

Évolutions BDD ENTREPRISES

Procédures de Mise à Jour de la BDD ENTREPRISES

ENTREPRISES V5

Ce document reprend intégralement le texte de la V3 de Jean-Marc. Les blocs précédés de ♦ sont des ajouts de Claude Code. Document entièrement éditable : modifiez, supprimez ou complétez librement avant de le redéposer. Accents français corrigés le 16 septembre 2026 (bug d'encodage sur les versions précédentes).

1. Interrogations de JM concernant les sièges et établissements secondaires

De ce que j'en ai compris, la récupération des entreprises automatisée comprend la récupération dans un département des établissements secondaires du moment que le siège, même situé dans un autre département, y possède un établissement secondaire.

Actuellement, l'adresse affichée pour cet établissement secondaire est celle du siège, ce qui peut porter à confusion et faire craindre à un doublon.

Toutefois, il apparaît que l'adresse de l'établissement secondaire est stockée dans les données JSON pas encore totalement exploitées (mais fort pertinemment conservées !).

J'en déduis donc que lorsque le traitement de correction sera apporté à l'adresse de l'établissement secondaire, celui-ci pourra apparaître en bonne place dans les recherches d'entreprise du département et c'est bien ce qui est important.

Toutefois, dans la mesure où le numéro SIREN sera le même pour le siège et ses établissements secondaires, il sera opportun de stocker aussi le SIRET de chaque entreprise qui est d'ailleurs la donnée très souvent utilisée pour identifier une entreprise.

Concernant la notion de doublons, il y a deux possibilités :

- Si au niveau du département du siège, sont stockées aussi les entreprises filiales, alors, il y a effectivement doublons. Dans le cas qui nous occupe, la BDD va servir à effectuer de la prospection qui va dans la majeure partie des cas intégrer une dimension géographique. Dans ce cas, il est inutile de garder les filiales, sauf si celles qui sont dans le même département que le siège ! Une sélection d'entreprises sur la base d'un autre critère que géographique, par type d'entreprise (ETI ou GE) par exemple, devra pouvoir proposer les établissements secondaires. Pour cela, la recherche pourra être organisée en se basant sur le numéro SIREN. Le fonctionnement des mécanismes de sélection des entreprises n'est pas le sujet de cette feuille route, il sera abordé plus tard.
- Si au niveau du département du siège ne sont pas stockées les établissements secondaires (ce que je n'ai pas vérifié, il faudrait le faire), alors il n'y a pas de doublons dans la BDD et il faut garder en l'état les données.

Dans tous les cas, il faudra stocker pour chaque entreprise, qu'elle soit siège ou établissement secondaire les informations de son adresse dans le département où elle est située.

◆ RÉPONSE CLAUDE CODE :

Vérifié sur données réelles (35 vs 44) : 7802 SIREN sont déjà présents comme ligne physique à la fois dans Entreprises35Phase1 ET Entreprises44Phase1 (exemple : « G2 (THE BRITISH CASHMERE) », même adresse de siège affichée dans les deux). C'est le 1er scénario qui se confirme : il y a bien des doublons aujourd'hui. Important : cette preuve repose sur l'EXISTENCE de la ligne dans chaque table (chaque table n'est peuplée qu'en interrogeant les communes de son propre département), pas sur le contenu du champ adresse — qui, lui, est encore erroné pour les filiales (affiche l'adresse du siège).

Sur la nuance « garder les filiales sauf celles du même département que le siège » : à confirmer avec Jean-Marc si l'intention est plutôt l'inverse en pratique — garder une filiale dans le département où elle est physiquement située (utile en recherche géographique locale), sans la dupliquer dans le département du siège si celui-ci est déjà représenté par sa propre ligne.

Piste technique pour identifier sans ambiguïté siège vs établissement secondaire : le JSON contient déjà un booléen explicite, `est_siege` (visible dans `matching_etablisements[]`) — directement exploitable, pas besoin de le déduire d'une comparaison de champs.

2. Tables mutualisées à créer

Ces tables devront être créées dans la BDD ENTREPRISES et mises à jour si besoin par les procédures de mises à jour quotidiennes.

Il s'agit de :

- La table des activités principales ;
- La table des natures juridiques ;
- La table des tranches des effectifs salariés ;
- La table des dirigeants ;
- Tout autre table nécessaire au décryptage des données stockées.

