Exploitation de Big Data Historiques pour les Humanités Numériques : application aux données financières HBDEX

Camille Guerry, Bertrand B. Coüasnon, Aurélie Lemaitre, Sébastien Adam, Thierry Paquet, Andres Rojas Camacho

NUMÉRIQUE ET PATRIMOINE
Enjeux et questions actuels

DIGITAL TECHNOLOGY AND HERITAGE - CHALLENGES AND ISSUES

11/12 MARS 2021

Conférence en ligne
Digital conference

GOUVERNEMENT
Liberté
Égalité
Fraternité

AGENCE NATIONALE DE LA RECHERCHE
ANR

Auteurs : Camille GUERRY, Andres ROJAS CAMACHO,
Bertrand COÜASNON, Aurélie LEMAITRE, Sebastien ADAM, Thierry PAQUET

Cultural Heritage
AND GLOBAL CHANGE

Exploitation de Big Data Historiques pour les Humanités Numériques : application aux données financières

HBDEX

Coordinateur(s): PIERRE CYRILLE HAUTCOEUR

Partenaires : PSE, LITIS, INSA-INRIA, CAMS

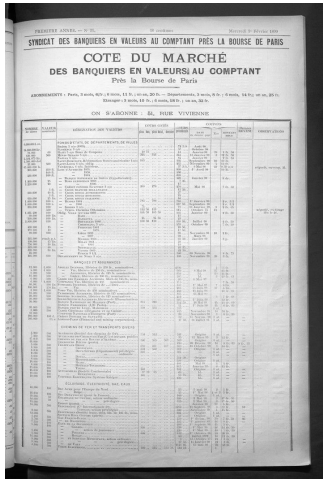
Présentation : Camille GUERRY, Andres ROJAS CAMACHO

Résumé :
L'objectif est d'étudier le comportement à long terme des marchés financiers, en extrayant une grande quantité de données à partir d'images de documents. Nous proposons une stratégie de reconnaissance qui exploite la séquentialité des documents.

Analyse de documents d'histoire financière

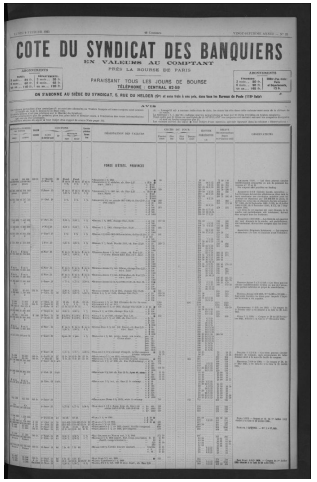
- Cotes boursières
  - Marché de La Coulisse à Paris
  - 1899 à 1939
- Intérêt de l'analyse des données
  - Modélisation des marchés financiers
- Une des problématiques du projet
  - Extraction des cotations dans des images de documents
- Difficultés classiques liées aux documents anciens
  - Documents très denses, abîmés, effacés
  - Impression de mauvaise qualité

