Extraction automatique d'informations issues des comptes sociaux d'unités légales

Laura GAIMARD/ Adem KHAMALLAH

Insee

Mesurer pour comprendre

Contexte

- Élaboration des statistiques structurelles d'entreprises à partir de deux sources de données :
  - enquête (ESA, EAP)
  - administratives (liasses fiscales)

- Problème : sources incomplètes pour la production de données en entreprises profilées (EP)

- Autre source de données : comptes sociaux (documentation comptable et financière des sociétés) sous forme structurée ou d'image scannées

  → Utilisation manuelle par les gestionnaires

- Depuis fin 2019, mise à disposition gratuite par l'INPI (open data) de ces documents

Contexte

- Tableaux de structures différentes selon les comptes sociaux

- Deux étapes :

  → récupérer directement la page contenant l'information recherchée parmi le document

  → récupérer directement les données recherchées dans un fichier plat pour automatiser le processus de calcul

Extrait d'un compte social comportant l'information recherchée

7.23. Filiales et participations

	Capital	Réserves et report à nouveau avant affectation du résultat (1)	Quote-part du capital détenu en %	Valeur d'inventaire des titres détenus		Prêt et avances consentis par la société et non remboursés	Cautions et avals donnés par la société	Chiffre d'affaires du dernier exercice écoulé	Bénéfice net (ou perte) du dernier exercice clos	Dividendes encaissés par la société au cours de l'exercice
				Brute	Nette					
RENSEIGNEMENTS DETAILLES				**114,2**	**25,5**					**8,2**
I-A FILIALES				**114,2**	**25,5**	**0,0**	**0,0**			**8,2**
SE ALPES	5,4	14,6	100%	6,4	6,4	0,0	0,0	277,1	8,4	7,2
France Transfo	1,7	(24,0)	100%	37,9	0,0	0,0	0,0	83,8	(23,1)	0,0
MG Ales	5,5	9,7	100%	4,3	4,3	0,0	0,0	118,5	1,5	0,0
Sté d'Appareillage Electrique Gardy	1,9	(5,3)	100%	28,8	0,0	0,0	0,0	2,0	(0,0)	0,0
SETBT	4,0	(0,6)	100%	4,0	0,0	0,0	0,0	47,0	1,7	0,0
Schneider Electric Telecontrol (Sorhodel Bardin)	2,4	(14,9)	100%	10,0	0,0	0,0	0,0	12,1	(3,5)	0,0
Scanelec	0,8	0,8	100%	5,1	5,1	0,0	0,0	42,5	0,7	0,8
Behar Sécurité	0,0	1,1	100%	4,2	4,2	0,0	0,0	5,1	0,0	0,2
Electro Porcelaine	1,3	5,8	100%	5,1	5,1	0,0	0,0	0,0	0,0	0,0
SE Manufacturing Bourguebus	2,1	(1,1)	100%	8,5	0,5	0,0	0,0	12,5	0,4	0,0
I-B PARTICIPATION				**0,0**	**0,0**					**0,0**
RENSEIGNEMENTS GLOBAUX				**7,3**	**6,4**	**0,2**	**0,0**			**3,3**
Filiales non reprises (en I-A)				6,2	5,4	0,0				2,4
Participations non reprises (en I-B)				1,1	0,9	0,2				0,9
AUTRES TITRES IMMOBILISES				**9,0**	**9,0**					**1,1**
Autres titres immobilisés (III)				9,0	9,0					1,1
AUTRES IMMOBILISATIONS FINANCIERES				**0,0**	**0,0**					
TOTAL				**130,5**	**40,9**					**12,6**

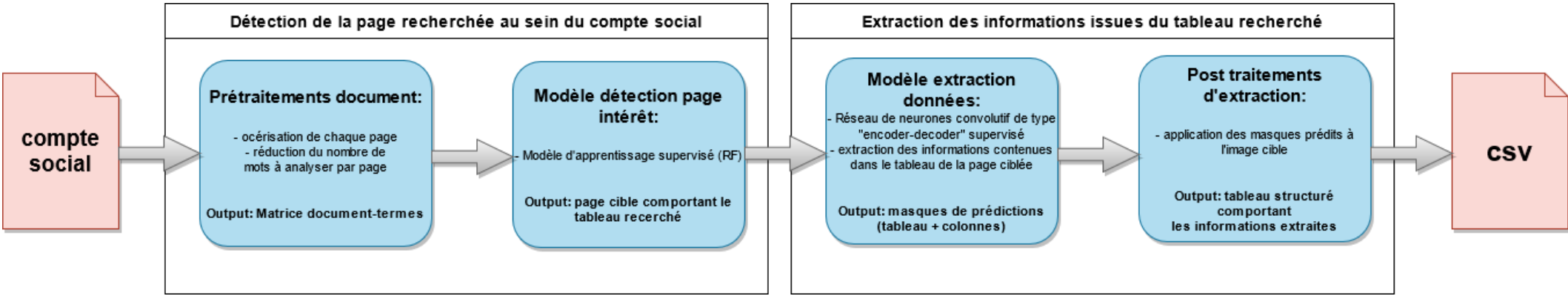
(1) Réserves + report à nouveau, y compris bénéfice net ou perte du dernier exercice clos + subventions d'investissements + provisions réglementées

