité ;
- Etc.

Quelles solutions envisager ?

- Des échos corrects via un moteur de recherche souvent dès la première page donc codage manuel simple
- Web-scraping + matching + testing
- Utilisation de corrélation spatiale dans le fichier d'entrée (mon voisin travaille dans la même boîte, mais lui il a tout bien saisi)
- Calcul de distance entre libellé de raison sociale et/ou d'adresse, commune et activité « proches » pour donner des échos
- Utiliser tous les champs de Sirene (ie les autres variables)
- Etc.