Historical big-data: modelization of strategies to analyze collections of documents

Camille Guerry

Supervisor : Bertrand Coüasnon, Aurélie Lemaitre, Sébastien Adam

SIFED : 07/06/2019

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

UMR IRISA

intuitive user interaction for document

litis

Introduction

Context of the project

ANR project: HBDEX - Exploitation of Big Data for the Historical Social Sciences application to Financial Data

Global aim: Understanding the long term dynamics of financial markets.

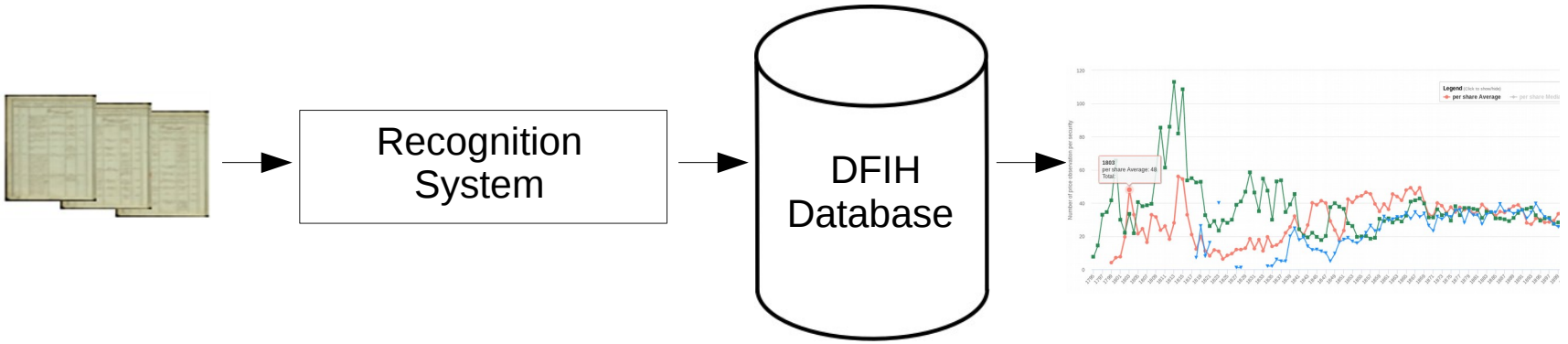
Collaboration with :

- DFIH team of the Paris School of Economics
- LITIS laboratory (Rouen)

Documents processing :

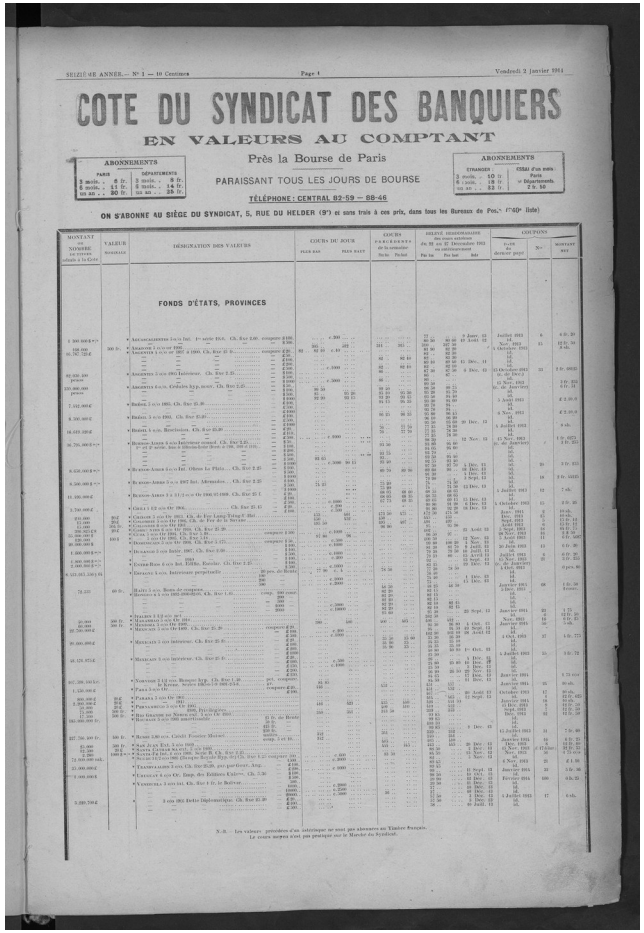
- Strategy for the recognition of collections
- Structure recognition
- Text recognition → LITIS

Global aim of the project



- Extraction of all information contained in the documents
- Limitation of human interaction
- Production of reliable data

The documents



COTE DU SYNDICAT DES BANQUIERS

EN VALEURS AU COMPTANT

Près la Bourse de Paris

PARAISSANT TOUS LES JOURS DE BOURSE

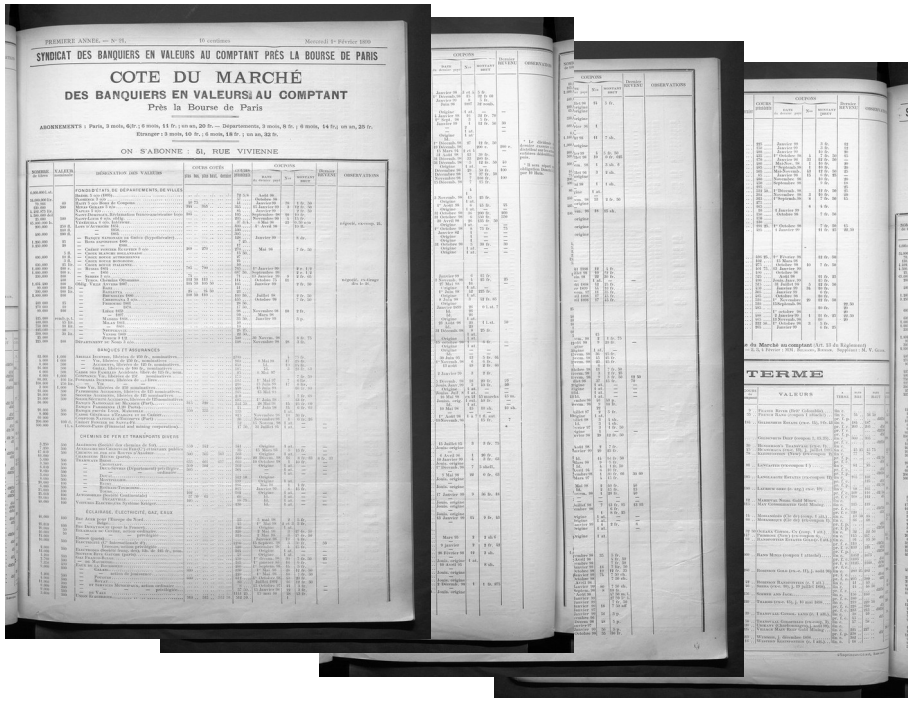
FONDS D'ÉTATS, PROVINCES

- Kind of documents: **Stock exchange daily list**
- In the late 19th and early 20th centuries
- Printed Documents
- Different Markets

Stock exchange daily list:

- List of all stocks
- Tabular structure

La Coulisse (Paris unofficial market)


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

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

ON S'ABONNE : 61, RUE VIVIENNE

TERME

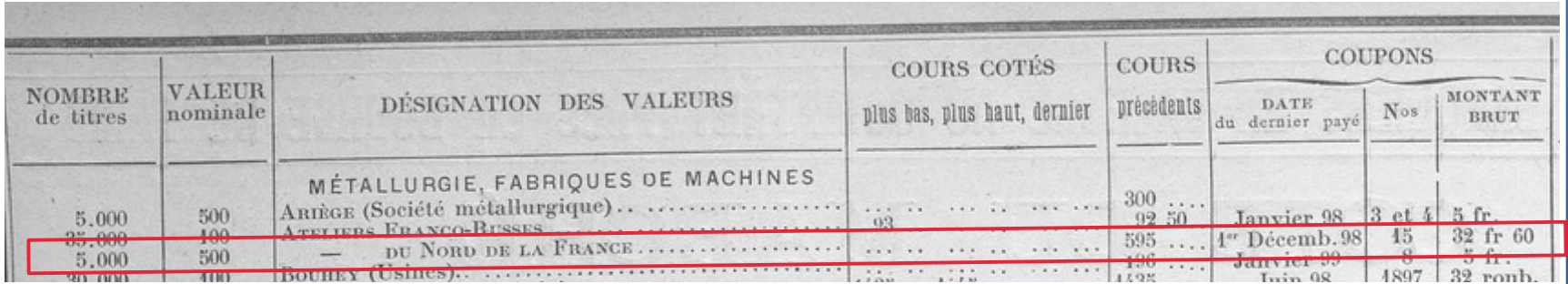
- Sequential data
- 1 day = several pages
- More than 140 000 images
  - 40 years (from 1899 to 1939)
  - between 4 and 15 pages/day

Work done on the Coulisse collection will be generic enough to be adapted to other collections

Presentation Content

Documents characteristics

Extraction example



NOMBRE de titres	VALEUR nominale	DÉSIGNATION DES VALEURS	COURS COTÉS plus bas, plus haut, dernier	COURS précédents	COUPONS DATE du dernier payé	COUPONS Nos	COUPONS MONTANT BRUT
		MÉTALLURGIE, FABRIQUES DE MACHINES					
5.000	500	Ariège (Société métallurgique)..	... .	300			
[illegible]	400	Ateliers Franco-Russes	93	92 50	Janvier 98	3 et 4	5 fr.
5.000	500	— du Nord de la France	... .	595	1er Décemb. 98	15	32 fr 60
[illegible]	400	Bouhey (Usines)..	... .	496	Janvier 99	8	5 fr.
				1435	Juin 98	1897	32 roub.