Les renseignements fournis sont issus des comptes sociaux des sociétés. Le taux de cloture est utilisé pour la conversion des montants en devises

Les valeurs de titres détaillées sont celles supérieures à 1% du capital de SCHNEIDER ELECTRIC FRANCE, soit 3,7M€

nd : données non disponibles

Processus général d'extraction des données



Prétraitements pour les pages scannées (images)

Étape 1 : Reconnaissance optique de caractères (OCR) sur les pages ayant un faible nombre de mots (inférieur à un seuil s)
Moteur OCR open source *tesseract* utilisé

Étape 2 : Retraitement des données textuelles afin de faciliter l'analyse des mots :

- passage en minuscule
- suppression de caractères
- suppression des nombres et chiffres
- suppression des mots vides

Étape 3 : Suppression des mots peu représentés au sein de la matrice documents-termes (matrice creuse)

Constitution de la base de travail du 1er modèle

Étape 1 : Étiquetage à la main de 50 pages issues de comptes sociaux (présence du tableau d'intérêt ou non)

Étape 2 : Entraînement d'un modèle (Random Forest) à partir de ces 50 comptes sociaux pour prédire la page ayant la plus grande probabilité de présence du tableau d'intérêt

Étape 3 : Utilisation d'un script R de contrôle de la prédiction permettant d'obtenir une base de donnée de taille supérieure (environ 450 pages) sur des comptes sociaux tirés aléatoirement parmi des UL appartenant à des très grands groupes

Modèle de détection de la page d'intérêt parmi d'autres pages

Pour chaque page de la base d'apprentissage :

Variable à prédire : indicatrice de présence du tableau des filiales et participations, variable binaire (0 si la page ne contient pas le tableau filiales et participations, 1 sinon)

Variables prédictives : mots de la matrice document-termes

Indicateurs de qualité calculé sur échantillon test avec modèle RF:

- **taux de faux positif** (pages classées à tort ayant le tableau d'intérêt) → 0 %
- **taux de faux négatif** → 2,4 %
- **taux d'erreur** → 1,2 %

Modèle de détection de la page d'intérêt au sein du compte social

- **But : retrouver la page comportant le tableau dans un compte social donné**

  → **mais le tableau n'est pas toujours présent dans un pdf**

Principe de prédiction de la page cible dans un compte social :

→ Récupération de la page du compte social avec la probabilité de présence du tableau la plus grande

→ Application d'un seuil minimal de présence de ce tableau (0,9)

Indicateurs de qualité de prédiction de la page cible (avec modèle RF) :

→ **taux de vrai positif** (est-ce que le CS contient le tableau recherché) : 95 %

→ **taux de pages bien prédites** (est-ce que la page du CS contient le tableau recherché) : 81 %

Limites du 1er modèle

- Base de travail avec peu d'observation (490) → la compléter par un travail de labellisation des pages comportant ou non le tableau

- Certains tableaux des filiales et participations peuvent être vide (ie. Ne pas contenir d'observation d'UL « filles »)

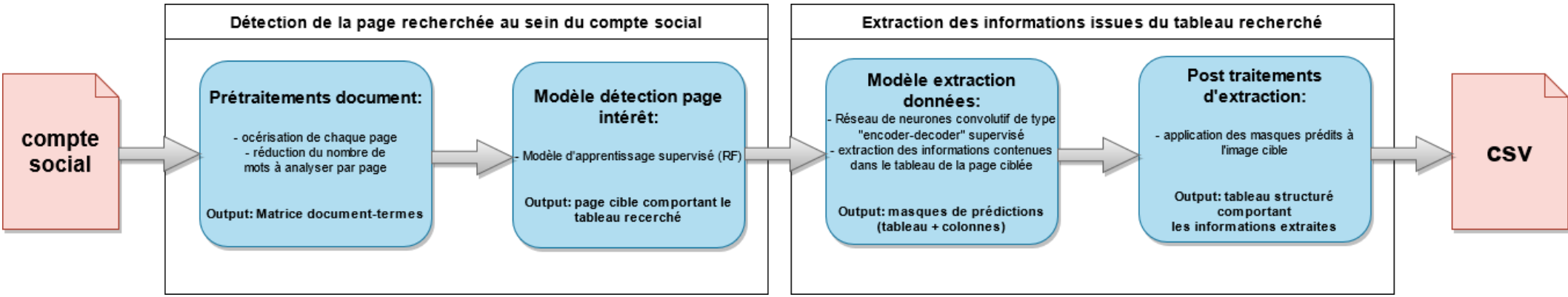
- Certains tableaux comportent des mots similaires à celui recherché, sans contenir l'ensemble des filiales et participations.