  => OCR commerciaux non satisfaisants

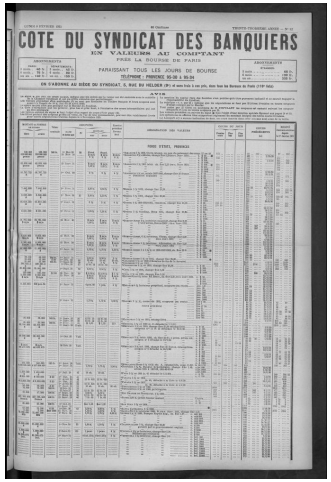

SYNDICAT DES BANQUIERS EN VALEURS AU COMPTANT PRES LA BOURSE DE PARIS

COTE DU MARCHÉ
DES BANQUIERS EN VALEURS AU COMPTANT
Près la Bourse de Paris

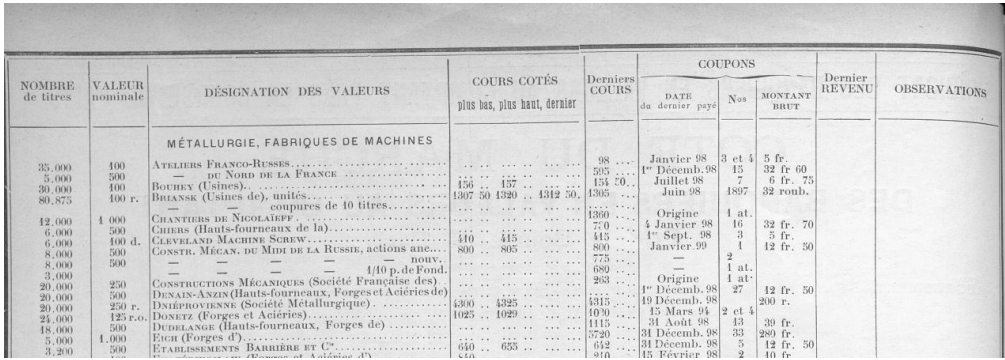
ON S'ABONNE RUE VIVIENNE


COTE DU SYNDICAT DES BANQUIERS
EN VALEURS AU COMPTANT
PRES LA BOURSE DE PARIS

PARAISSANT TOUS LES JOURS DE BOURSE

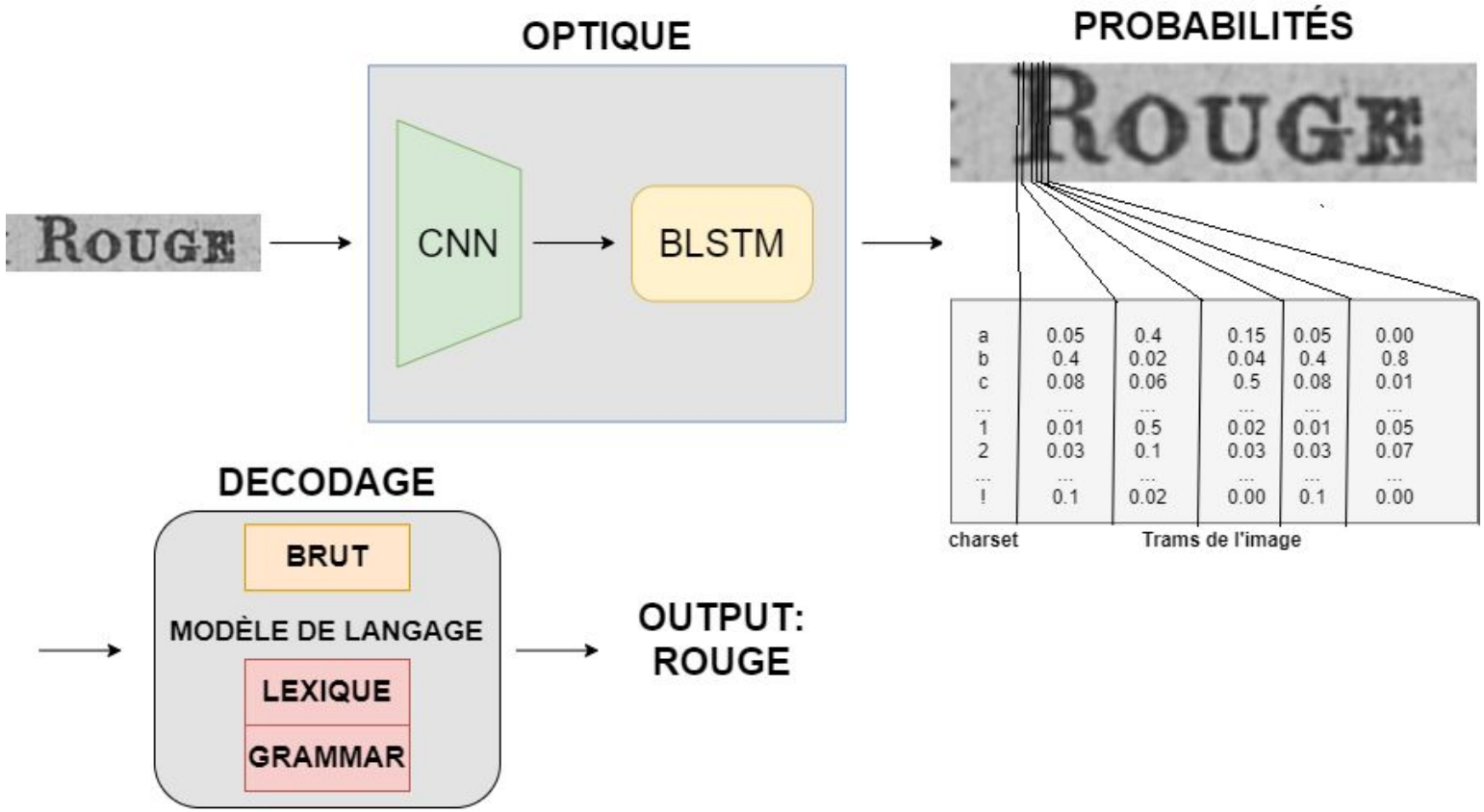

COTE DU SYNDICAT DES BANQUIERS
EN VALEURS AU COMPTANT
PRES LA BOURSE DE PARIS

PARAISSANT TOUS LES JOURS DE BOURSE

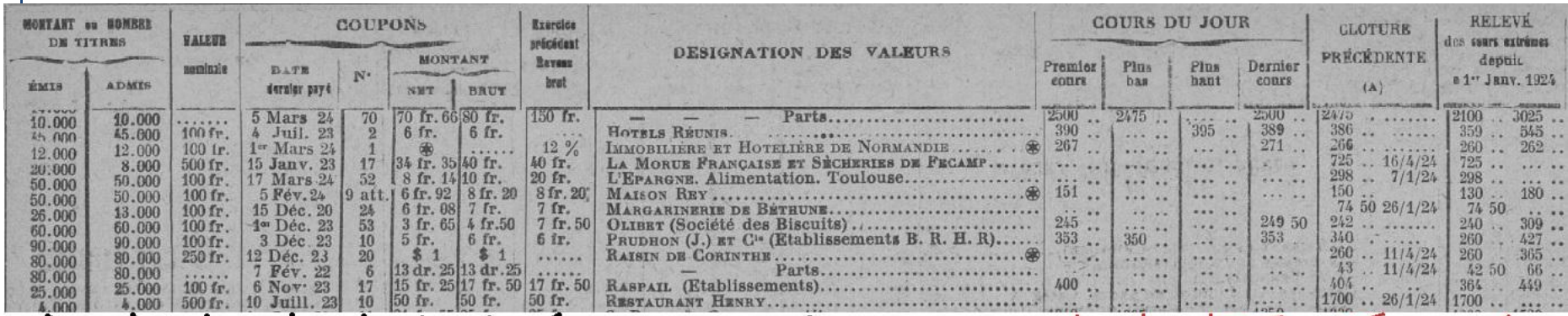


NOMBRE de titres	VALEUR nominale	DÉSIGNATION DES VALEURS	COURS COTÉS plus bas, plus haut, dernier	Derniers COURS	COUPONS DATE du dernier payé	COUPONS Nos	COUPONS MONTANT BRUT	Dernier REVENU	OBSERVATIONS
		MÉTALLURGIE, FABRIQUES DE MACHINES							
35.000	100	Ateliers Franco-Russes		98	Janvier 98	3 et 4	5 fr.		
5.000	500	— du Nord de la France		505	1er Décemb. 98	15	32 fr 60		
30.000	100	Bouhey (Usines)	156 .. 157	154 50	Juillet 98	7	6 fr. 75		
80.875	100 r.	Briansk (Usines de), unités	1307 50 1320 .. 1312 50	1305	Juin 98	1897	32 roub.		
		— — coupures de 10 titres							
12.000	1 000	Chantiers de Nicolaïeff		1350	Origine	1 at.			
6.000	500	Chiers (Hauts-fourneaux de la)		770	4 Janvier 98	16	32 fr. 70		
6.000	100 d.	Cleveland Machine Screw	410 .. 415	445	1er Sept. 98	3	5 fr.		
6.000	500	Constr. Mécan. du Midi de la Russie, actions anc.	800 .. 805	800	Janvier 99	1	12 fr. 50		
8.000	500	— — — — — nouv.		775	—	2			
8.000		— — — — — 1/10 p. de Fond.		680		1 at.			
3.000		Constructions Mécaniques (Société Française des)		263	Origine	1 at.			
20.000	250	Denain-Anzin (Hauts-fourneaux, Forges et Aciéries de)			1er Décemb. 98	27	12 fr. 50		
20.000	500	Dniéprovienne (Société Métallurgique)	4300 .. 4325	4315	19 Décemb. 98		200 r.		
20.000	250 r.	Donetz (Forges et Aciéries)	1025 .. 1029	1030	15 Mars 94	2 et 3			
24.000	125 r.o.	Dudelange (Hauts-fourneaux, Forges de)		1115	31 Août 98	43	30 fr.		
18.000	500	Eich (Forges d')		5720	31 Décemb. 98	33	280 fr.		