Stock name : Ateliers du Nord de la France

Sector : MÉTALLURGIE, FABRIQUES DE MACHINES

Stock Quantity :

Value : 5.000 **Currency :**

Stock Nominal :

Value : 500 **Currency :** **Comments :**

Prices :

Min : **Max :** **Close :** **Previous:** 595 **Comments :**

Dividend :

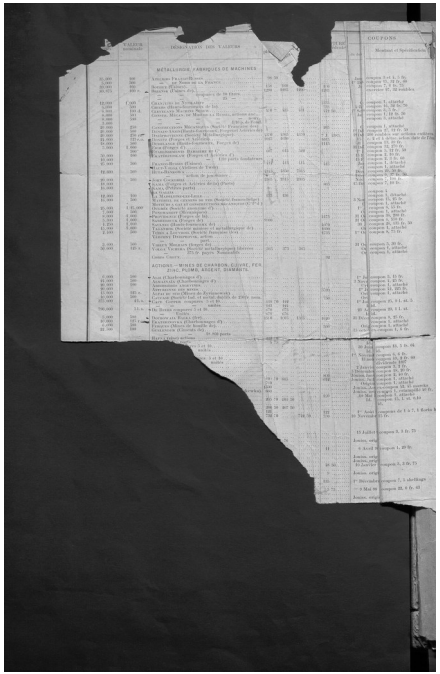
Last payment date : 1898 - 12 - 01 **Last_payment_date_precision :** DAY

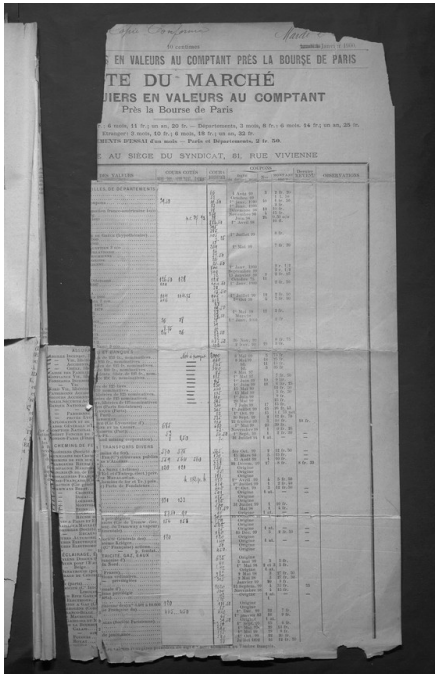
Coupon_number : 15 **Coupon_is_attached :** N **Gross value :** 17.50

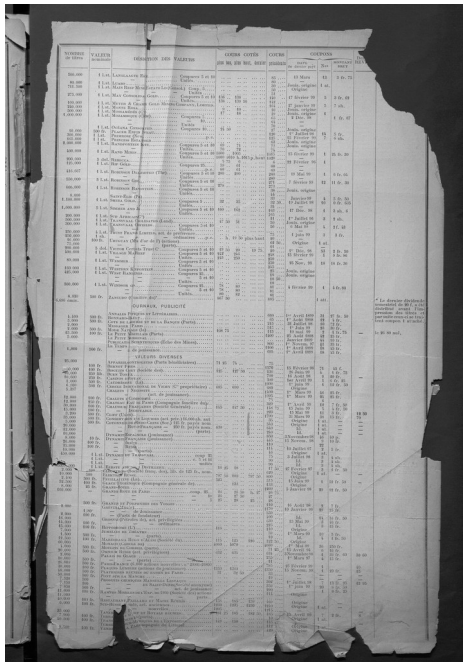
Coupon_number_comments :

Difficulties

- Some documents are strongly deteriorated

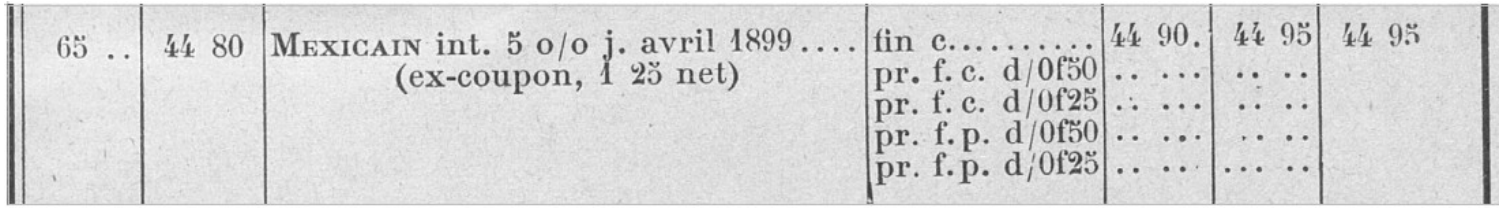






Difficulties

- Some documents are strongly deteriorated
- Complex logical structure



65 ..	44 80	MEXICAIN int. 5 o/o j. avril 1899 .... (ex-coupon, 1 25 net)	fin c.....	44 90.	44 95	44 95
			pr. f. c. d/0f50	.. ...	.. ..	
			pr. f. c. d/0f25	.. ...	.. ..	
			pr. f. p. d/0f50	.. ...	.. ..	
			pr. f. p. d/0f25	.. ..	... ..	

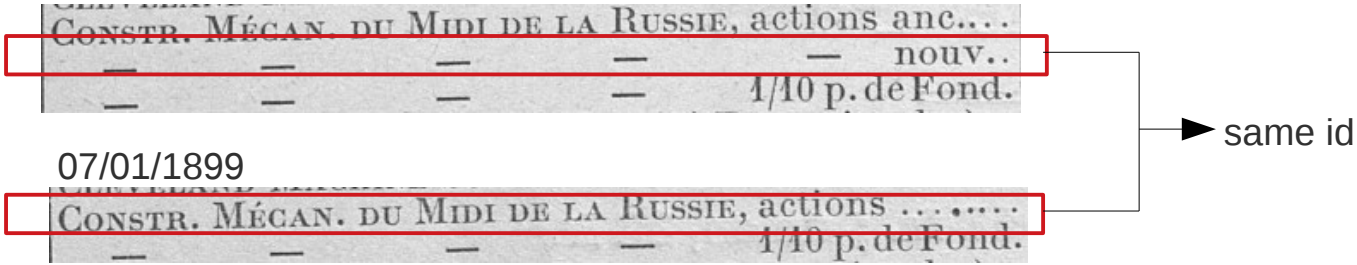
Difficulties

- Some documents are strongly deteriorated
- Complex logical structure
- Stock identification

Different cases:

- different name for the same stock
- close name for different stock
- specificity of an corporate action

Ex : 05/01/1899



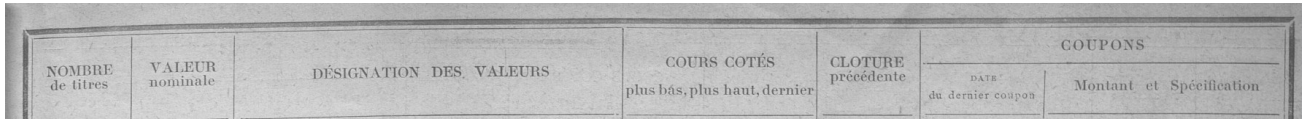
Difficulties

- Some documents are strongly deteriorated
- Complex logical structure
- Stock identification

➡ **Use of the collection characteristics**

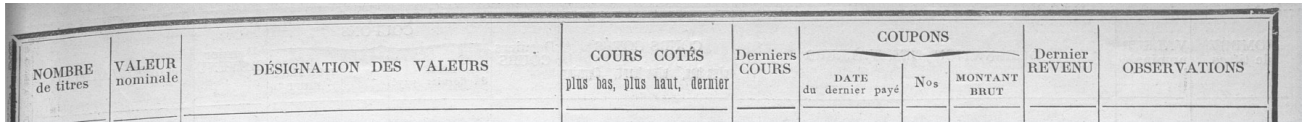
Collection characteristics
Structural stability of column

10/01/1899



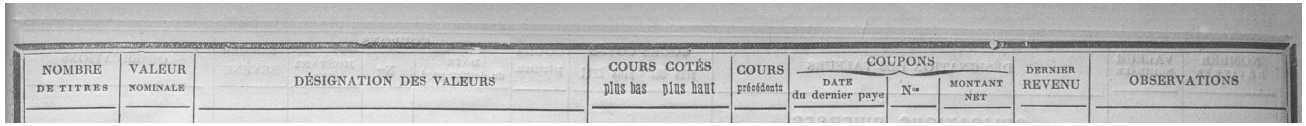
. . .

16/01/1899



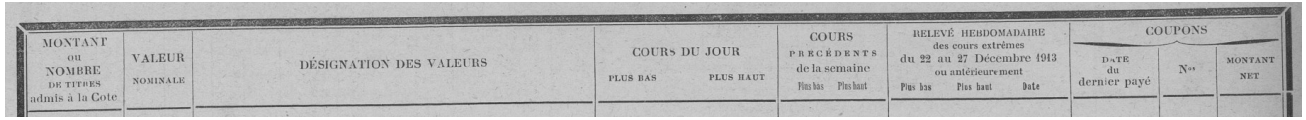
. . .

01/01/1905



. . .