3.4 Modèle de détection de la page d'intérêt au sein du compte social

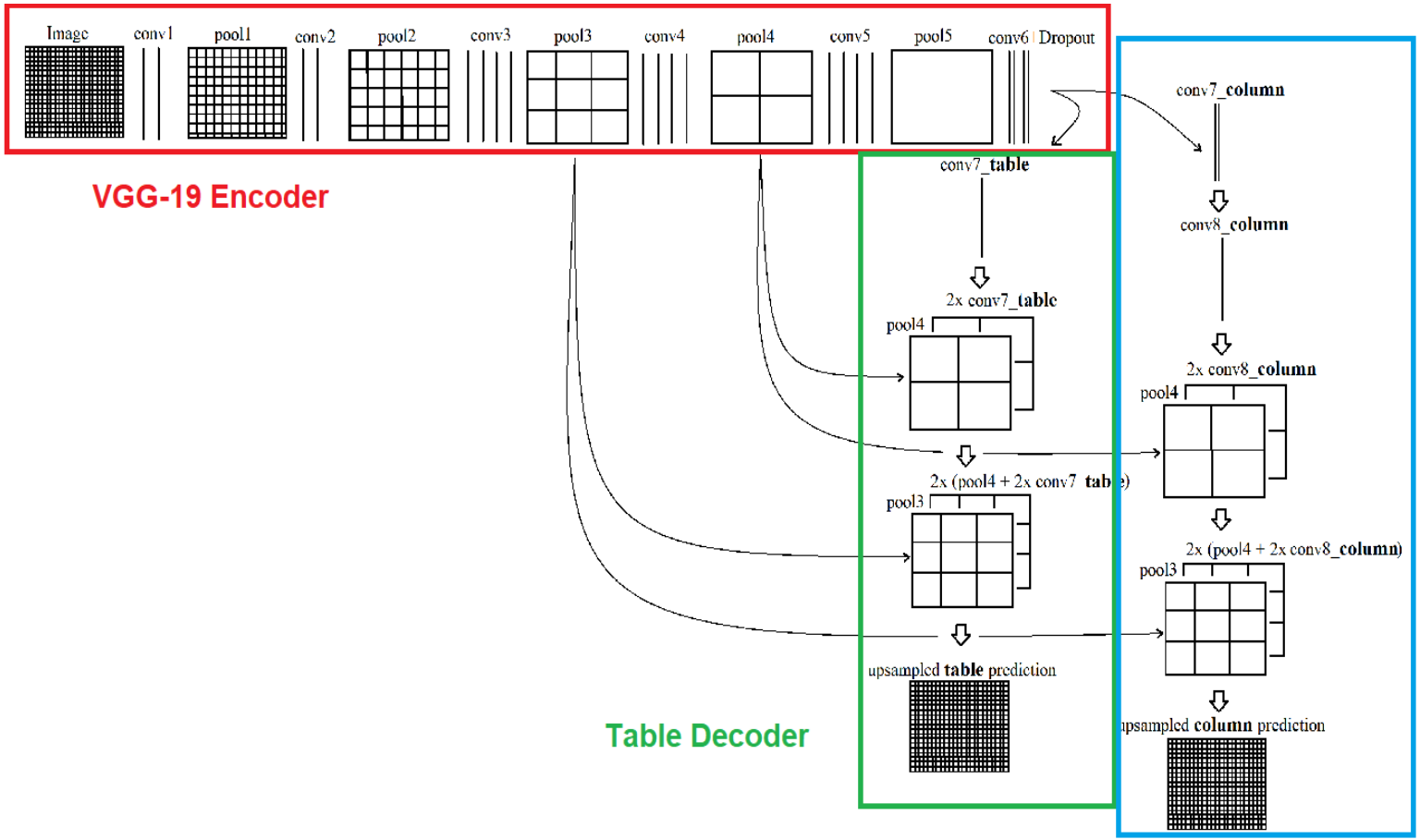
On applique ce modèle à 100 UL appartenant à des grands groupes tirées aléatoirement :

- Nombre de comptes sociaux renvoyés par la base RNCS : 74/100 (certains numéros SIREN ne renvoient pas de .pdf via l'API)
- Temps moyen de téléchargement : 6,3 secondes
- Temps moyen d'océrisation : 135,9 secondes sur AUS vs 35 secondes sur le SSPCloud
- Part moyenne des pages océrisées : 87%
- Nombre de tableaux trouvés par le modèle : 46/74 (dont 6 erreurs, sans prendre en compte que certains comptes sociaux ne possèdent pas ce tableau)

