**Brief projet**

**L'IA au service de la sécurité alimentaire**

**\*\*\***

**Groupe 1**

**\*\*\***

**Retour sur MCD**

**\*\*\***

Jean-Michel DIDIER

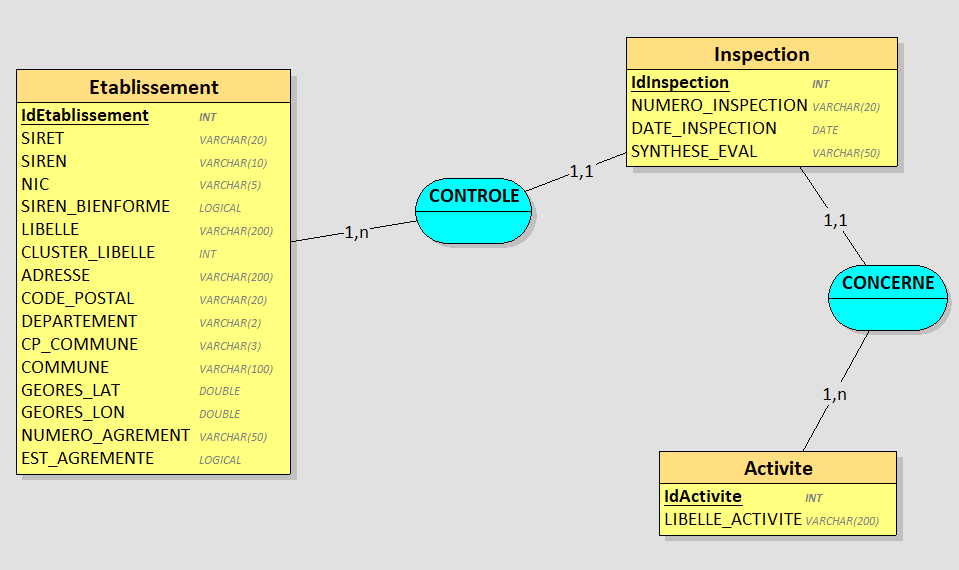
Version du 11/01/2022

# Ressources

## Le format du fichier csv en entrée

|  |
| --- |
| APP\_Libelle\_etablissement |
| SIRET |
| Adresse\_2\_UA |
| Code\_postal |
| Libelle\_commune |
| Numero\_inspection |
| Date\_inspection |
| APP\_Libelle\_activite\_etablissement |
| Synthese\_eval\_sanit |
| Agrement |
| geores |
| filtre |
| ods\_type\_activite |

## Le MCD proposé par le groupe 1



# Le modèle de données proposé

## Justification du stockage en base de données et du modèle

Vous disposez d'un jeu de données sous la forme d'un fichier csv, vous devrez les analyser d'une part, et d'autre part les transformer pour les exploiter via un ou plusieurs moteurs d'IA.

### Pourquoi proposez-vous une base de données ?

Nous estimons que l’utilisation d’une base de données permettra un traitement plus efficace des données ainsi qu’une restitution plus aisée à destination de l’IA. Nous pourrons également collecter les réponses du modèle, et ajouter une interface en front et en back qui pourra s’appuyer sur cette base.

### Quelle est votre intention en les stockant en base ?

Notre intention est d’avoir un support stable et efficace pour nous permettre d’effectuer les transformations que nous souhaitons sur les données. Grâce à la base de données nous pourrons avoir une grande flexibilité sur les choix de stockage, ainsi qu’une dimension relationnelle indispensable pour ce genre de projet. Nous avons en effet repéré des entités différentes au sein du même fichier CSV.  
De plus, le stockage en base nous assure que nous conserverons l’intégrité du fichier CSV.

### Pourquoi voulez-vous éclater un format unique et simple en un modèle relationnel ?

Nous le faisons car, bien que ce fichier se base sur les inspections faites dans des établissements alimentaires, nous savons que cela peut concerner plusieurs fois un même établissement, et plusieurs fois des mêmes activités. Ainsi nous pouvons extraire ces entités afin d’en avoir un traitement et une visualisation plus aisée par la suite.

### Quelle(s) plus-value(s) attendez-vous de cette transformation ?

Nous souhaitons procéder à cette transformation pour mieux organiser nos données. Nous pourrons ainsi mieux les identifier pour les fournir à l’IA ou un utilisateur, qui auront des données propres et pertinentes. Aussi, cette transformation nous permettra de les exploiter au sein d’une application Front et Back basée sur Flask, qui présentera les résultats du modèle IA en front, et fournira une interface d’administration pour le back.

### Quels traitements appliquerez-vous sur ces données une fois en base ?

Nous souhaitons dédoublonner certaines données et regrouper les activités similaires même si elles ne sont pas clairement identifiées comme telle dans le fichier CSV initial.

Par ailleurs, nous effectuerons un nettoyage et un formatage adéquat des données.

### En quoi le modèle proposé vous aidera-t-il ?

Ce modèle conceptuel de données nous permet de visualiser les différentes relations entre les entités, ainsi que les cardinalités correspondantes, puis il nous permettra de produire le modèle physique de données, et ainsi de créer les tables dans notre système de gestion de base de données.