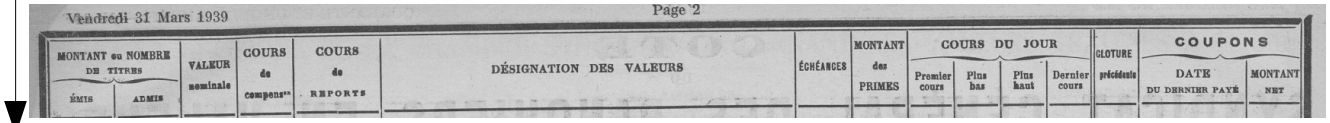
02/01/1914



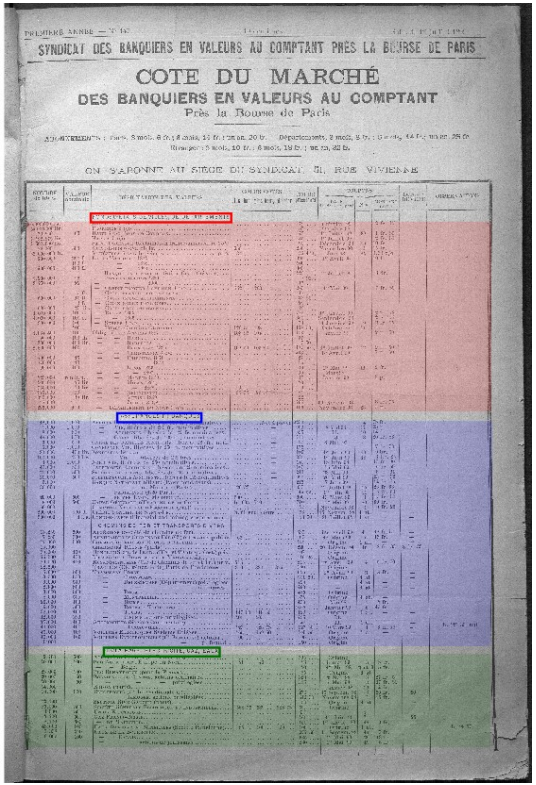
. . .

. . .

28/12/1939



Collection characteristics
Structural stability of sections


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

COTE DU MARCHÉ

DES BANQUIERS EN VALEURS AU COMPTANT

Près la Bourse de Paris

Sections layout

Sections names evolution :

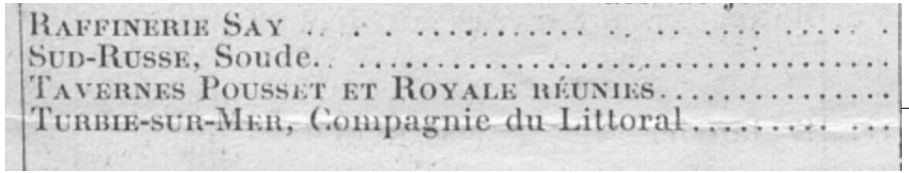
Ordre d'apparition	1899 (07/01/1899)	1905 (21/07/1905)
1	FONDS D'ÉTATS, DE VILLES, DE DÉPARTEMENTS	FONDS D'ÉTATS, VILLES, DÉPARTEMENTS
2	ASSURANCES ET BANQUES	ASSURANCES & BANQUES
3	CHEMINS DE FER ET TRANSPORTS DIVERS	CHEMINS DE FER & TRANSPORTS
4	ÉCLAIRAGE, ÉLECTRICITÉ, GAZ, EAUX	ÉCLAIRAGE, ÉLECTRICITÉ, GAZ, EAUX
5	METALLURGIE, FABRIQUES DE MACHINES	INDUSTRIES TEXTILES, SOIES ARTIFICIELLES, TEINTURERIES
6	ACTIONS.— MINES DE CHARBON, CUIVRE, FER, ZINC, PLOMB, ARGENT, DIAMANTS.	MÉTALLURGIE, CONSTRONS MÉCANIQUES
7	MINES D'OR ET SOCIÉTÉS D'EXPLORATIONS	MINES DE CHARBONS, CUIVRE, FER, ZINC, PLOMB, ARGENT DIAMANTS
. . .	. . .	. . .

Collection characteristics
Structural stability of stock

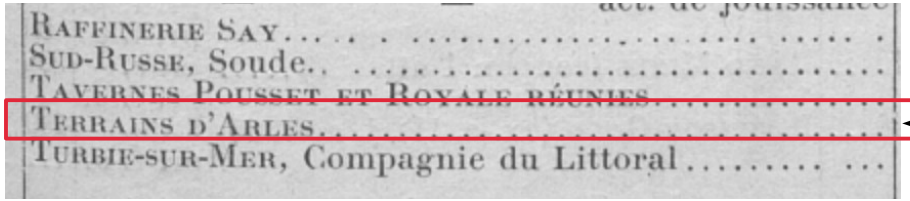
16/02/1899

. . .

21/02/1899



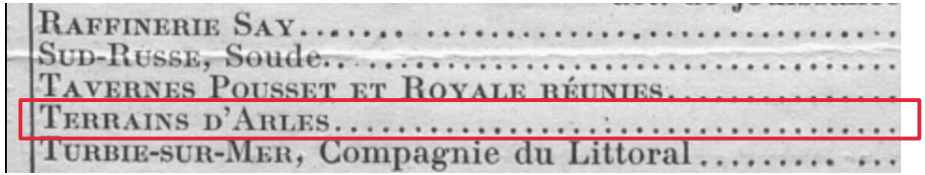
23/02/1899



"Terrain d'Arles" is added

. . .

28/02/1899



Difficulties and Challenges

Difficulties

- Some documents are strongly deteriorated
- Complex logical structure
- Stock identification

Challenges

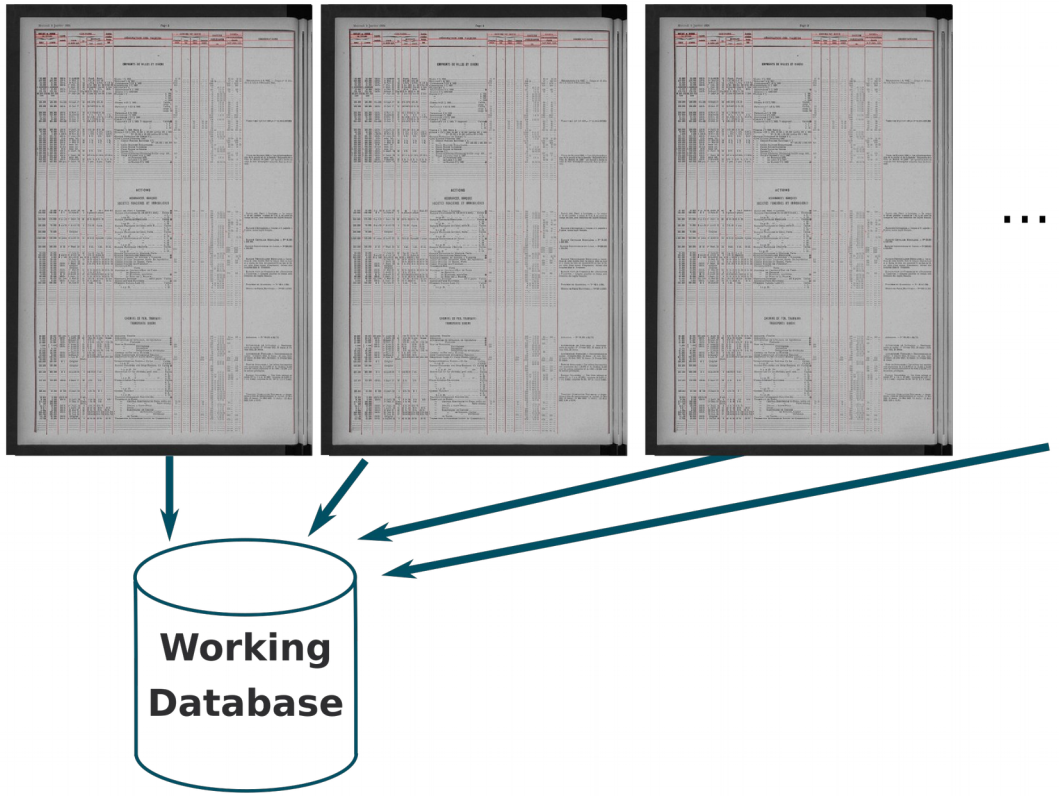
- Use the structural stability of the collection and take the evolution into account
- Use economical rules that apply to the collection
- Propose a strategy usable for other collections

