|  |
| --- |
| Master 1 DSC - Groupe 2 - Interopérabilité |
| Extraction des sources |
| 21/02/2020 |

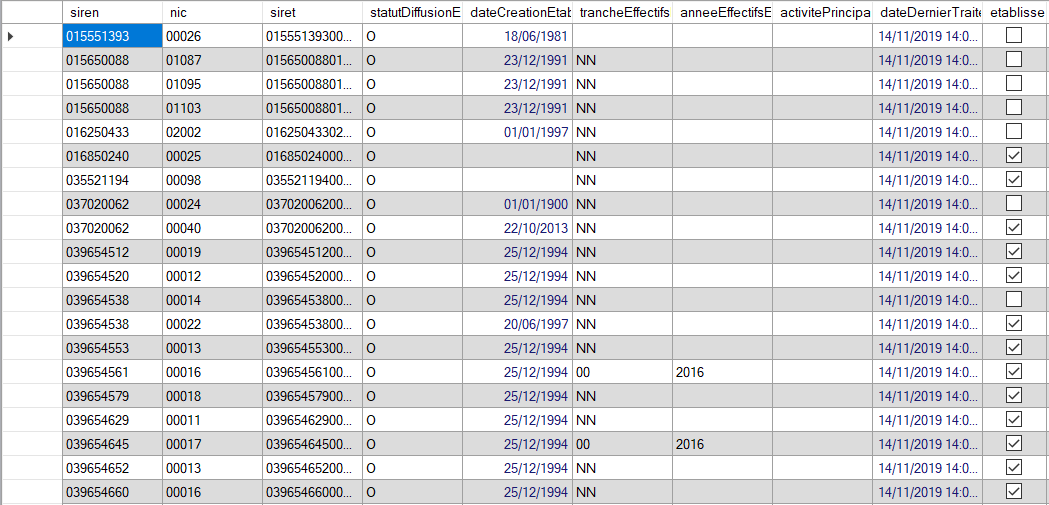
|  |
| --- |
| ROMDAN Elias  TROTTA Nicolas |

1. Introduction

Nous avons déjà abordé la partie du principe d’extraction des données dans le document intitulé « [2020-02-07] - Groupe 02 - Rendu 02 - D4 - Définition et description des sources ». Ce document sera un complément au dernièrement cité.

2. Extraction des données de type CSV

Voici un rappel de la structure de nos sources CSV :



L’idée d’extraction dans ce cas est simple vu que les données sont déjà organisées dans un tableau. La conversion d’un CSV vers une liste Java conservera la forme de ce tableau. Il suffit par la suite de parcourir le tableau résultant et d’enregistrer le contenu des cases voulues dans des objets Java, puis de stocker ces objets dans la base de données.

3. Extraction des données de type HTML

Voici un rappel de la structure de nos sources HTML :



Sachant que les pages des sources HTML contiennent d’autres informations, mais pour notre projet, on s’intéressera uniquement à la partie du tableau avec les entreprises. L’idée d’extraction se base sur la détection de la ou les balises qui contiennent les informations sur les entreprises et de pointer sur eux. Par la suite il sera possible de récupérer le contenu de ces balises et de le mettre dans des objets Java qui seront ensuite stockés dans la base de données.

4. Extraction des données de type PDF

Voici un rappel de la structure de nos sources PDF :



Parmi les 3 sources, l’extraction des données PDF sera la plus compliquée, à la différence des sources CSV ou HTML où les données ont une structure qui nous permettra d’accéder aux informations voulues facilement. À l’extraction d’une source PDF, le contenu sera considéré comme étant du texte en succession. Une idée est de repérer le contenu qui se répète à chaque fois comme SIREN ou le format unique des dates et d’organiser l’extraction des données suivant ces éléments statiques.