## Questions et remarques sur le modèle

*(il s’agit davantage de demandes d’explication que de critiques)*

### Questions sur votre démarche Tête avec engrenages

* Fonctionnellement, pourquoi avoir ajouté des colonnes contenant des informations alors que ces informations sont déjà présentes (indirectement) dans les colonnes existantes ? par exemple le département dans le code postal.

Fonctionnellement, l’ajout de ces colonnes n’est pas justifié d’un point de vue modélisation pure.

* Techniquement, pourquoi créer des colonnes dédiées à partir d’une colonne lorsque l’information peut être aisément déduite par requête ad-hoc ? Utilisation d’espace de stockage versus utilisation de temps de calcul ?

Techniquement, nous pensons qu’en précalculant certaines données notre application sera bien plus performante.

* Le modèle de données permet de décrire votre base de données, mais il est important de « documenter » comment vous l’avez conçue : par exemple, pour chaque nouvelle colonne, où avez-vous « documenté » comment elle est calculée ou déduite d’autre(s) donnée(s) ?

Afin d’aboutir au MCD que nous avons fourni, nous avons analyser les données du fichier CSV en créant un dictionnaire de données.

*(« documenter » : selon le manifeste agile, je ne parle pas ici d’une « documentation exhaustive » sous forme d’un document externe au MCD qui pourrait être difficile à maintenir cohérente avec le MCD au fil du projet, mais d’une façon simple et opérationnelle d’ajouter ces indications dans l’outil utilisé pour construire ce MCD, afin de faciliter les futures interventions sur la base de données)*

### Etablissement

* Extraire le NIC ou le SIREN à partir du SIRET est une bonne idée fonctionnelle, mais pourquoi NIC et SIREN ?

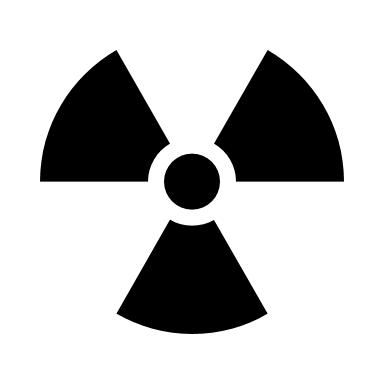
Nous avons imaginé qu’il était important de récupérer le SIREN indépendamment car il indique l’appartenance d’un établissement à une chaine, ce qui permettra une meilleure granularité des informations données à l’IA.

* Extraire le département du code postal est de même une bonne idée, mais à quoi sert la colonne CP\_COMMUNE ?

Cette colonne est en effet inutile, nous venons de nous en apercevoir. Par contre, le numéro de département reste pertinent.

* Pourquoi séparer longitude et latitude ?

En effet, nous réalisons que séparer ces deux données n’est pas utile car elles sont systématiquement utilisées conjointement.

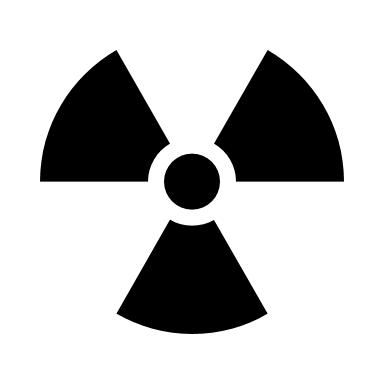
* Attention à l’appauvrissement de la sémantique, la colonne LIBELLE doit a priori être le libellé d’établissement, il est préférable de conserver cette information dans le nom de la colonne.

Nous tenons compte de cette remarque, et renommerons les colonnes trop vagues pour coller mieux aux données représentées.

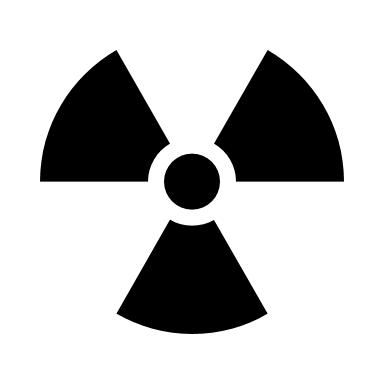
### Inspection

* R.A.S.

### Activité

* 3 colonnes du fichier brut contiennent des informations sur l’activité. Quelle colonne du fichier brut avez-vous choisie pour alimenter cette table, et pourquoi ?

Nous pensons rassembler les données des 3 colonnes, en évitant les doublons, pour alimenter la table activité de manière cohérente.

* Si une seule colonne retenue, pourquoi avoir éliminé les 2 autres colonnes ?
* Pourquoi une table spécifique pour cette information, et pas, par exemple, de table pour le code postal et la commune ?

Nous venons, de part cette remarque, de constater que nous aurons en effet tout intérêt à créer une table des communes afin d’éviter de perdre des informations de corrélation. (Car pour un même code postal, il y a potentiellement plusieurs communes, et de plus, les communes ne sont pas toujours orthographiées de la même façon).