Collection recognition strategy

Iterative process

Structural Analysis

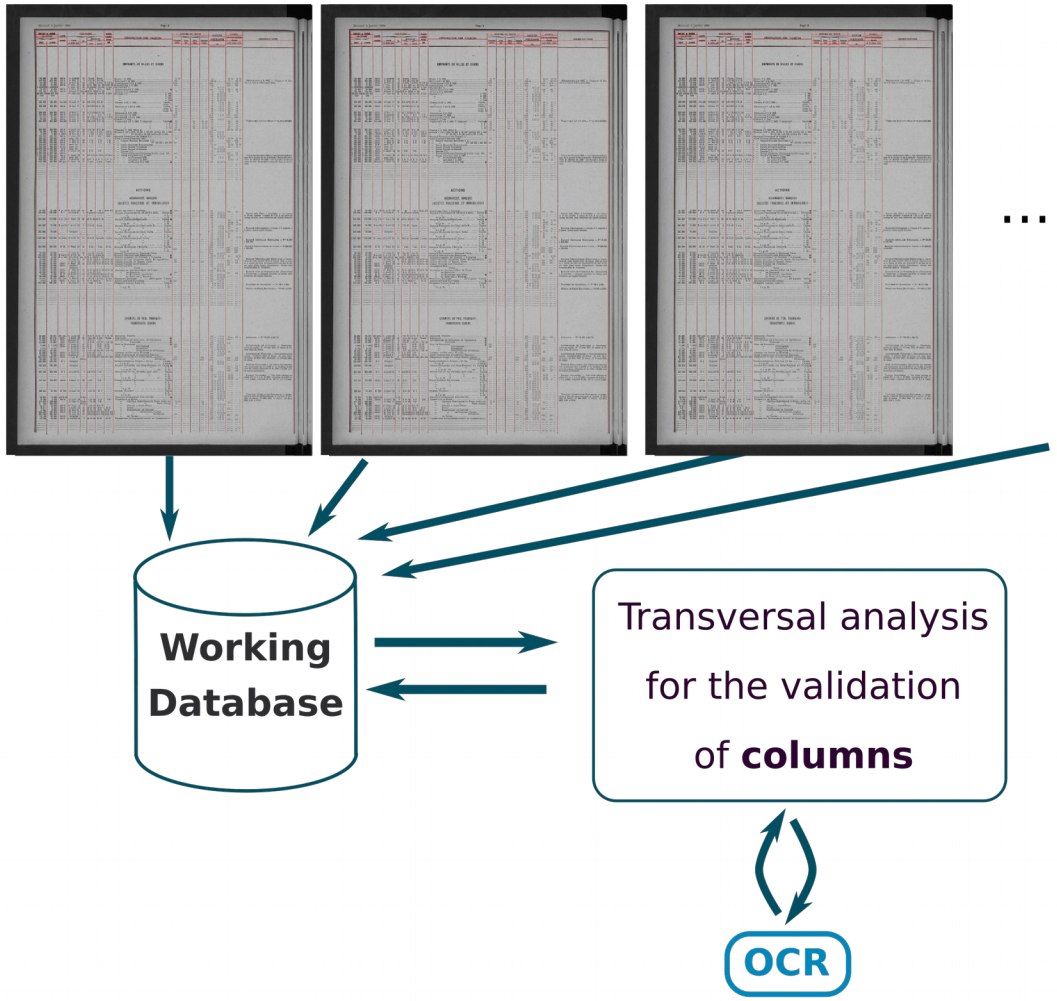
Columns



Iterative process

Structural Analysis

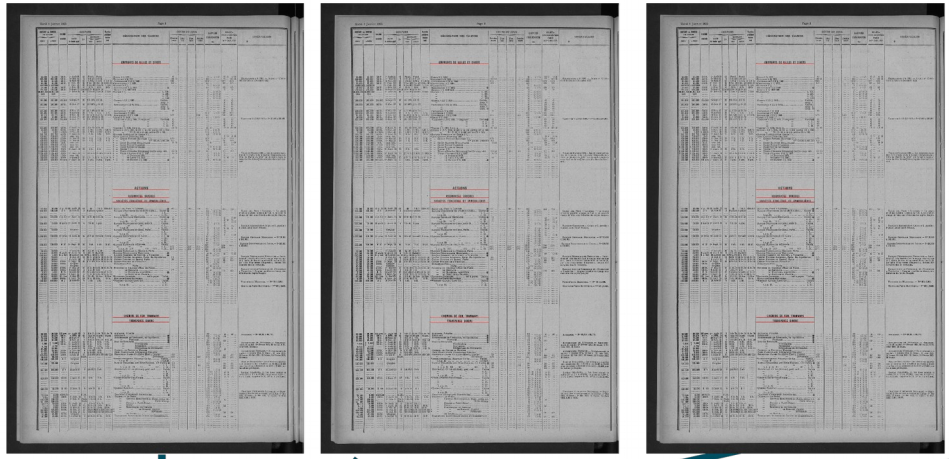
Columns



Iterative process

Structural Analysis

- Columns (reliable)
- Sections



...

Working Database

Transversal analysis for the validation of **sections**

OCR

Iterative process

Structural Analysis

- Columns (reliable)
- Sections (reliable)
- Stocks

...

Working Database

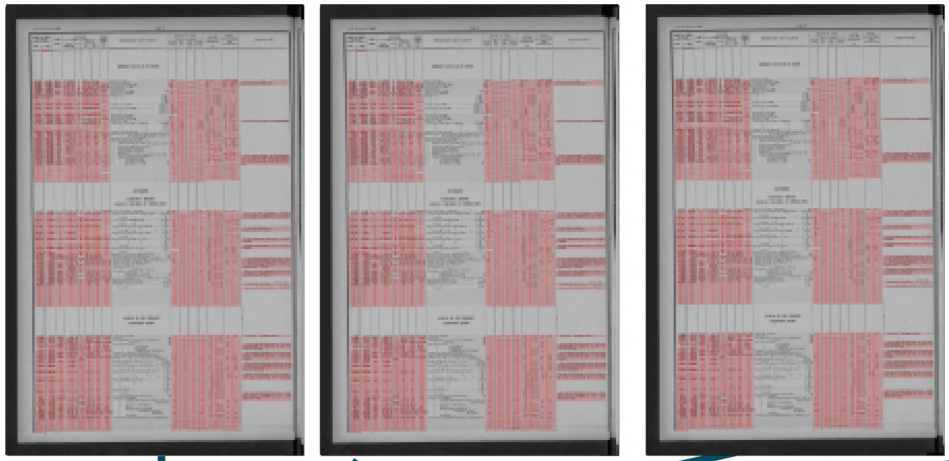
Transversal analysis for the validation of **stocks**

OCR

Iterative process

Structural Analysis

- Columns (reliable)
- Sections (reliable)
- Stocks (reliable)
- Other values



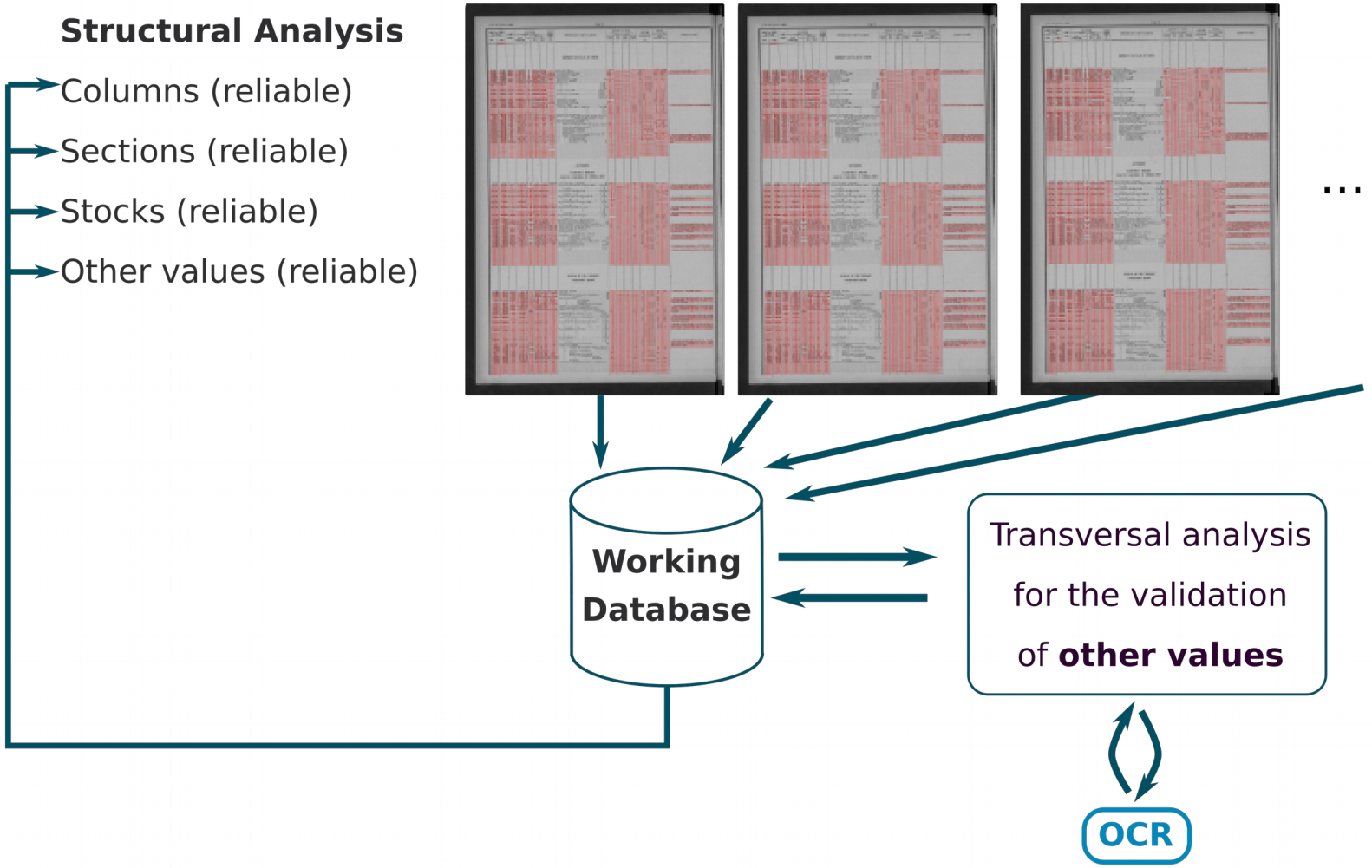
...