5.000	1.000	Etablissements Barrière et Cie	640 .. 655	642	31 Décemb. 98	5	12 fr. 50		
3.200	500			810	15 Février 98	3	10 fr.		

OCR : Schéma fonctionnement



OCR: Exemple de grammaire



MONTANT ou NOMBRE DE TITRES ÉMIS	MONTANT ou NOMBRE DE TITRES ADMIS	VALEUR nominale	COUPONS DATE dernier payé	COUPONS N°	COUPONS MONTANT NET	COUPONS MONTANT BRUT	Exercice précédent Revenu brut	DESIGNATION DES VALEURS	COURS DU JOUR Premier cours	COURS DU JOUR Plus bas	COURS DU JOUR Plus haut	COURS DU JOUR Dernier cours	CLOTURE PRÉCÉDENTE (A)	RELEVÉ des cours extrêmes depuis le 1er Janv. 1924
10.000	10.000	.....	5 Mars 24	70	70 fr. 66	80 fr.	150 fr.	— — — Parts	2500	2475	....	2500	2475	2100 3025
45.000	45.000	100 fr.	4 Juil. 23	2	6 fr.	6 fr.	....	HOTELS RÉUNIS	390	....	395	389	386	359 545
12.000	12.000	100 fr.	1er Mars 24	1	❀	.....	12 %	IMMOBILIÈRE ET HOTELIÈRE DE NORMANDIE ❀	267	....	....	271	266	260 262
20.000	8.000	500 fr.	15 Janv. 23	17	34 fr. 35	40 fr.	40 fr.	LA MORUE FRANÇAISE ET SÉCHERIES DE FECAMP	....	....	....	....	725 16/4/24	725 ....
50.000	50.000	100 fr.	17 Mars 24	52	8 fr. 14	10 fr.	20 fr.	L'EPARGNE. Alimentation. Toulouse	....	....	....	....	298 7/1/24	298 ....
50.000	50.000	100 fr.	5 Fév. 24	9 att.	6 fr. 92	8 fr. 20	8 fr. 20	MAISON REY ❀	151	....	....	....	150	130 180
26.000	13.000	100 fr.	15 Déc. 20	24	6 fr. 08	7 fr.	7 fr.	MARGARINERIE DE BÉTHUNE	....	....	....	....	74 50 26/1/24	74 50 ....
60.000	60.000	100 fr.	1er Déc. 23	53	3 fr. 65	4 fr.50	7 fr. 50	OLIBET (Société des Biscuits)	245	....	....	249 50	242	240 309
90.000	90.000	100 fr.	3 Déc. 23	10	5 fr.	6 fr.	6 fr.	PRUDHON (J.) ET Cie (Etablissements B. R. H. R)	353	350	....	353	340	260 427
80.000	80.000	250 fr.	12 Déc. 23	20	$ 1	$ 1	.....	RAISIN DE CORINTHE ❀	....	....	....	....	260 11/4/24	260 365
80.000	80.000	.....	7 Fév. 22	6	13 dr. 25	13 dr. 25	.....	— Parts	....	....	....	....	43 11/4/24	42 50 66
25.000	25.000	100 fr.	6 Nov. 23	17	15 fr. 25	17 fr. 50	17 fr. 50	RASPAIL (Etablissements)	400	....	....	....	404	364 449
4.000	4.000	500 fr.	10 Juill. 23	10	50 fr.	50 fr.	50 fr.	RESTAURANT HENRY	....	....	....	....	1700 26/1/24	1700 ....