Processus général d'extraction des données



Modèle d'extraction de données contenus dans des tableaux : Tablenet

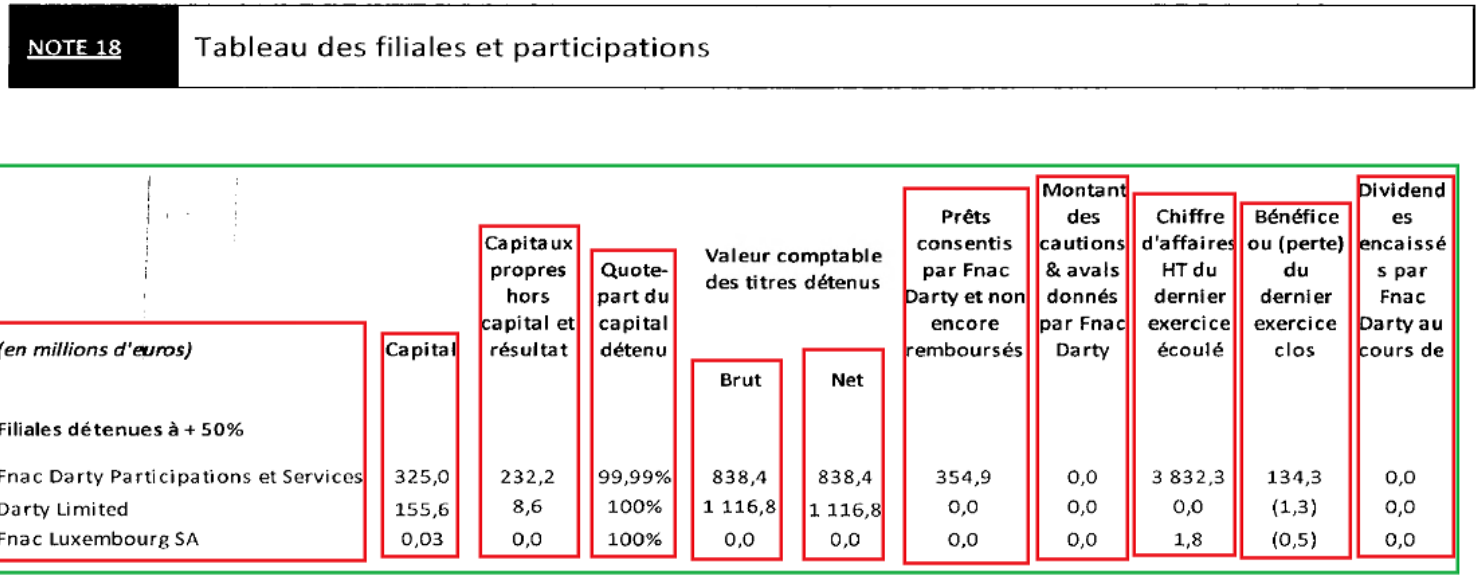


(b) Proposed TableNet architecture

Constitution base de travail du modèle d'extraction de données

Base apprentissage et test issue de la base de données Marmot :

- images issues de documents pdf annotés (tout type de tableaux)
- ajout d'images comportant des tableaux des comptes sociaux annotés manuel



NOTE 18 Tableau des filiales et participations

(en millions d'euros)	Capital	Capitaux propres hors capital et résultat	Quote-part du capital détenu	Valeur comptable des titres détenus		Prêts consentis par Fnac Darty et non encore remboursés	Montant des cautions & avals donnés par Fnac Darty	Chiffre d'affaires HT du dernier exercice écoulé	Bénéfice ou (perte) du dernier exercice clos	Dividendes encaissés par Fnac Darty au cours de
				Brut	Net					
Filiales détenues à + 50%										
Fnac Darty Participations et Services	325,0	232,2	99,99%	838,4	838,4	354,9	0,0	3 832,3	134,3	0,0
Darty Limited	155,6	8,6	100%	1 116,8	1 116,8	0,0	0,0	0,0	(1,3)	0,0
Fnac Luxembourg SA	0,03	0,0	100%	0,0	0,0	0,0	0,0	1,8	(0,5)	0,0