Working Database

Transversal analysis for the validation of **other values**

OCR

Iterative process

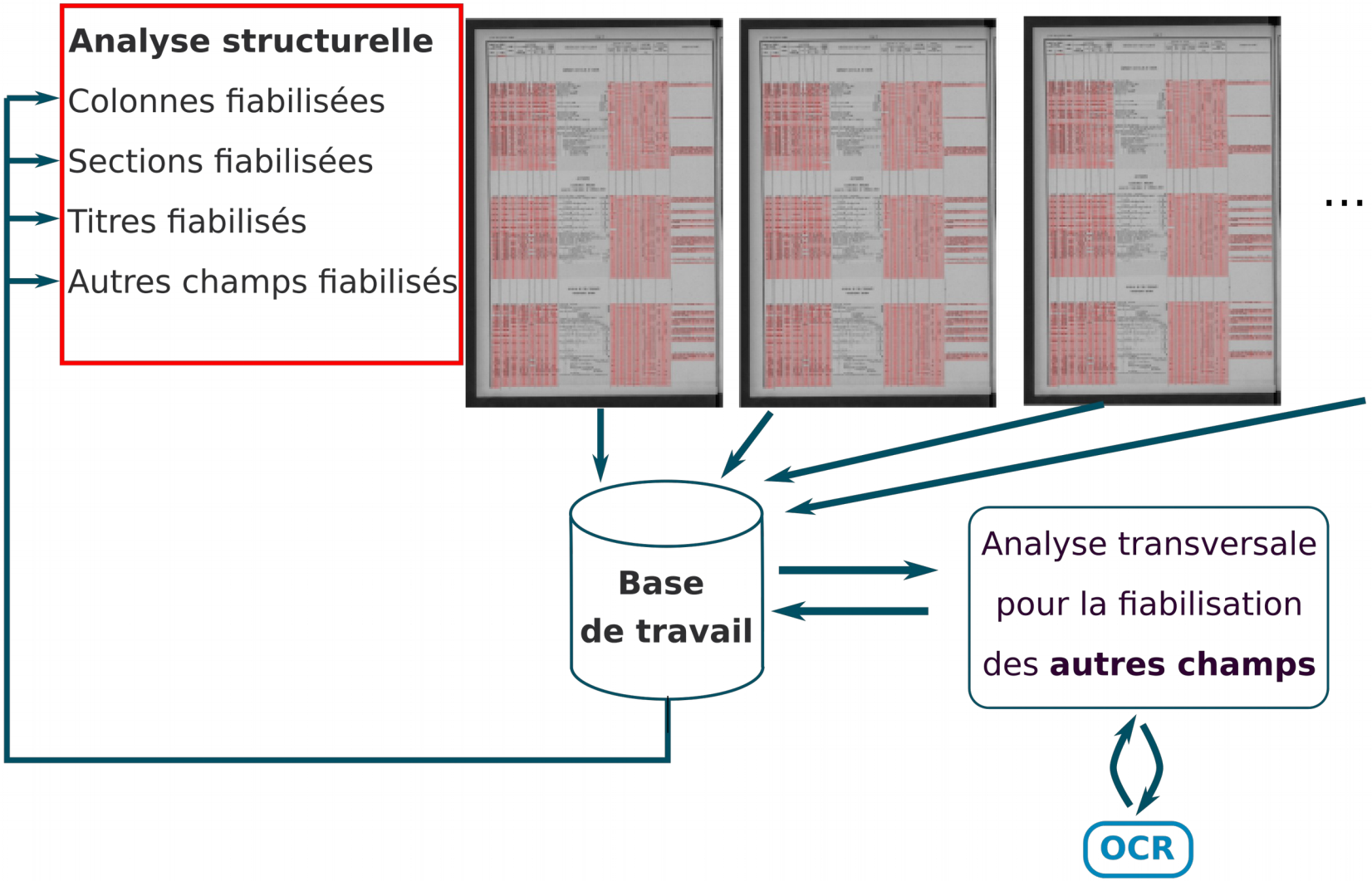


First implementation

-

Analysis of individual pages

Intégration dans la stratégie globale



Purpose and Principle

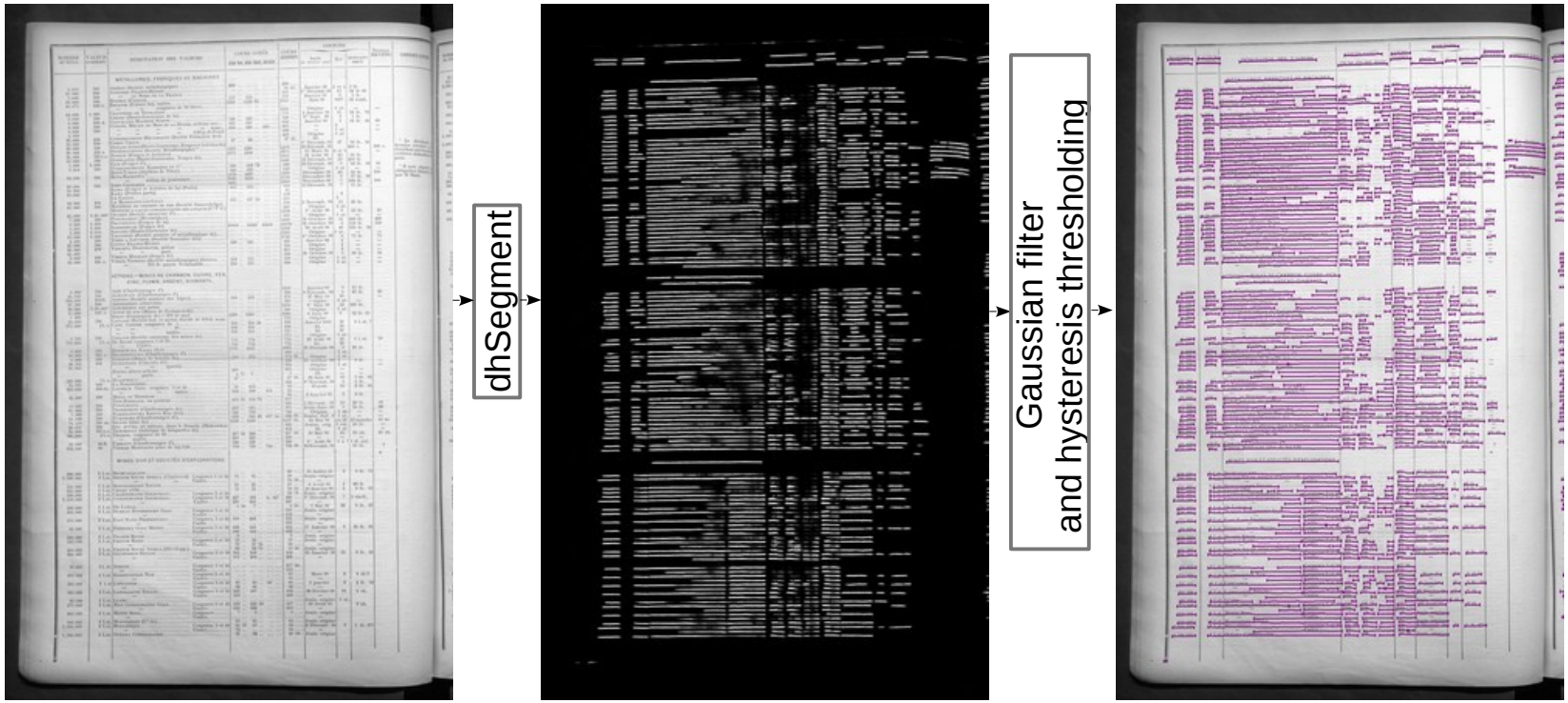
Purpose:

- First step of the global strategy
- Data generation for OCR training
- Study of the vocabulary contained in the documents

Principle: Combination of a deep learning and a syntactical approach

- Text-lines extraction (deep learning part)
- Structural description of the documents (syntactical part)

Deep Learning for text-lines extraction
dhSgment [S. A. Oliveira (2018)]



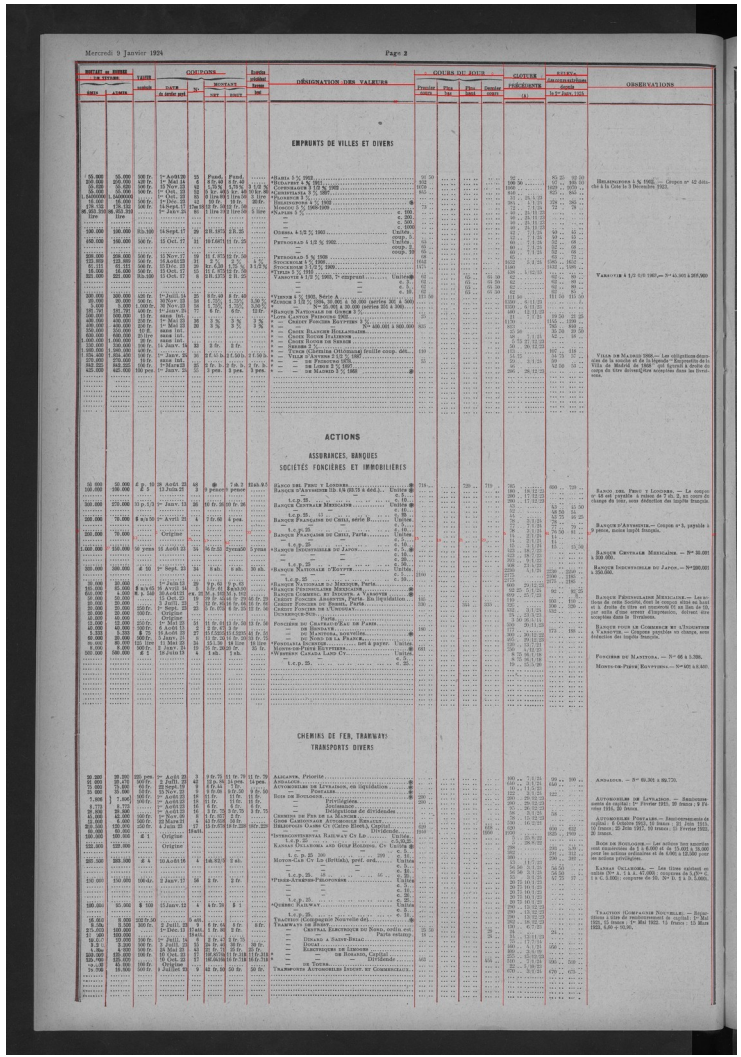
- Prediction of the classes of each pixel thanks to a CNN
- Simple post-processing for the extraction of text-lines

Grammatical Description

structure ::=

banniereDeTitre Banniere EnsColonnes *&&*
lesDonnees EnsLignes.