Lexique répétitif

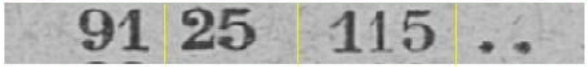
grammaire_cours_du_jours

montant, centimes

grammaire_cloture_précédente

montant, centimes, date

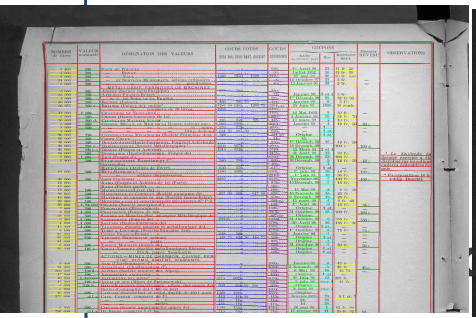
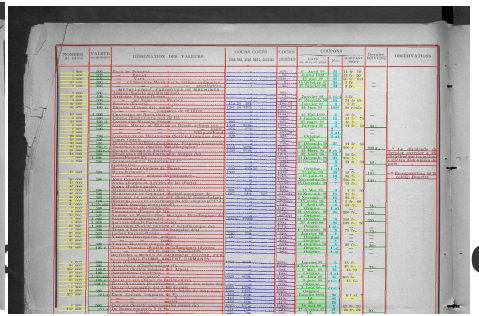
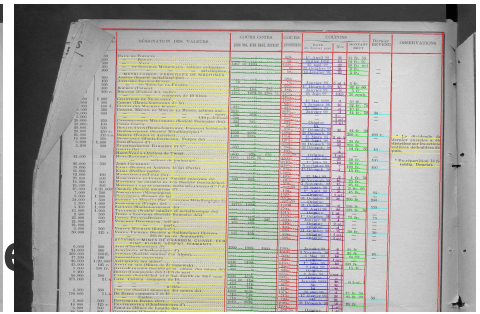
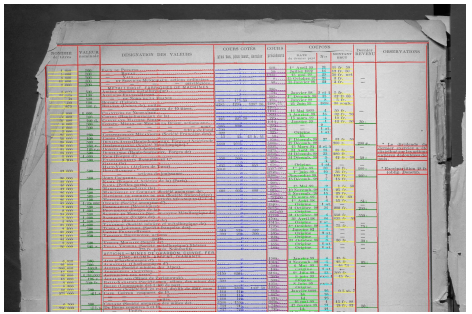
grammaire_relevé_cours_ext



montant, centimes, montant, centimes

Stratégie : importance de la collection

- Exploiter les stabilités pour pallier les erreurs de reconnaissances
- Détecter automatiquement les séquences stables / points de rupture

Marcinelle et Couillet (Société anonyme de)	340
Moteurs a gaz et constructions mécaniques (Cie fse d.)	160
Ougrée-Marihaye (Société anonyme d')	1170
Providence (Forges de la)	...
Sambre et Moselle (Soc. anonyme Métallurgique de)	...
Sarrebruck (Forges de)	...
Taganrog (Société minière et métallurgique de)	...
Tubes a Louvroil (Société française des) estampillées	...
Usines Franco-Russes	553
Verchny Dnieprovsk	..
— — parts	.
Villelongue (Usines Électro-Métallurgiques de)	48
Vireux Molhain (forges de)	...
Volga Vichera (Société métallurgique) libérées	..
Wattelar-Francq (Usines)	...

Marcinelle et Couillet (Société anonyme de)	302 50
Moteurs a gaz et constructions mécaniques (Cie fse d.)	... ..
Ougrée-Marihaye (Société anonyme d')	... ..
Providence (Forges de la)	2900 ..
Sambre et Moselle (Soc. anonyme Métallurgique de)	...
Sarrebruck (Forges de)	8850 ..
Taganrog (Société minière et métallurgique de)	650 .
Tubes a Louvroil (Société française des) estampillées	... ..
Usines Franco-Russes	560 ..
Verchny Dnieprovsk	27 ..
— — parts	. .
Villelongue (Usines Électro-Métallurgiques de)	.. ..
Vireux Molhain (forges de)	... ..
Volga Vichera (Société métallurgique) libérées	... ..
Wattelar-Francq (Usines)	... ..

...

Marcinelle et Couillet (Société anonyme de)	320
Moteurs a gaz et des const. mécaniques (Cie fse des) (3)	115
Ougrée-Marihaye (Société anonyme d')	...
Providence (Forges de la)	2300
Sambre et Moselle (Soc. anonyme Métallurgique de)	...
Sarrebruck (Forges de)	6935
Taganrog (Société métallurgique de)	.
Tubes a Louvroil (Soc. fse. pl la fabricat[on] des) estamp.	1085
Usines Franco-Russes (Anc. Etabl. Baird)	425
Verchny-Dnieprovsk (Compagnie Métallurgique de)	119
— — parts	..
Vireux-Molhain (forges de)	...
Volga-Vichera (Société minière et métallurgique)	..
Wattelar-Francq (Usines)	...

Conclusion

- **Résultats**
  - **OCR**
    - Evaluation sur une page de 1899

      -> Meilleur OCR du commerce: 84.72%

      -> OCR (réseau de neurones) + Modèle de langage: 99.12%
  - **Identification des titres boursiers**
    - Evaluation sur 4 jours (16 images) de 1899, 1696 titres boursiers

      -> Apport de la collection : F-mesure : passe de 91,4% à 98.8 % (100% visé)
- **Perspectives**
  - **Analyse des données par le CAMS**
  - **Etudes de modèles économiques par PSE**

ANR AGENCE NATIONALE DE LA RECHERCHE

HBDEX

Slides supplémentaires

Visio - 27/01/2021

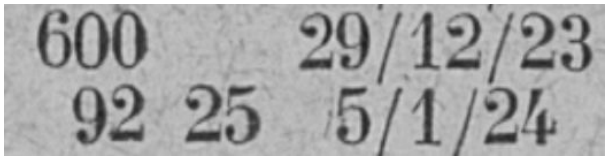
OCR: Création d'une grammaire et lexique

La prédiction brute se limite au visuel, elle est sensible aux imperfections dans l'image. L'utilisation d'un modèle de langage permet de corriger la sortie à l'aide d'un lexique ou d'une grammaire pré-établis.