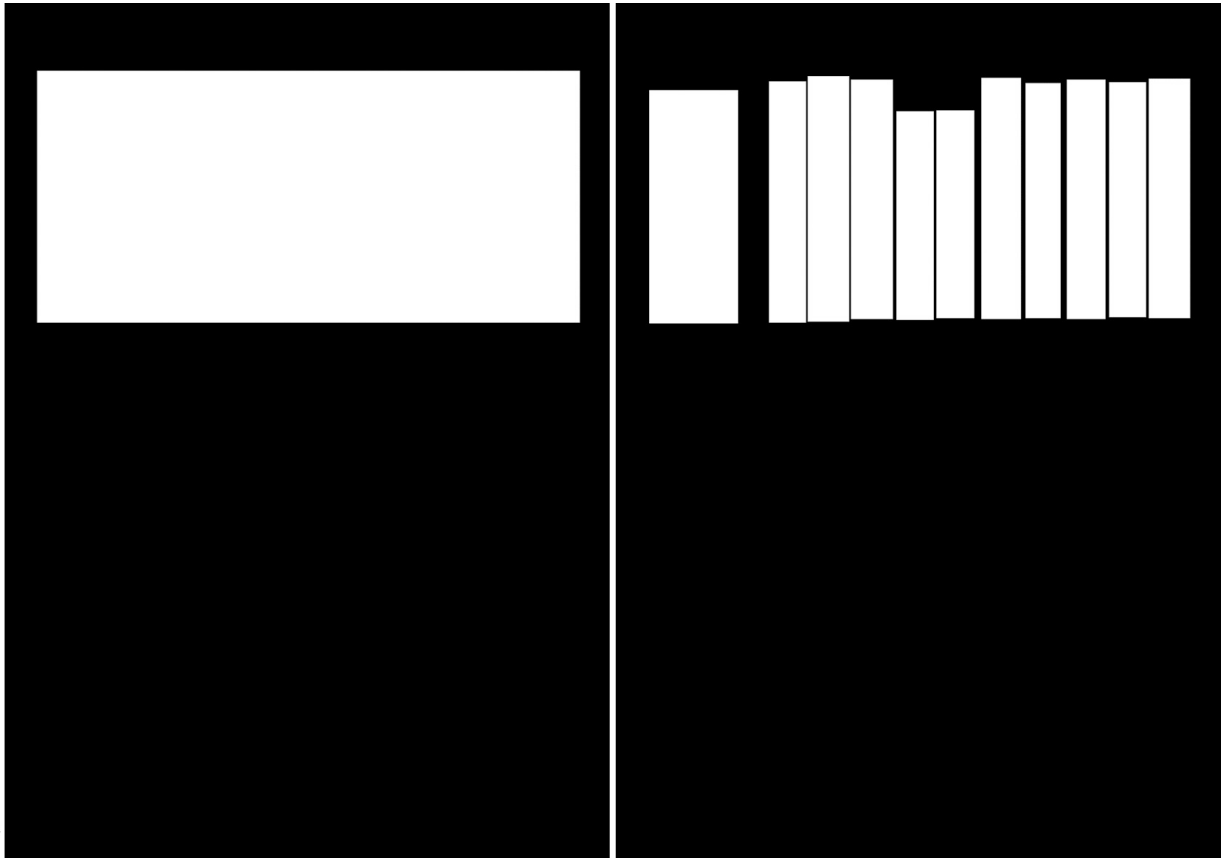
Constitution base de travail du modèle d'extraction de données

EIFFAGE METAL — Filiales et participations — Annexe au 31/12/2019

Dénomination	Capital	Capitaux Propres Hors Capital	Q.P. capital détenue (%)	Valeur comptable des titres détenus		Prêts, avances	Cautions	Chiffre d'affaires	Résultat	Dividendes encaissés
				Valeur brute	Valeur nette					
FILIALES (plus de 50%)										
SEH	102	18 002	100,00%	17 133	17 133	0		127 169	4 877	600
STAHLBAU ENGINEERING	294	562	100,00%	741	741			703	145	0
Eiffel Nigéria	25	89	100,00%	51	0			0	-139	768
Eiffage Métal Guyane	6 028	-4 648	100,00%	3 000	1 180			8 465	-429	0
Seruitiers Group NV	10 000	2 403	67,20%	75 000	75 000			11 875	1 345	400
Eiffage Métal UK	7	-32	100,00%	7	0			0	-2	0
Eiffage Métal Gestion y desarrollo	3	5 028	100,00%	3	3			10 608	4 883	3 360
Eiffage Métal Espana	31 877	4 269	100,00%	31 912	31 912	14 900		59 583	3 714	1 640
PARTICIPATIONS (10 à 50%)										
UNIBRIDGE	9 040	-2 779	49,00%	5 000	3 067			0	-545	0
AUTRES PARTICIPATIONS										