...

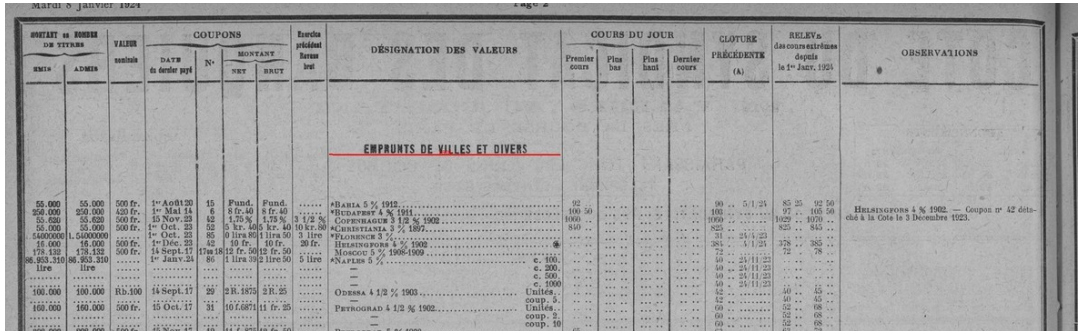


Grammatical Description

structure ::=
bannièreDeTitre Banniere EnsColonnes &&
lesDonnees EnsLignes.

lesDonnees ::=
lesSections EnsSection.

lesSections [Section | R] ::=
uneSection Section &&
lesSections R.



Mardi 8 janvier 1924 — Page 2

NOMBRE DE TITRES émis	NOMBRE DE TITRES admis	VALEUR nominale	COUPONS Date du dernier payé	COUPONS N°	COUPONS Montant net	COUPONS Montant brut	Exercice précédent Revenu brut	DÉSIGNATION DES VALEURS	COURS DU JOUR Premier cours	COURS DU JOUR Plus bas	COURS DU JOUR Plus haut	COURS DU JOUR Dernier cours	CLÔTURE PRÉCÉDENTE (A)	RELEVé des cours extrêmes depuis le 1er Janv. 1924	OBSERVATIONS
								EMPRUNTS DE VILLES ET DIVERS							
55.000	55.000	500 fr.	1er Août 20	15	Fund.	Fund.	.....	Bahia 5 % 1912	92				90 .. 5/1/24	85 25 92 50	
250.000	250.000	450 fr.	1er Mai 14	6	8 fr. 40	8 fr. 40	.....	Budapest 4 % 1911	100 50	....	....	....	101	97 .. 105 50	
55.620	55.620	500 fr.	15 Nov. 23	42	1,75 %	1,75 %	3 1/2 %	Copenhague 3 1/2 % 1902	1060 ..	....	....	....	1060	1029 .. 1070 ..	
55.000	55.000	500 fr.	1er Oct. 23	52	5 kr. 40	5 kr. 40	10 kr. 80	Christiania 3 % 1897	840 ..	....	....	....	825	825 .. 845 ..	
54000000	1.54000000	.....	1er Oct. 23	85	0 lira 80	1 lira 50	3 lire	Florence 3 %	....	....	....	....	31 .. 24/1/23	.....	
16.000	16.000	500 fr.	1er Déc. 23	42	10 fr.	10 fr.	20 fr.	Helsingfors 4 % 1902	✱	....	....	....	384 .. 4/1/24	378 .. 385 ..	
178.132	178.132	500 fr.	14 Sept. 17	17 bis	12 fr. 50	12 fr. 50	.....	Moscou 5 % 1908-1909	....	....	....	....	72	72 .. 78 ..	
86.953.310 lire	86.953.310 lire	.....	1er Janv. 24	86	1 lira 39	2 lire 50	5 lire	Naples 5 %	....	....	....	....	50 .. 24/11/23	.....	
.....	.....	.....	.....	.....	.....	.....	.....	—	c. 100.	....	....	....	50 .. 24/11/23	.....	
.....	.....	.....	.....	.....	.....	.....	.....	—	c. 200.	....	....	....	50 .. 24/11/23	.....	
.....	.....	.....	.....	.....	.....	.....	.....	—	c. 500.	....	....	....	50 .. 24/11/23	.....	
.....	.....	.....	.....	.....	.....	.....	.....	—	c. 1000	....	....	....	50 .. 24/11/23	.....	
100.000	100.000	Rb. 100	14 Sept. 17	29	2 R. 1875	2 R. 25	.....	Odessa 4 1/2 % 1903	Unités.	....	....	....	42	40 .. 55 ..	
.....	.....	.....	.....	.....	.....	.....	.....	—	coup. 5.	....	....	....	42	40 .. 45 ..	
160.000	160.000	500 fr.	15 Oct. 17	31	10 f. 6871	11 fr. 25	.....	Petrograd 4 1/2 % 1902	Unités.	....	....	....	60	52 .. 68 ..	
.....	.....	.....	.....	.....	.....	.....	.....	—	coup. 5.	....	....	....	60	52 .. 68 ..	
.....	.....	.....	.....	.....	.....	.....	.....	—	coup. 10	....	....	....	60	52 .. 68 ..	

Helsingfors 4 % 1902. — Coupon n° 42 détaché à la Cote le 3 Décembre 1923.

uneSection Section ::=
***intituleDeSection UnIntitule** &&*
titreBoursier EnsTitreBoursier&&
consSection UnIntitule EnsTitreBoursier Section.

...

Grammatical Description

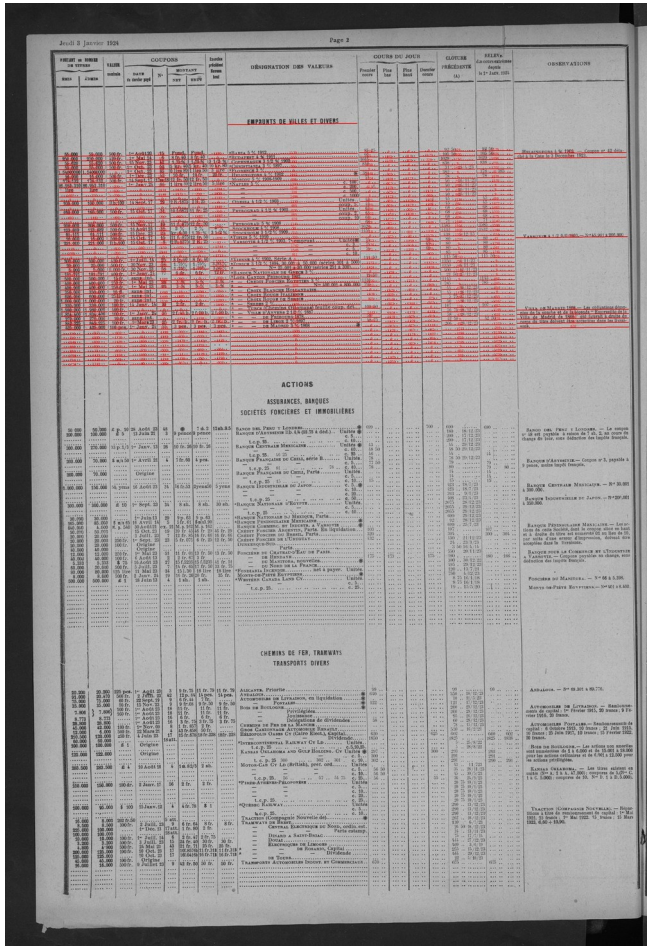
structure ::=
banniereDeTitre Banniere EnsColonnes &&
lesDonnees EnsLignes.

lesDonnees ::=
lesSections EnsSection.

lesSections [Section | R] ::=
***uneSection Section** &&*
lesSections R.

uneSection Section ::=
intituleDeSection UnIntitule &&
titresBoursiers EnsTitreBoursier&&
consSection UnIntitule EnsTitreBoursier Section.

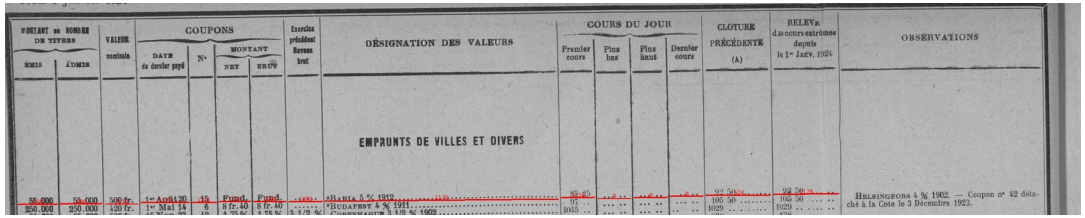
...



Grammatical Description

structure ::=
bannièreDeTitre Banniere EnsColonnes &&
lesDonnees EnsLignes.

lesDonnees ::=
lesSections EnsSection.



lesSections [Section | R] ::=
uneSection Section &&
lesSections R.