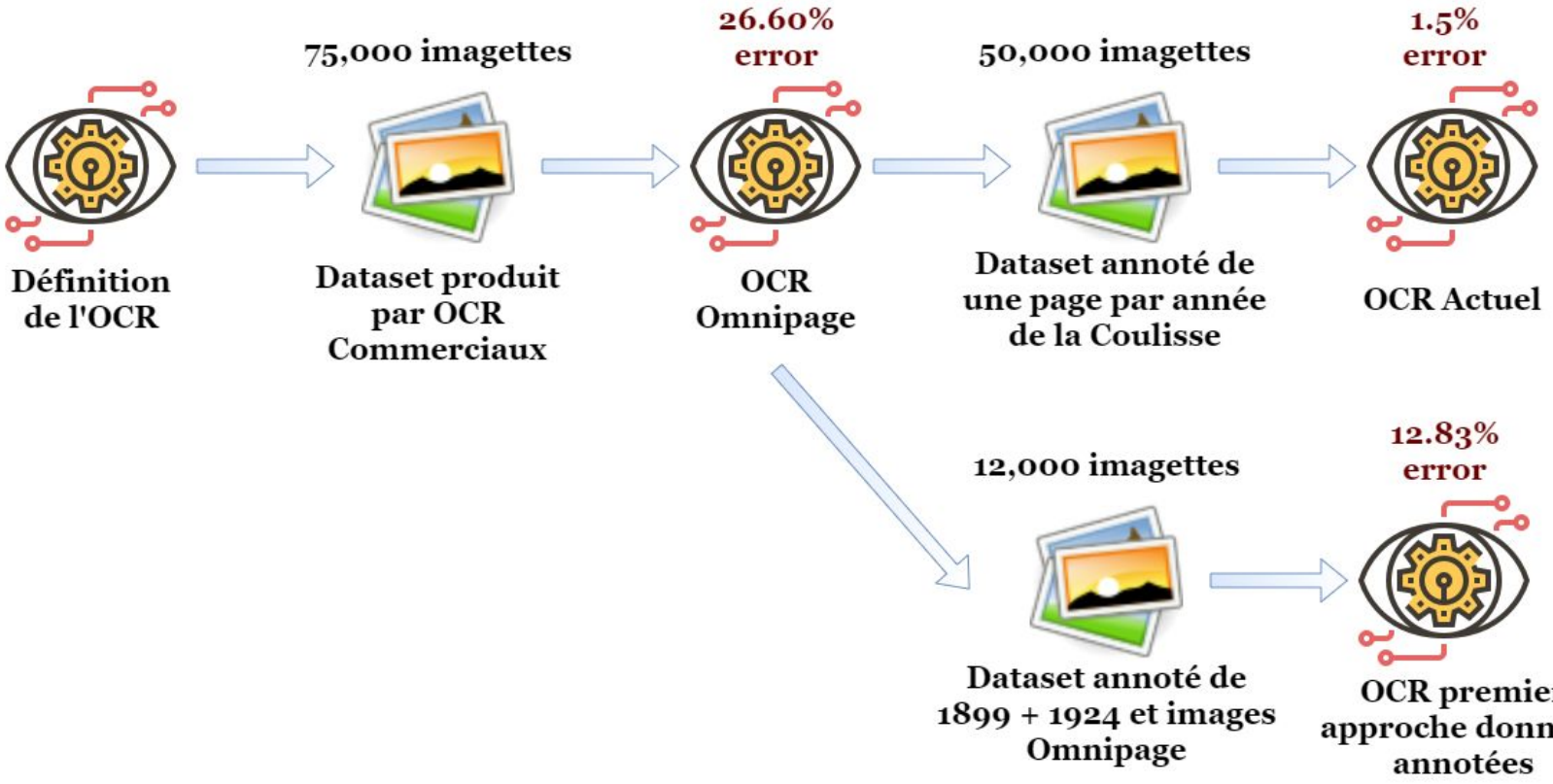
COmLACIAUX	COMMERCIAUX ... COMMUNAUTÉ	COMMERCIAUX

Le lexique s'utilise dans les situations où l'objet en question appartient à une liste de candidats. Pour les données qui changent un d'un jour à l'autre, il faut utiliser une grammaire pour définir un format de sortie pour avoir une prédiction correcte.

{##} {##} ##/##/## = 222 25 12/1/23



OCR : Apprentissage



OCR: Outil d'annotation

HBDEX - Côte Syndicat 1924_page8_chaque15

(Threshold : 0.9)