Page 29/30

Page 65 sur 66

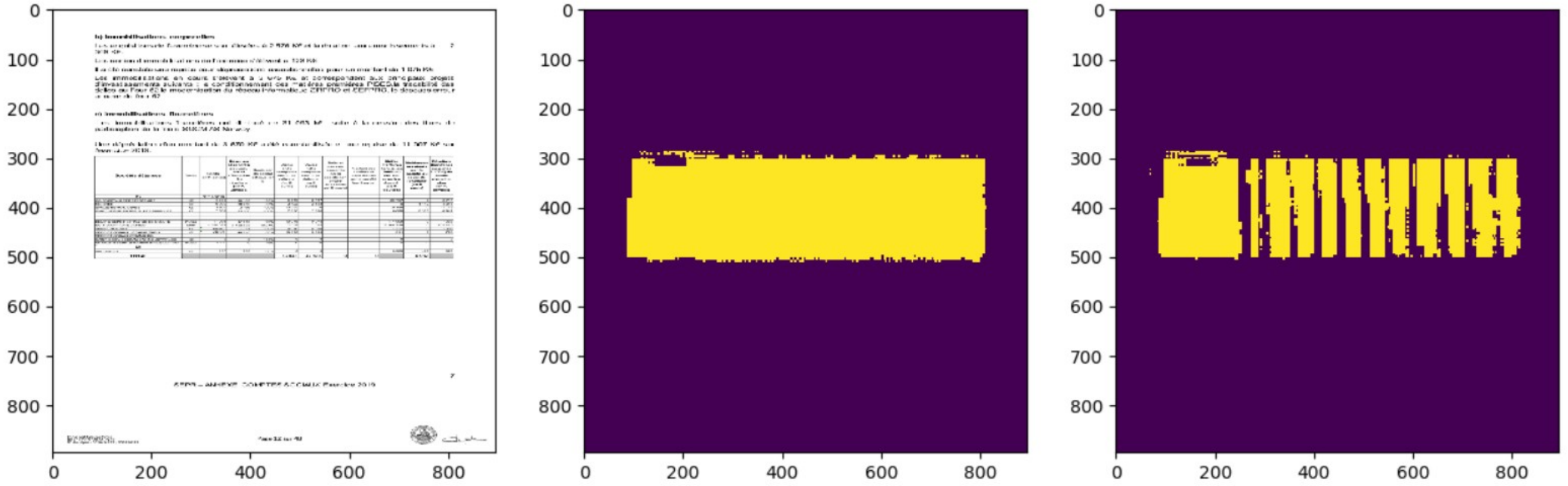


Exemple d'extraction d'informations

Sociétés détenues	Devise	Capital (en K devise)	Réserves et report à nouveau avant affectation des résultats (en K Devises)	Quote-part de capital détenue (en %)	Valeur brute comptable des titres détenus (en K euros)	Valeur nette comptable des titres détenus (en K euros)	Prêts et avances consentis par la société non encore remboursés (en K euros)	Montant des cautions et avals donnés par la société (en K euros)	Chiffre d'affaires hors taxes (100%) du dernier exercice écoulé (en K devises)	Dividendes encaissés par la société au cours de l'exercice (en K euros)	Résultats (bénéfices ou pertes à 100%) du dernier exercice clos (en K devises)
(1)		31/12/2019									
SG CRISTAUX ET DETECTEURS	K€	6 000	10 180	100%	5 519	2 747			23 767	0	-2 747
SG CREE	K€	6 900	28 950	35%	2 422	2 422			0	1 142	4 348
SAVOIE REFRACTAIRES	K€	3 905	3 166	100%	23 535	0			28 798	0	-1 118
SAINT GOBAIN MATERIAUX CERAMIQUES	K€	7 508	22 700	100%	7 630	7 630			16 688	5 375	6 947
BEIJING SEPR REFRACTORIES CO. LTD	KUSD	17 098	17 815	66%	10 075	10 075			174 909	0	7 332
SG INDIA PRIVATE LIMITED	KINR	1 355 263	3 452 633	10,21%	7 025	7 025			6 366 728	0	805 449
SEPR ITALIA SPA	K€	#DIV/0!	7 217	100%	4 020	4 020			2 372	0	2 333
SEPR ST GOBAIN KERAMIK GMBH	K€	#DIV/0!	46 815	100%	13 615	13 615			311	0	-728
SEPR ST GOBAIN KERAMIK KG											
SAINT-GOBAIN CERAMICAS INDUSTRIALES	K€	0	0	100%	0	0			0		0
SHANGHAI SEPR ZIRCONIUM PRODUCT CO. L	KUSD	2 200	0	76%	0	0			0	0	0
(2)											
VALOREF S.A.	K€	167	1635	100%	0	0			8 283	465	620
TOTAL					73 841	47 534	0	0		6 982	

Masques de prédictions du modèle entraîné avec les images Marmot

Résultat de la prédiction :



Exemple de masque prédit avec le modèle entraîné à l'aide de la base de données Marmot

Métriques mesurant la qualité de la prédiction

Résultat des indicateurs de qualité de la prédiction des emplacements de la table et de ses colonnes :

Table		Colonnes	
Intersection over Union	Taux de Perte	Intersection over Union	Taux de Perte
0.87	0.06	0.80	0.10