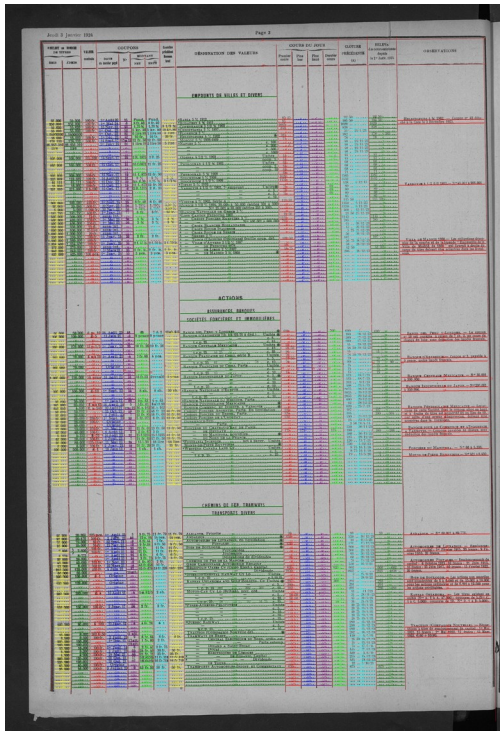
uneSection Section ::=
intituleDeSection UnIntitule &&
titresBoursiers EnsTitreBoursier&&
consSection UnIntitule EnsTitreBoursier Section.

titresBoursiers [unTitre|R]::=
designationDuTitre DesignationTitre &&
champsColonnesDroite EnsChamps1 &&
champsColonnesGauche EnsChamps2 &&
consTitre DesignationTitre EnsChamps1 EnsChamps2 unTitre&&
titresBoursiers R.

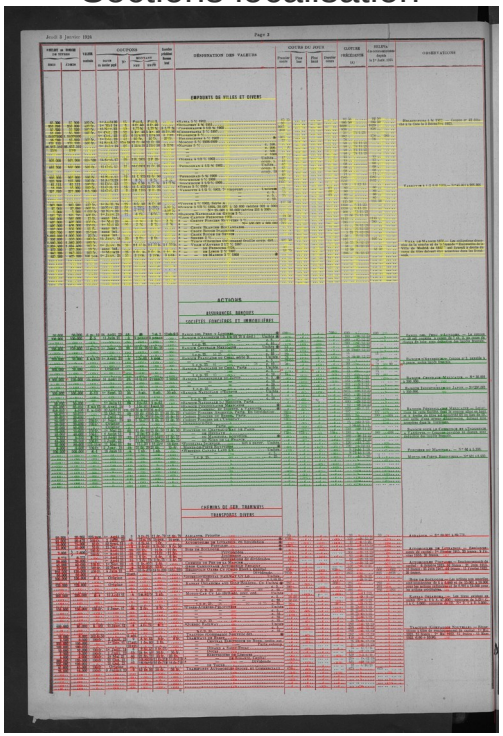
...

Structural Recognition

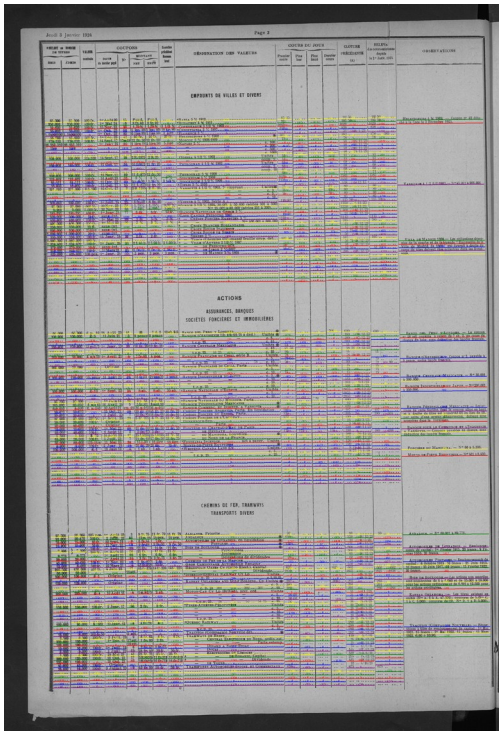
Column localization



Sections localisation



Table's rows



Result's format :
(a text-line)


```
{ "type" : "line",
"polyline":"1032,5175;1039,5175;1052,5173;1066,5173; ...",
 "column" : "2",
 "sectionTitle" :"False" ,
 "sectionNumber" : "2",
 "rowNumber" : "15",
 "title" : "False" }
```

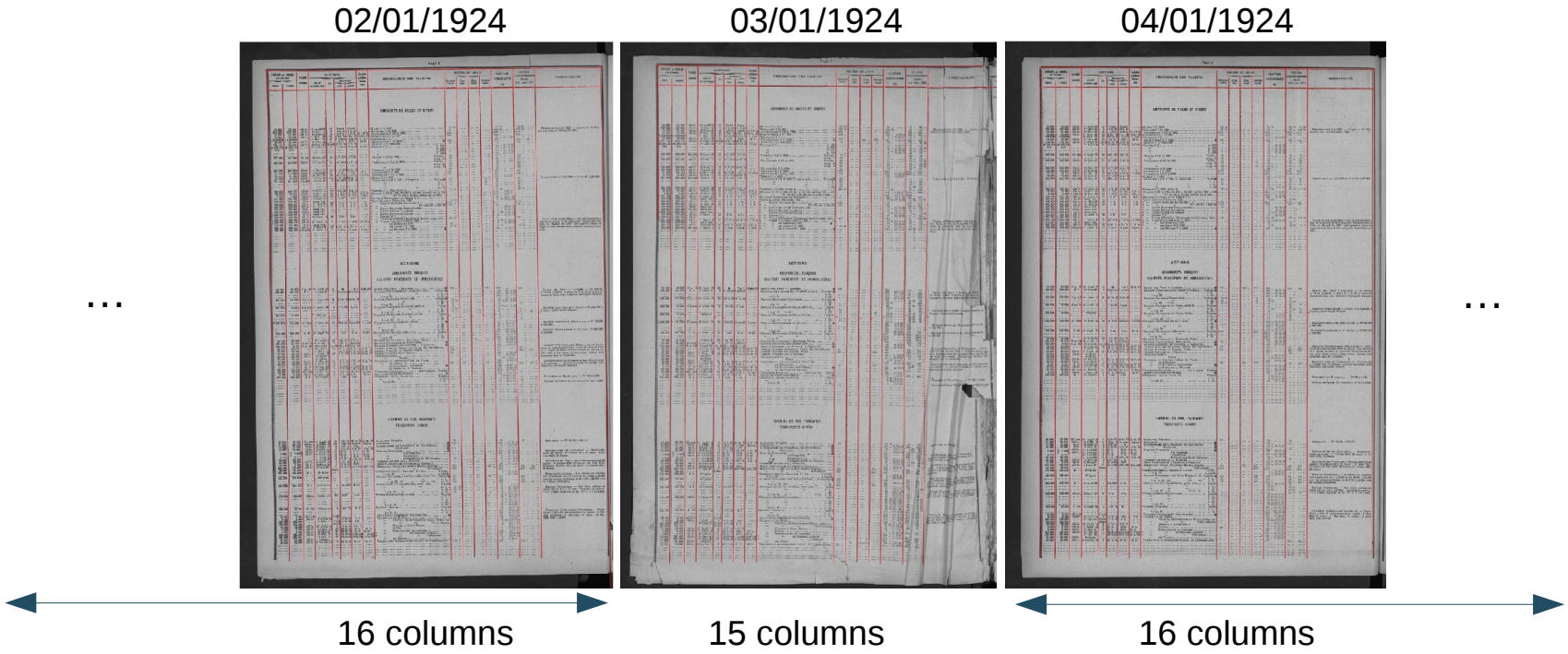
First evaluation : Column Recognition

Période	Du 09/01 au 14/01 1899 (page 2)	Du 14/01 au 31/03 1899 (page 2)	Du 02/01 au 15/04 1924 (page 2)	Du 03/01 au 31/03 1939 (page 2)
Nb d'images	6	64	72	64
Nb d'images avec le bon nb de colonnes détectées	**5**	**58**	**71**	**58**
Nb d'images avec un nb erroné de colonnes détectées	1	6	1	6

Notes :

- Purpose: - having a first idea of the system performances
  - identifying the errors
- Errors will be corrected by the global strategy
- A more precise metric is necessary to correctly access the performances of our system

Transversal analysis to improve columns recognition



- Error detection :
  - → column number
  - → OCR results for the columns names
- Identifying the missing column
- Modifying the structural recognition

Conclusion

Context

- Stock exchange daily list
- A considerable number of documents

Difficulties

- Some documents are strongly deteriorated
- Complex logical structure
- Stock identification

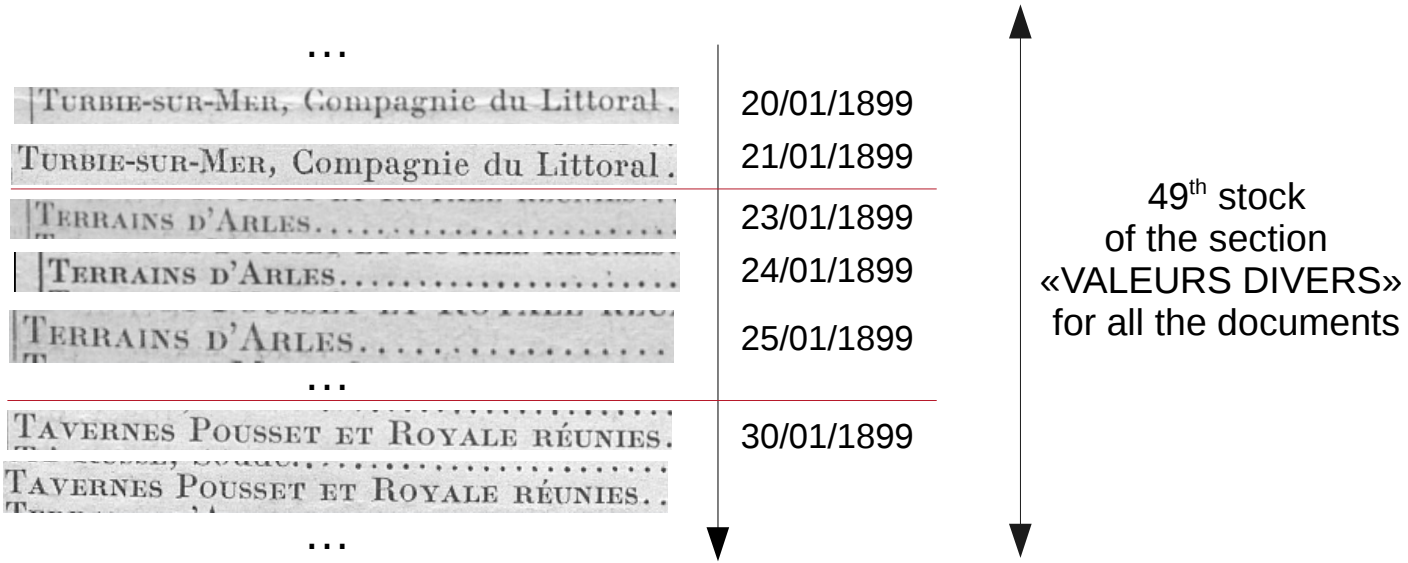
Strategy

- Iterative process
- Traversal analysis to improve the recognition
- A Strategy adaptable to other collection