Exit without saving

ALIMENTATION, BRASSERIES, HOTELS,

SUCRERIES

ALIMENTATION (Société d'), en liquidation

ALIMENTATION DE PROVENCE

AU PLANTEUR DE CAIFFA, privil. 8% cum

- - ordinaires §

- - Parts

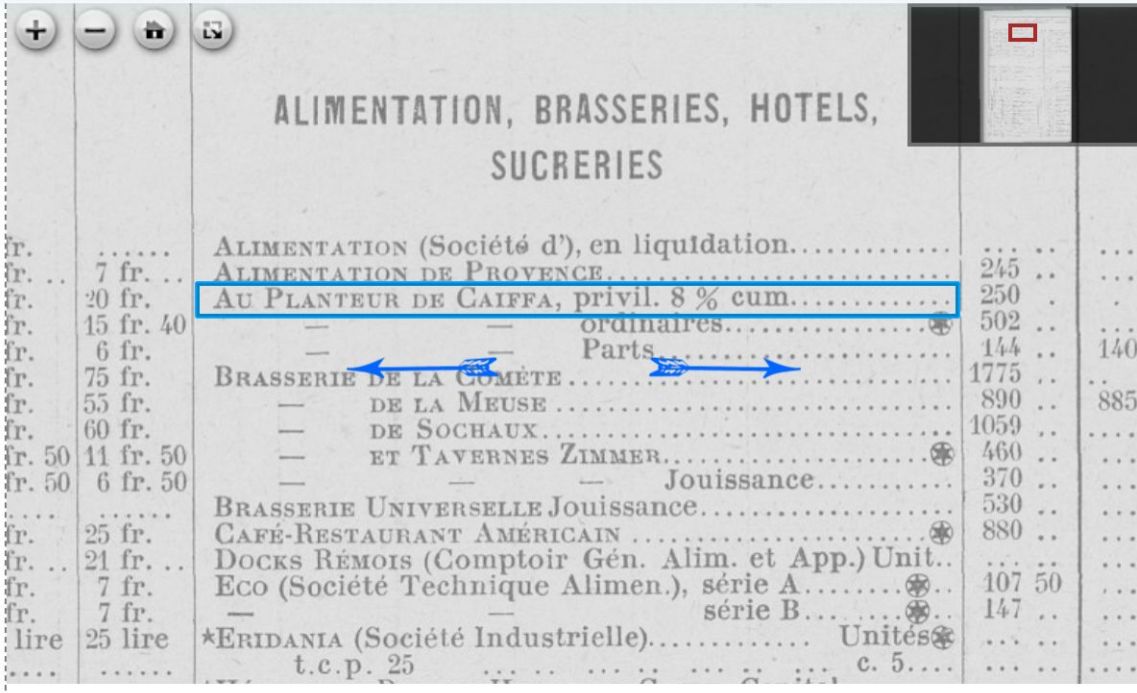
BRASSERIE DE LA COMÈTE

- DE LA MEUSE

- DE SOCHAUX

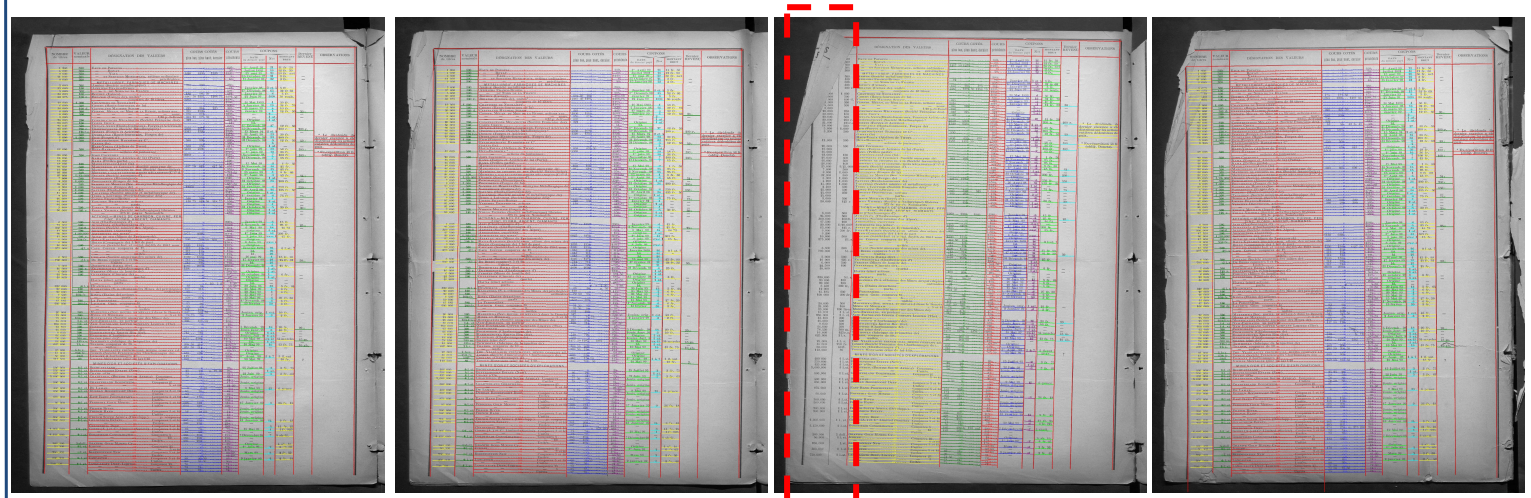
- ET TAVERNES ZIMMER §

- - - Jouissance



Détection de ruptures

1. Analyse d'image (grammaire et OCR) : extraction des colonnes sans a priori sur tout le corpus
2. Fiabilisation transversale : détection de ruptures et de plages stables
3. Analyse d'image : extraction des colonnes fiabilisées



Alignement de séquences

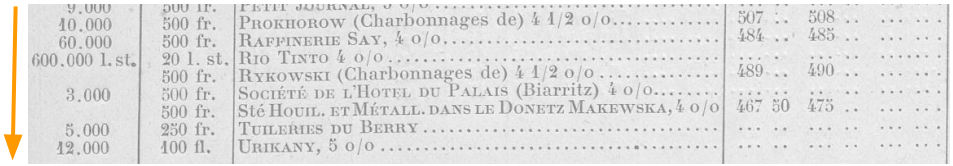
1. **Analyse d'image (grammaire et OCR) : extraction des titres sans a priori sur tout le corpus**
2. **Fiabilisation transversale : alignement de séquences**