Mode de fonctionnement en réalisation :

- Me proposer pour chacune de ces tables une structure
- Attendre que je valide avant de les remplir de valeurs

3. Questions préalables et remarques concernant les tables Phase1

- À quoi correspond `ENT_SectionActivite` → une section NAF. À conserver !
- A-t-on récupéré des entreprises avec plusieurs noms de dirigeant ? → Oui, alors créer une table des dirigeants (voir point 2)
- Il faut identifier les entreprises sièges et établissement secondaire

- Interrogations à suivre concernant le champ JSON ;
- Quid du caractère_employeur ?
- Quid du statut diffusion ?
- Quid de compléments{} ?
- Quid de matching établissements ? → important à voir ensemble
- Quid de epci ? → à conserver
- Quid de liste_rge → à conserver
- Quid de liste idcc → à conserver
- Quid de liste finess → à conserver
- Quid de liste uai → à conserver

De manière générale, il est utile de conserver toute information discriminante qui pourrait servir demain à moi ou un autre utilisateur d'ASPIC

Mode de fonctionnement en réalisation :

- Me proposer pour chacun de ces champs une structure
- Attendre que je valide avant de lancer les traitements

◆ RÉPONSE CLAUDE CODE :

caractère_employeur : valeurs « O »/« N », au niveau établissement (a des salariés ou non) — vérifié dans matching_etablissements[[]].

statut_diffusion : « O » dans l'exemple vérifié (diffusion publique complète), lié au mécanisme [NON-DIFFUSIBLE] déjà géré ailleurs dans le script.

compléments{} : que des booléens qualitatifs (bio, ESS, RGE, Qualiopi, association...) — déjà inspecté en détail, faible intérêt a priori pour la prospection.

matching_établissements : en attente d'être étudié ensemble comme demandé — c'est là où se trouve, entre autres, le champ est_siege et l'adresse propre de chaque établissement matché (voir point 1).

4. Modifications à apporter aux tables Phase1

Ces modifications porteront sur des champs :

- À modifier de nom ;
- À ajouter ;
- À supprimer.

Opérations à réaliser :

- ENT_AdresseLig1 constitué comme suit numéro_voie+", "+type_voie+" "+libellé voie. *** Attention ** je tiens à ce que la ligne 1 soit bien écrite « 15, rue des Châtaignes ». Cette petite différence a une utilité pour moi !
- ENT_AdresseLig2 = complément_adresse
- Renommer Ent_CodePostal en ENT_AdresseCP

- Renommer ENT_Commune en ENT_AdresseCommune
- Renommer ENT_CodeCommune en ENT_AdresseCodeCommune
- Ajouter ENT_AdresseCedex et y mettre la partie du champ JSON correspondant
- Renommer ENT_Departement en ENT_AdresseDepartement
- Renommer ENT_Region en ENT_AdresseRegion
- Ne pas conserver les détails de l'adresse du siège si établissement secondaire
- Ajouter le champ ENT_SiegeSIREN avec le numéro SIREN du siège
- Ajouter le champ ENT_SiegeDept avec le numéro de département du siège
- Ajouter le champ ENT_SIRET avec le numéro SIRET de l'entreprise
- Récupérer tous les champs idcc, finess, uai, epci, ess, rge, qualiopi, association...
- Supprimer le champ ENT_AdresseSiege
- Créer un champ ENT_CapitalSocial
- ENT_FinanceAnnee, ENT_FinanceCa, ENT_FinanceResultatNet doivent être reportés dans Phase3 et après vérification suppression de ces champs
- ENT_CanalProspectio, ENT_Phase2, ENT_ScoreCiblage, ENT_StatutQualification, ENT_Notes à supprimer
- Créer un champ ENT_NumeroTVA
- À récupérer dans le JSON (si ce n'est pas déjà le cas) toutes les informations pouvant être discriminantes (voir point 3)

Mode de fonctionnement en réalisation :

- Me proposer pour chacun de ces champs une structure
- Attendre que je valide avant de lancer les traitements

◆ RÉPONSE CLAUDE CODE :

nombre_etablisements_ouverts et tva existent déjà dans le JSON stocké, jamais mappés en colonne typée — pas besoin d'appel API supplémentaire, juste étendre map_entreprise().

Pour « identifier les entreprises sièges et établissement secondaire » (point 3) : voir la piste est_siege notée au point 1, plus simple qu'une colonne enum à maintenir en synchro.

5. Création de la table Phase5

Cette table reprendra la liste de tous les dirigeants connus de l'entreprise (si possible en différenciant siège des établissements secondaires)

Mode de fonctionnement en réalisation :

- Me proposer pour chacun de ces champs une structure
- Attendre que je valide avant de lancer les traitements

◆ RÉPONSE CLAUDE CODE :

Point d'attention avant de concevoir la structure : dans le JSON, les dirigeants sont rattachés à l'entreprise (niveau racine du document), pas à un établissement précis — la donnée source ne permet donc

probablement pas de savoir « ce dirigeant est associé à cette filiale plutôt qu'au siège ». Le SIREN suffira à relier un dirigeant à l'entreprise, mais la différenciation par établissement risque de ne pas être réalisable avec cette source. À confirmer en creusant le JSON si le sujet est important.

