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 :

	siren	nic	siret	statutDiffusionE	dateCreationEtat	trancheEffectifs	anneeEffectifsE	activitePrincipa	dateDernierTraite	etablisse
▸	015551393	00026	01555139300...	O	18/06/1981				14/11/2019 14:0...	☐
	015650088	01087	01565008801...	O	23/12/1991	NN			14/11/2019 14:0...	☐
	015650088	01095	01565008801...	O	23/12/1991	NN			14/11/2019 14:0...	☐
	015650088	01103	01565008801...	O	23/12/1991	NN			14/11/2019 14:0...	☐
	016250433	02002	01625043302...	O	01/01/1997	NN			14/11/2019 14:0...	☐
	016850240	00025	01685024000...	O		NN			14/11/2019 14:0...	☑
	035521194	00098	03552119400...	O		NN			14/11/2019 14:0...	☑
	037020062	00024	03702006200...	O	01/01/1900	NN			14/11/2019 14:0...	☐
	037020062	00040	03702006200...	O	22/10/2013	NN			14/11/2019 14:0...	☑
	039654512	00019	03965451200...	O	25/12/1994	NN			14/11/2019 14:0...	☑
	039654520	00012	03965452000...	O	25/12/1994	NN			14/11/2019 14:0...	☑
	039654538	00014	03965453800...	O	25/12/1994	NN			14/11/2019 14:0...	☐
	039654538	00022	03965453800...	O	20/06/1997	NN			14/11/2019 14:0...	☑
	039654553	00013	03965455300...	O	25/12/1994	NN			14/11/2019 14:0...	☑
	039654561	00016	03965456100...	O	25/12/1994	00	2016		14/11/2019 14:0...	☑
	039654579	00018	03965457900...	O	25/12/1994	NN			14/11/2019 14:0...	☑
	039654629	00011	03965462900...	O	25/12/1994	NN			14/11/2019 14:0...	☑
	039654645	00017	03965464500...	O	25/12/1994	00	2016		14/11/2019 14:0...	☑
	039654652	00013	03965465200...	O	25/12/1994	NN			14/11/2019 14:0...	☑
	039654660	00016	03965466000...	O	25/12/1994	NN			14/11/2019 14:0...	☑

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 :

Raison sociale	CP	Ville	C.A.
DISTRIBUTION CASINO FRANCE	42000	Saint-Ã‰tienne	8 998 000 000 €
FLOREAL	42000	SAINT ETIENNE	811 095 540 €
SNF SAS	42160	ANDREZIEUX BOUTHEON	750 433 864 €
NEXTER SYSTEMS	42300	Roanne	588 907 000 €
EASYDIS	42000	SAINT ETIENNE	569 716 426 €
LOCAM - LOCATION AUTOMOBILES ET MATERIEL	42000	Saint-Ã‰tienne	532 478 338 €
CASINO CARBURANTS	42000	Saint-Ã‰tienne	332 311 978 €
TRADIVAL	42300	ROANNE	328 174 468 €
HAULOTTE GROUP	42152	L'Horme	279 519 000 €
DELTAGRO EXPORT	42300	ROANNE	215 018 634 €
ATELIER FOREZIEN DU FRAIS	42350	La TalaudiÃ¨re	202 524 693 €
CENTRALE D'ACHATS KIDILIZ	42400	Saint-Chamond	181 992 320 €
BECKER INDUSTRIE SA	42600	Savigneux	178 315 798 €

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 :

Liste des 82 immatriculations d'entreprises

validées au Répertoire des Métiers entre le 16/03/2019 et le 31/03/2019

Identification de l'entreprise	Activité exercée	Adresse	Evènement(s)	Date d'effet
ABABOU Driss Né le 01/01/1967 à AIT HADDOU CHAIB (MAROC) SIREN : 495 143 521	ETANCHEITE ISOLATION	6-8 ALLEE HENRY PURCELL 42000 SAINT ETIENNE	Création PP individuelle ayant déjà exercé act. non salariée	21/03/2019
ABDECHAKOUR née BENHADDAD Katia, Djedjia Née le 13/06/1983 à SAINT ETIENNE SIREN : 848 154 704	RESTAURATION RAPIDE (PIZZAS...)	19 RUE PRAIRE 42000 SAINT ETIENNE	Entrée de champ RCS ou RM	04/03/2019
AGLM AUTRE SARL à associé unique SIREN : 848 609 400	POSE ET MONTAGE DE MENUISERIES EXTERIEURES EN BOIS, NOTAMMENT DE TERRASSES	437 RUE GEORGES CLEMENCEAU 42153 RIORGES	Création d'une société avec activité	01/03/2019
AIT LHO Karim Né le 27/07/1975 à MEKNES (MAROC) SIREN : 849 150 057	PLATRERIE PEINTURE	22 RUE THIMONNIER 42100 SAINT ETIENNE	Début d'activité non salariée	18/03/2019
AJT RESEAUX AUTRE SARL à associé unique SIREN : 849 033 311	DEPLOIEMENT ET MAINTENANCE DES RESEAUX DE TELECOMMUNICATION CUIVRE ET FIBRE OPTIQUE Travaux d'installation electrique dans tous locaux	6 CHEMIN DE LA RIVIERE 42111 LA VALLA	Création d'une société avec activité	11/03/2019
AKCAKOC Irfan Né le 01/05/1956 à NOHUTLU (TURQUIE) SIREN : 428 711 311	RETOUCHE DE VETEMENTS	55 RUE JEAN JAURES 42700 FIRMINY	Création PP individuelle ayant déjà exercé act. non salariée	01/04/2019
ARGOMANIZ Julien Né le 12/02/1982 à SAINT ETIENNE SIREN : 799 824 149	MACONNERIE GENERALE ET COFFREUR	644 ROUTE DE SAVIGNEUX 42450 SURY LE COMTAL	Entrée de champ RCS ou RM	14/03/2019
BARLET Olivier, Michel, Bernard Né le 25/09/1981 à SAINT ETIENNE SIREN : 849 329 438	ENTRETIEN DE VEHICULES DE COMPETITION ET ENGINS DE TRAVAUX PUBLICS ET AGRICOLES	34 AVENUE DE SAINT MARCELLIN 42160 BONSON	Début d'activité non salariée	25/02/2019
BBM 41015 LA RICAMARIE AUTRE SARL à associé unique SIREN : 849 491 063	RESTAURATION RAPIDE A EMPORTER	4 AVENUE JEAN MERMOZ 42160 ANDREZIEUX BOUTHEON	Création d'une société avec activité	14/03/2019
BBM 41016 NIORT AUTRE SARL à associé unique SIREN : 849 488 192	RESTAURATION RAPIDE A EMPORTER	4 AVENUE JEAN MERMOZ 42160 ANDREZIEUX BOUTHEON	Création d'une société avec activité	14/03/2019
BBM 41017 AUTRE SARL à associé unique SIREN : 849 355 623	RESTAURATION RAPIDE A EMPORTER	4 AVENUE JEAN MERMOZ 42160 ANDREZIEUX BOUTHEON	Création d'une société avec activité	14/03/2019

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.