S1	PETIT JOURNAL, 5 %	PROKHOROW	RAFFINERIE	RYKOWSKI'	Ste HOUIL. ET METALL.	TUILERIES	URIKANY, 5 %	URIKANY, 4 %
S2	PETIT JOURNAL	PROKHOROW	RAFFINERIE	RYKOWSKI	Ste HOUIL. ET METALL.	TUILERIES	URIKANY, 5 %	
S3	PETIT JOURNAL, 5 %	KHOROW	RAFFINERIE	RYKOWSKI	Ste HOUIL. ET METALL.	ATUILERIES	URIKANY, 5 %	
S4	PETIT JOURNAL, 5 %	PROKHOROW	RAFFINERIE	RIO TINTO 4 %	RYKOWSKI	Ste HOUIL.METALL.	TUILERIES	URIKANY, 5 %

	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6	ID 7	ID 8	ID 9
S1'	PETIT JOURNAL, 5 %	PROKHOROW	RAFFINERIE	-	RYKOWSKI'	Ste HOUIL. ET METALL.	TUILERIES	URIKANY, 5 %	URIKANY, 4 %
S2'	PETIT JOURNAL	PROKHOROW	RAFFINERIE	-	RYKOWSKI	Ste HOUIL. ET METALL.	TUILERIES	URIKANY, 5 %	-
S3'	PETIT JOURNAL, 5 %	KHOROW	RAFFINERIE	-	RYKOWSKI	Ste HOUIL. ET METALL.	ATUILERIES	URIKANY, 5 %	-
S4'	PETIT JOURNAL, 5 %	PROKHOROW	RAFFINERIE	RIO TINTO 4 %	RYKOWSKI	Ste HOUIL.METALL.	TUILERIES	URIKANY, 5 %	-

3. **Rapprochement avec le vocabulaire pour l'attribution d'un identifiant DFIH**
4. **Sollicitation utilisateur**

1. **Analyse d'image (grammaire et OCR) : extraction des titres sans a priori sur tout le corpus**



10.000	500 fr.	Prokhorow (Charbonnages de) 4 1/2 o/o	507 .. 508 ..	... ...
60.000	500 fr.	Raffinerie Say, 4 o/o	484 .. 485 ..	... ...
600.000 l. st.	20 l. st.	Rio Tinto 4 o/o	... .	... ...
	500 fr.	Rykowski (Charbonnages de) 4 1/2 o/o	489 .. 490 ..	... ...
3.000	500 fr.	Société de l'Hotel du Palais (Biarritz) 4 o/o	... .	... ...
	500 fr.	Sté Houil. et Métall. dans le Donetz Makewska, 4 o/o	467 50 475 ..	... ...
5.000	250 fr.	Tuileries du Berry	... .	... ...
12.000	100 fl.	Urikany, 5 o/o	... .	... ...

2. **Fiabilisation transversale : alignement de séquences**

S1	PETIT JOURNAL, 5 %	PROKHOROW	RAFFINERIE	RYKOWSKI'	Ste HOUIL. ET METALL.	TUILERIES	URIKANY, 5 %	URIKANY, 4 %
S2	PETIT JOURNAL	PROKHOROW	RAFFINERIE	RYKOWSKI	Ste HOUIL. ET METALL.	TUILERIES	URIKANY, 5 %	
S3	PETIT JOURNAL, 5 %	KHOROW	RAFFINERIE	RYKOWSKI	Ste HOUIL. ET METALL.	ATUILERIES	URIKANY, 5 %	
S4	PETIT JOURNAL, 5 %	PROKHOROW	RAFFINERIE	RIO TINTO 4 %	RYKOWSKI	Ste HOUIL.METALL.	TUILERIES	URIKANY, 5 %

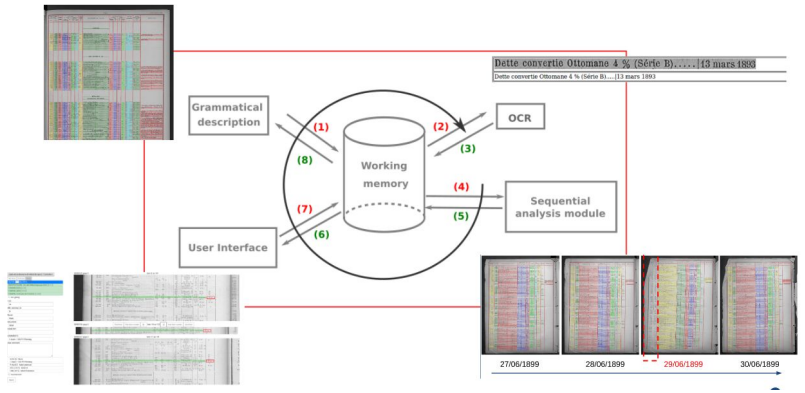
	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6	ID 7	ID 8	ID 9
S1'	PETIT JOURNAL, 5 %	PROKHOROW	RAFFINERIE	-	RYKOWSKI'	Ste HOUIL. ET METALL.	TUILERIES	URIKANY, 5 %	URIKANY, 4 %
S2'	PETIT JOURNAL	PROKHOROW	RAFFINERIE	-	RYKOWSKI	Ste HOUIL. ET METALL.	TUILERIES	URIKANY, 5 %	-
S3'	PETIT JOURNAL, 5 %	KHOROW	RAFFINERIE	-	RYKOWSKI	Ste HOUIL. ET METALL.	ATUILERIES	URIKANY, 5 %	-
S4'	PETIT JOURNAL, 5 %	PROKHOROW	RAFFINERIE	RIO TINTO 4 %	RYKOWSKI	Ste HOUIL.METALL.	TUILERIES	URIKANY, 5 %	-

3. **Rapprochement avec le vocabulaire pour l'attribution d'un identifiant DFIH**
4. **Sollicitation utilisateur**