First implementation: Analysis of individual pages

OCR 1 Specifications

Input



Output

range 1:

- **Interval:** 01/01/1899 - 21/01/1899
- **Transcription Hypothesis:** “Turbie-sur-Mer, Compagnie du Littoral”

range 2 :

- **Interval:** 23/01/1899 – 29/01/1899
- **Transcription Hypothesis:** “Terrains d’Arles”

range 3 :

- **Interval:** 30/01/1899 - …
- **Transcription Hypothesis:** “Tavernes Pousset et Royale Réunies”

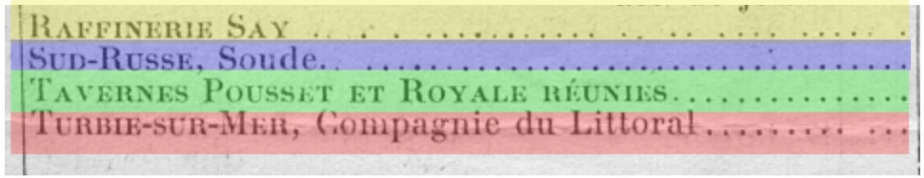
…

Using OCR 1

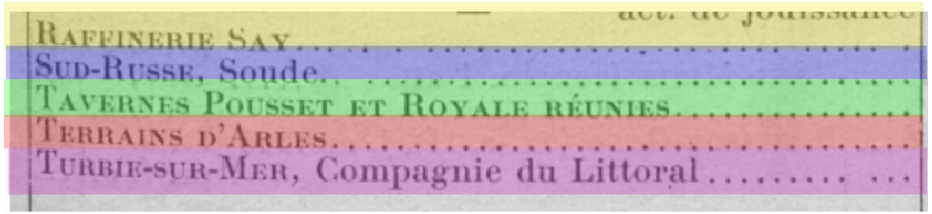
01/01/1899

...

21/01/1899

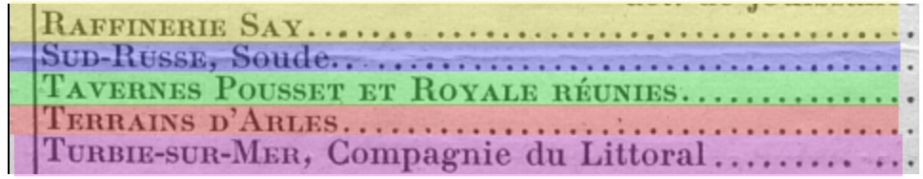


23/01/1899



...

28/01/1899



Transversal analysis

OCR 1: put together same stocks

Detecting potential errors:

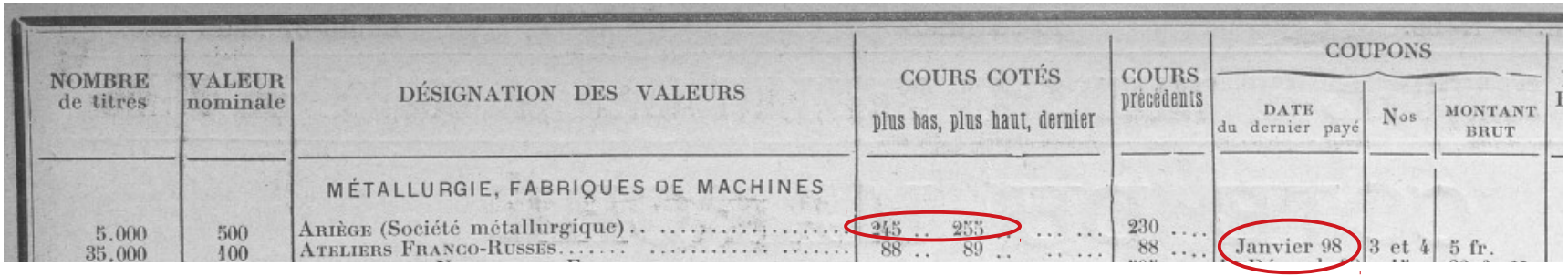
- 1 stock appear just for 1 day
- 1 stock is missing just for 1 day
- 1 stock disappear and reappear during a short period

Processing:

- Ask for a confidence score (OCR 2)
- Take a decision during the next iteration
- Ask an human operator

Collection characteristics

1) Rules for a same day



NOMBRE de titres	VALEUR nominale	DÉSIGNATION DES VALEURS	COURS COTÉS plus bas, plus haut, dernier	COURS précédents	COUPONS DATE du dernier payé	COUPONS Nos	COUPONS MONTANT BRUT
		MÉTALLURGIE, FABRIQUES DE MACHINES					
5.000	500	Ariège (Société métallurgique)	245 .. 255	230			
35.000	100	Ateliers Franco-Russes	88 .. 89	88	Janvier 98	3 et 4	5 fr.

- ***Interdependence between values***

  ex : 'Cours plus bas' ⩽ 'Cours plus haut'

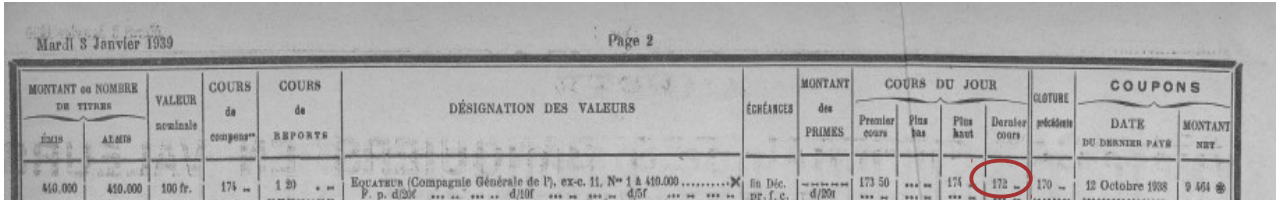
- ***Take the date into account***

  ex : 'Date du dernier payé' ⩽ 'Date of the day'

Collection characteristics

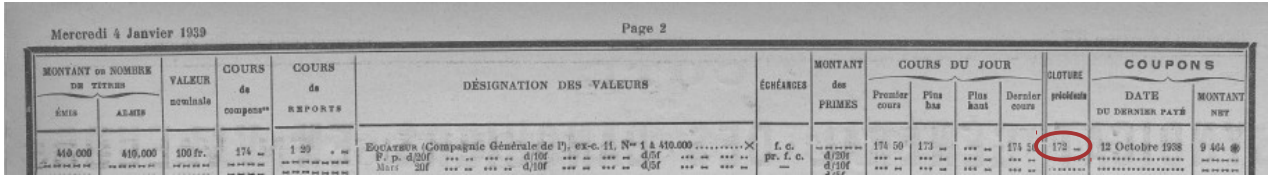
2) Rules between days

03/01/1939


Mardi 3 Janvier 1939 — Page 2

MONTANT ou NOMBRE DE TITRES ÉMIS	MONTANT ou NOMBRE DE TITRES ALMIS	VALEUR nominale	COURS de compens^on	COURS de REPORTS	DÉSIGNATION DES VALEURS	ÉCHÉANCES	MONTANT des PRIMES	COURS DU JOUR Premier cours	Plus bas	Plus haut	Dernier cours	CLOTURE précédente	COUPONS DATE DU DERNIER PAYÉ	COUPONS MONTANT NET
410.000	410.000	100 fr.	174 ..	1 20 ..	Equateur (Compagnie Générale de l'), ex-c. 11, N^os 1 à 410.000	fin Déc.	d/20f	173 50	... ..	174 ..	172 ..	170 ..	12 Octobre 1938	9 464

04/01/1939


Mercredi 4 Janvier 1939 — Page 2

MONTANT ou NOMBRE DE TITRES ÉMIS	MONTANT ou NOMBRE DE TITRES ALMIS	VALEUR nominale	COURS de compens^on	COURS de REPORTS	DÉSIGNATION DES VALEURS	ÉCHÉANCES	MONTANT des PRIMES	COURS DU JOUR Premier cours	Plus bas	Plus haut	Dernier cours	CLOTURE précédente	COUPONS DATE DU DERNIER PAYÉ	COUPONS MONTANT NET
410.000	410.000	100 fr.	174 ..	1 20 ..	Equateur (Compagnie Générale de l'), ex-c. 11, N^os 1 à 410.000	f. c. pr. f. c.	d/20f	174 50	173 ..	... ..	174 50	172 ..	12 Octobre 1938	9 464

- ***Interdependence between values***

ex : *'Dernière Clôture'* is equal to '*Dernier COURS'* of previous day

- **Stable Values**

ex: 'VALEUR NOMINALE'

- **Change at a precise date**

ex : 'COURS de Compensation'