Résumé des résultats pour le modèle TableNet pour la base de données Marmot

→ plus le taux de perte est proche de 0, meilleure est la prédiction

→ plus l'IoU est proche de 1, meilleure est la prédiction

Application des masques de prédiction sur l'image

SGILILS LU 1YZ.		Valeur	Valeur	Montant des	Chiffre	Dividendes Résultats	
		brute	nette	cautions et	d'affaires		
Sociétés détenues		>omptable	somptable	avals donnés	hors taxes		
a)		des titres	des titres	par la société	(100%) du		
SG CRISTAUX ET DETECTEURS		détenus	détenus	(en K euros)	dernier		
SGCREE		(en K	(en K		exercice		
SAVOIE REFRACTAIRES		euros)	euros)		écoulé		
SAINT GOBAIN MATERIAUX CERAMIQUES		[5519	2747		(enK		
		7 2422	2422		devises)		
		7 23535	7630		23761		
		7630	10075		28 798]		
		10075	7025		16 68€		
		7 7025	4020		174 90€		
		4020			6 366 72€		
		13615	13615		2372		
BENING SEPR REFRACTORIES CO. LTD	Réserves	73841	47534		ET	ces bénéfices	
	et report à		Prêts et			encaissés	
SG INDIA PRIVATE LIMITED	pos Quote-part		avances		8 28:	ar la pu pertes	
	avan -		consentis			pere [à 100%) du	
SEPR ITALIA SPA	ï de capital		par la			société au :	
	Capital affectation Le Sep'		société non			dernier	
SEPR ST GOBAIN KERAMIK GMBH	en K devise) détenue (en		encore			cours de :	
	des %)		emboursés			. : exercice	
SEPR ST GOBAIN KERAMIK KG	résultats		en K euros			l'exercice	
	(enK		ol			(enk clos	
SAINT-GOBAIN CERAMICAS INDUSTRIALE	Devises)					euros) (enK	
	31/12/2019					devises)	
	6 000 10 180 100%					—	
SHANGHAI SEPR ZIRCONIUM PRODUCT C	6 900 28 950 35%					0 -2747	
r	3 905 3 166 100%					1142 4 348	
(2)	7 508 22 700 100%					0 -1118	
	17 098 17 815 66%					5 375 6 947	
VALORFF S.A.	1355 263| 3452633 10	21%					0 7 332
	#DIV/0! 7217 100%					805449	
TOTAL	#DIV/0! 46 815 100%					0 2 333	
	0 0 100%					0 -728	
	2200 0 76%					0	
	167 1635 100%					0 0	
						465620	
						6982	

→ résultat final du modèle dans un fichier csv

→ mots, expressions et chiffres bien retranscrits

→ **problème d'alignement des lignes et colonnes**

Application des masques de prédiction sur l'image après retraitements

xercice 2019.									
		Réserves						Chiffre	
						Prêts et			jividendes | Résultats
		et report à		Valeur	Valeur			d'affaires	
						avances			: La l(pénéfices encaissés
		nouveau	Quote-part	brute	nette	consentis	Montant des	hors taxes	Par 4 |3 400%) du ar la ou pertes
Sociétés détenues	Devise	Capital affectation : avant	de capital	comptable des titres	comptable des titres	par la	cautions et avals donnés	(100%) du dernier	société au : dernier
		(en K devise) des c	détenue (en			société non			cours de :
		résultats	%)	détenus (en K	détenus (en K	encore	par la société (en K euros)	exercice écoulé	l'exercice . : exercice
						remboursés			@nk clos
		(en K		euros)	euros)			(en K	
						(en K euros)			euros) tenK
		Devises)						devises)	devises)
(a)		31/12/2019 [| |							
SG CRISTAUX ET DETECTEURS	K€	6 000	100%	5519	2747			23 767	Q -2747
SGCREE	K€	6 900	35%	2 422	2 422			0	1142 4 348
SAVOIE REFRACTAIRES	K€	3 905	100%	23 535	Q			28 798	Q -1118
SAINT GOBAIN MATERIAUX CERAMIQUES	K€	7 508	100%	7 630	7 630			16 688	5375 6 947
		[1							
		[7]							
BENING SEPR REFRACTORIES CO. LTD	KUSD	17 098	66%	10075	10075			174 909	Q 7 332
SG INDIA PRIVATE LIMITED	KINR	1 355 263	10,21%	7 025	7 025			6 366 728	Q 805 449
SEPR ITALIA SPA	K€	#DIVO!	100%	4 020	4 020			2 372	Q 2333
SEPR ST GOBAIN KERAMIK GMBH	K€	#DIVO!	100%	13615	13615			311	0
SEPR ST GOBAIN KERAMIK KG		[7]							
SAINT-GOBAIN CERAMICAS INDUSTRIALES	K€	07 0	100%	Q	Q			0	
SHANGHAI SEPR ZIRCONIUM PRODUCT CC	KUSD	2200] 7777 0	76%	Q	Q			0	0
(2)		|						
VALOREF S.A.	K€		100%	Q	0			8 283	
TOTAL				73 841	47 534	0	0		