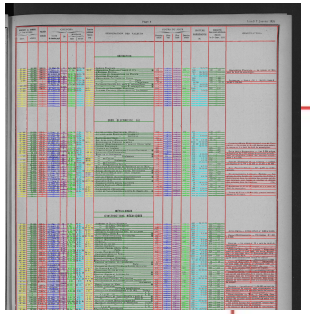
Application à la collection de listes de prix

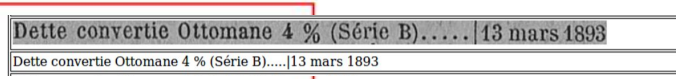
Stratégie complète de reconnaissance

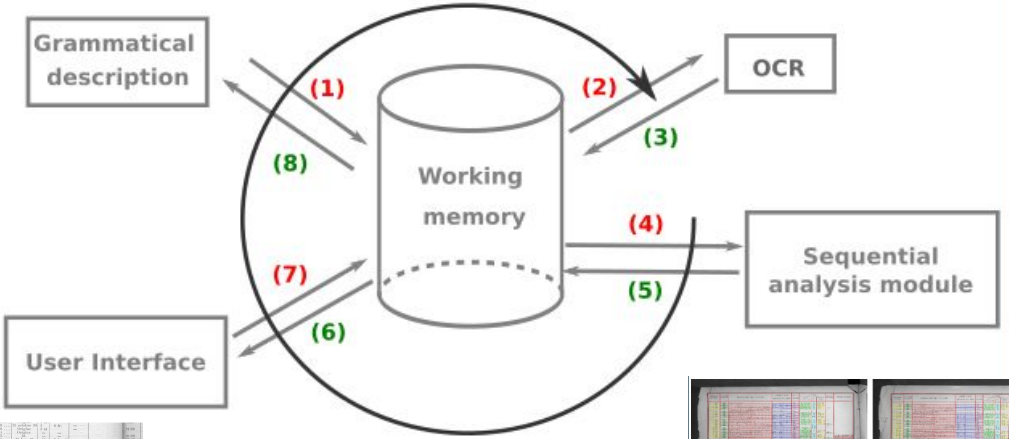
- Itération 1 : colonnes
  - Extraction des colonnes sans a priori sur tout le corpus
  - Fiabilisation transversale
  - Extraction des colonnes fiabilisées
- Itération 2 : sections
  - Extraction des sections sans a priori sur tout le corpus
  - Fiabilisation transversale
  - Extraction des sections fiabilisées
- Itération 3 : titres
  - Extraction des titres sans a priori sur tout le corpus
  - Fiabilisation transversale
  - Attribution d'un identifiant DFIH
  - Sollicitation utilisateur
- Itération 4 : autres champs
  - Fiabilisation transversale
  - Identification des devises
  - Sollicitation utilisateur
- Production du XML

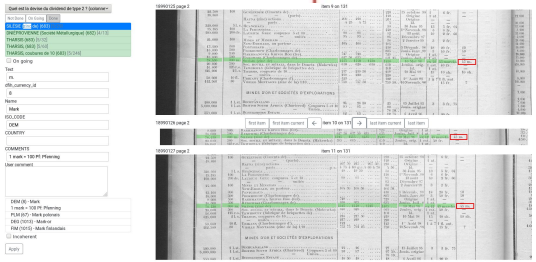


Système d'analyse de collection









OCR: Résultats OCR + Modèle de langage

Test page 1899

OCR	Character Error Rate
Tesseract	31.62%
Omnipage	15.28%
OCR Omni + 1924 + 1899	11.03%
OCR Omni + 1924 + 1899 + Modèle de langage	3.60%
OCR une page par année	1.41%
OCR une page par année + Modèle de langage	0.88%

Résultats quantitatifs

Nombre d'images	Nb d'erreurs avant fiabilisation	Nb d'erreur après fiabilisation
331	14	0

Identification des titres boursiers

	Précision	Rappel	F-mesure
Pas de prise en compte de la collection (pages isolées)	0.913	0.914	0.914
Prise en compte de la collection (pas d'interaction)	0.931	0.932	0.931
Prise en compte de la collection et interactions	**0.989**	**0.988**	**0.988**

Evaluation sur 4 jours (16 images) de 1899 -> 1696 titres boursiers

Utilisation de OCR avec erreur de 0.88%

0% d'erreur visé suite au dernières questions

Résultats quantitatifs

Identification des titres boursiers

Evaluation sur 4 jours (16 images) de 1899
-> 1696 titres boursiers

Vocabulaire utilisé :
ensemble des titres sans les variations dans les intitulés

Utilisation de OCR avec erreur de 0.88%

F-mesure sur titres boursiers
Sans collection : 0,914
Avec collection : 0,988

0% d'erreur visé suite au dernières questions

	Pas de prise en compte de la collection (pages isolées)	Prise en compte de la collection Pas d'interaction	Prise en compte de la collection Après interactions
Nb de vrai positif	1550	1580	**1675**
Nb faux positif	147	117	**13**
Nb faux négatif	146	116	**21**
Précision	0.913	0.931	**0.989**
Rappel	0.914	0.932	**0.988**
F-mesure	0.914	0.931	**0.988**

Conclusion

- Conception d'une stratégie d'analyse
  - Réalisation d'itérations successives
  - Construction d'un OCR adapté
  - Analyse transverse et fiabilisation
  - Applications à des éléments de granularité variable

- Résultats :
  - Moins de 1% pour l'OCR
  - Correction automatique des colonnes
  - Correction automatique des sections
  - Identifications des titres par fiabilisation avec la collection
    - F-mesure : passe de 91,4% à 98.8 % (100% visé)