6. Déplacement des champs JSON

- Supprimer la table Phase2
- Créer la table Phase4
- En ajoutant le champ ENT_SIREN et ENT_ReponseBruteJson
- Déplacer les champs ENT_ReponseBruteJson de Phase1 vers Phase4
- Supprimer après vérification du bon déplacement du champ ENT_ReponseBruteJson de Phase1

Mode de fonctionnement en réalisation :

- Attendre que je valide avant de lancer les traitements
- Effectuer ce traitement pour tous les Phase1

7. Modification de la table ENT_Departements

- Ajouter le champ DEP_NbCommunes
- Ajouter les champs DEP_EI, DEP_TPE, DEP_PME, DEP_ETI, DEP_GE

Mode de fonctionnement en réalisation :

- Attendre que je valide avant de lancer les traitements
- Effectuer ce traitement pour tous les départements
- Prévoir d'intégrer la mise à jour dans les traitements de mise à jour

◆ RÉPONSE CLAUDE CODE :

DEP_ETS (nombre d'établissements secondaires) présent en V1/V2 n'apparaît plus ici — volontaire, ou oubli pendant la réécriture ? Vu que tout le point 1 tourne autour de cette distinction, ça semble une info utile à garder. Même remarque au point 8 (REG_ETS).

8. Création de la table des régions ENT_Regions

- Créer le champ REG_Code
- Créer le champ REG_NbDepartements
- Créer le champ REG_NbCommunes
- Créer le champ REG_NbEntreprises
- Créer les champs REG_EI, REG_TPE, REG_PME, REG_ETI, REG_GE

Mode de fonctionnement en réalisation :

- Attendre que je valide avant de lancer les traitements

- Effectuer ce traitement pour tous les départements
- Prévoir d'intégrer la mise à jour dans les traitements de mise à jour

9. À propos de la mise à jour des tables

Le site <https://nouvelles-entreprises.com/statistiques> communique que cette page en disant : « Chaque nuit : l'ingestion récupère les nouvelles immatriculations SIRENE, retire les entreprises qui ont fermé, puis recalcule tous les agrégats affichés sur ces pages. »

Il faut absolument que l'on comprenne comment ils ont pour limiter leurs traitements aux seules mises à jour (et auquel cas voir si l'on a trouvé toutes les pistes pour le faire puisqu'il était annoncé que cela était faisable) ou bien si cela est juste une affirmation marketing !

De plus, je note que le business modèle de cette entreprise repose en grande partie sur ce que je m'appête à offrir dans ASPIC. La lecture de toute leur communication est très instructive. Ils font du fait qu'ils ne connaissent pas l'adresse mail un argument ! C'est fort ! Il me faut absolument rebondir là-dessus et mettre en avant que pour ASPIC c'est gratuit. Toutefois, j'envisage de positionner un quota max par mois pour les extractions, avec option payante au-delà en tablant sur leurs tarifs !

À noter : j'ai essayé de voir quel est leur chiffre d'affaires mais la société éditrice de ce site est Ysae SLU, immatriculée en Andorre. On a affaire à des malins !

◆ RÉPONSE CLAUDE CODE :

Très probablement pas du marketing : l'API Sirene officielle expose nativement un filtre par date de dernière modification (`dateDernierTraitementUniteLegale`), le même mécanisme déjà identifié pour notre propre future synchro incrémentale (réflexions stratégiques, section 25). Rien de mystère à percer, capacité standard de l'API.

Limite à connaître cependant : ce filtre ne couvre que les champs sourcés Sirene (identité, effectif...), pas le capital social qui vient du RNE (INPI), registre à part avec son propre canal (`/companies/diff` ou le flux FTP). Confirmé par Jean-Marc le 16/09 : trou jugé acceptable vu la rareté des changements de capital social.

10. Faire le point sur les procédures de mise à jour de toutes ces tables

Pour cela nous devons :

- Revoir les procédures de mise à jour de toutes les tables
- Organiser leur traitement : la nuit, étendue du traitement...
- Étudier la mise à jour du capital social de chaque entreprise avec fichiers PDF

11. Création d'une table mutualisée dans la BDD ENTREPRISES

Cette table aura pour objectif de servir de recensement des temps et coûts de traitement des opérations automatiques réalisées par ASPIC pour qualifier des entreprises et ce pour l'ensemble des utilisateurs, en faisant en sorte de pouvoir faire ces analyses par type d'entreprise.

En effet, je suis presque certain que ce sera plus difficile de qualifier des EI que des PME ! On ne va pas aller jusqu'à identifier les différences de ces informations par région, mais je veux que ASPIC soit capable d'annoncer, par l'intermédiaire de Viktor que « pour qualifier cette sélection de xx entreprises, j'estime que je vais mettre xx heures, que je vais réussir à en qualifier xx (yy %) et que cela va coûter xx € sur votre crédit de xx €. Confirmez vous le lancement de la qualification ? »

Pour conserver le maximum d'information, je souhaite que les données enregistrées soient le plus détaillées possible : par quel utilisateur ASPIC, description du lot qualifié, traitement automatique utilisé (recherche web, recherche GoogleMaps) répartition par type d'entreprise (EI, TPE...) avec les nombres des types concernés, temps de traitement global et coût en appels API.

On lancera ensuite plusieurs tests de qualification pour commencer à alimenter cette table, mais cela fera l'objet d'un futur chantier !

Mode de fonctionnement en réalisation :

- Me proposer pour chacun de ces champs une structure
- Le bon enregistrement des données sera testé au fur et à mesure des essais de prospection !