Application des masques après retraitements

ETAT DES FILIALES ET PARTICIPATIONS

Dénomination Siège Social	Capital Capitaux Propres	Q.P. Détenue Dividendes Encaissés	Val. Brute Titres Val. Nette Titres	Chiffre d'Affaires en Euros	Résultat de l'exercice en Euros
FILIALES (plus de 50%)					
ACCORINVEST AUSTRIA GMBH	250 000	100,00%	250 000	87 507 000	2 040 573
Autriche	4 671 000	0	250 000		
PORTIS SA	1 239 895	100,00%	99 350 000	67 035 851	10 557 036
Portugal	65 596 989	0	99 350 000		
HOTELINVEST HOLDING GMBH	25 000	100,00%	274 000 000		-5 764 857
Allemagne .	287 832 370	0	274 000 000		
SOCIETE HOTELIERE ATHENES CENTRE	9 164 000	100,00%	7 662 161	6 400 871	826 009
Gréce	1 445 751	0	7 662 161		
ACCOR HOSPITALITY ITALIA SRL	305 300 000	100,00%	419 023 330	101 598 079	21 127 569
Italie	107 155 346	0	419 023 330		
ECO HOTELS BUSSIGNY	890 720	100,00%	6 473 615	2 139 444	-805 716
Suisse	-675 449	0	6 473 615		
SHRE	21 266 999	100,00%	26 848 063		882 284
91080 EVRY-COURCOURONNES	2 614 429	0	26 848 063		
SOCIETE DE GESTION HOTELINVEST	200 000	100,00%	1 700 000	107 031 540	2 995 169
75012 PARIS	2 976 896	0	1 700 000		
SPH	627 996 800	100,00%	879 261 162	0	5 120 471
75012 PARIS	262 116 609	0	879 261 162		
ACCORINVEST SWITZERLAND	2 175 584	100,00%	172 777 735	97 194 898	51 452 771
Suisse	59 049 265	0	172 777 735		
FRANCIMMO	7 963 000	100,00%	66 739 975	0	25 951
75012 PARIS	826 000	0	66 739 975		
AURORA HOLDING	25 000	94,90%	1 229 370	0	-973 872
Allemagne	144 609	0	1 229 370		
BALHOTEL	4 453 600	70,00%	23 254 333	4 821 835	1 330 938
Suisse	2 530 363	0	20 446 333		
ACCORHOSPITALITY NEDERLAND	6 929 597	58,09%	140 975 532	172 864 000	-2 648 000
PAYS-BAS	73 193 403	0	140 975 532		

Application des masques de prédiction sur l'image

LL LR LR 22) V7					
. PARIS 12E					
TRES	Caphai en				
	sr À		TR LES .R	| RAD E GX	AO
L CRETE)					
EILIALES {plus de 50%)					
ACCORINVEST AUSTRIA GMEH	250 000	100,00%	250 000	87 507 000	2 040 573
Autriche	4 671 000	0	250 000		
PORTIS SA	1 239 895	100,00%	99 350 000	67 035 851	10 557 036
Portugal	65 596 989	0	99 350 000		
HOTELINVEST HOLDING GMBH	25 000	100,00%	274 000 000		-5 764 857
Allemagne .	287 832 370	0	274 000 000		
SOCIETE HOTELIERE ATHENES CENTR	9 164 000	100,00%	7 862 161	6 400 871	826 009
Grége	1445751	0	7 662 161		
ACCOR HOSPITALITY ITALIA SRL	305 300 000	100,00%	419 023 330	101 598 079	21 127 569
Italie	107 155 346	0	419 023 330		
ECO HOTELS BUSSIGNY	890 720	100,00%	6 473 615	2 139 444	-805 716
Suisse	-675 449	0	6473615		
SHRE	21 266 999	100,00%	26 848 063		882 284
91080 EVRY-COURCOURONNES	2614 429	0	26 84B 063		
SOCIETE DE GESTION HOTELINVEST	200 000	100,00%	1 700 000	107 031 540	2 995 169
75012 PARIS	2 976 896	0	1 700 000		
SPH	627 996 800	100,00%	879 261 162	0	5120471
75012 PARIS	262 116 609	0	879 261 162		
ACCORINVEST SWITZERLAND	2 175 584	100,00%	172 777 735	97 194 898	51452771
Suisse	59 049 265	0	172777 735		
FRANCIMMO	7 963 000	100,00%	66 739 975	D	25 951
75012 PARIS	826 000	0	66 739 975		
AURORA HOLDING	25 000	94,90%	1 229 370	0	-973 872
Allemagne	144 609	0	1 229 370		
BALHOTEL	4 453 600	70,00%	23 254 333	4 821 835	1 330 938
Suisse	2 530 363	0	20 446 333		
ACCORHOSPITALITY NEDERLAND	6 929 597	58,09%	140 975 532	172 864 000	-2 648 000
PAYS-BAS	73 193 403	0	140 975 532		

Limite du processus d'extraction

→ Retraitement et application complexe des masques prédits (lissage des frontières)

→ Prédiction moins précise pour les images en couleur (ou avec beaucoup de nuances de gris)

→ Quelques erreurs de reconnaissance de caractères avec le moteur tesseract (lettres à la place de chiffres, virgules oubliées...)

→ Certains tableaux sont trop « dégradés » pour pouvoir extraire leurs informations (scan de mauvaise qualité)

Conclusion

Perspectives d'application de cette expérimentation :

- extraire d'autres informations des comptes sociaux (explications sur des évolutions/ informations sur des restructurations…)

- application à d'autres documents comportant des tableaux (réduction de la charge d'enquête)

- application à la facturation des entreprises (facturation électronique obligatoire pour l'ensemble des entreprises dès 2026)

- application dans d'autres domaines que les entreprises