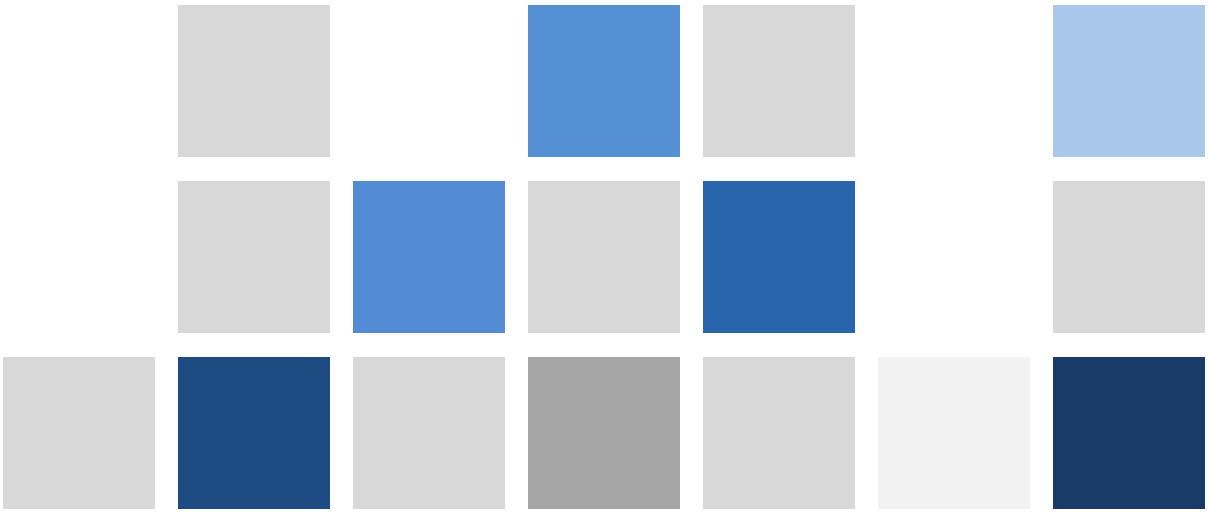
EurHisFirm M7.1: First version of the data extraction system

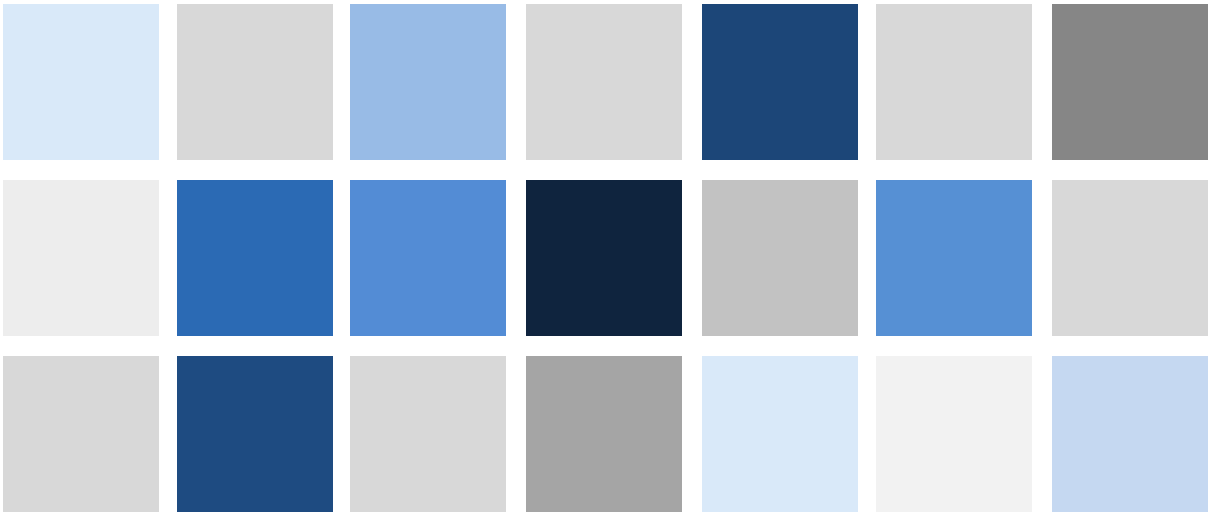
Sébastien Adam, Simon Bouvier, Bertrand B. Coüasnon, Camille Guerry, Aurélie Lemaitre, Iwan Le Floch, Thierry Paquet, Andres Rojas Camacho, Achille Fedioun, Wassim Swaileh



Long-term data for Europe

EURHISFIRM

M7.1: First version of the data extraction system



This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

Deliverable	**M7.1: First version of the data extraction system**
Due Date of Deliverable	**Month 24, 31/03/2020**
Work Package	**WP7: Data extraction and enrichment system**
Tasks	**T7.2, T7.3**
Type	**Software Libraries**

AUTHORS:

Sébastien ADAM (UNIVERSITÉ DE ROUEN NORMANDIE)
Simon BOUVIER (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Bertrand COÜASNON (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Camille GUERRY (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Aurélie LEMAITRE (UNIVERSITÉ RENNES 2, INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Iwan LE FLOCH (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Thierry PAQUET (UNIVERSITÉ DE ROUEN NORMANDIE)
Andres ROJAS CAMACHO (UNIVERSITÉ DE ROUEN NORMANDIE)
Achille FEDIOUN (UNIVERSITÉ DE ROUEN NORMANDIE)
Wassim SWAILEH (UNIVERSITÉ DE ROUEN NORMANDIE)

APPROVED IN 2020 BY:

Jan ANNAERT *(Universiteit Antwerpen)*

Wolfgang KÖNIG *(Goethe-Universität Frankfurt)*

Angelo RIVA *(École d'Économie de Paris)*

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

Table of Contents

This project has received funding from

the European Union’s Horizon 2020 research and innovation programme
under grant agreement N° 777489

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

1 Introduction

Work Package 7 develops an intelligent and collaborative system for the extraction of structured information from images of historical documents related to companies' financial and economic activities. The system focuses on printed serial sources related to listed companies such as stock exchange yearbooks and price lists. In the deliverable D7.1, we provided general software libraries, which can be used to build different prototypes of document recognition and understanding systems adapted to different kinds of documents. This milestone M7.1 is composed of the first version of two recognition systems: one for yearbook information extraction, and one for price list data extraction. Those systems have been applied to several corpora: the German Yearbook 1913-1914 (Handbuch der deutschen Aktiengesellschaften), the French Desfossés Yearbook 1962, the official price lists for Brussels 1912 and Paris 1961-1962, which are part of the document samples dataset validated by the Steering Committee. This document samples dataset is made of three yearbooks, three stock price lists, with three different languages, on three time periods: before WWI, interwar and post WWII. The two remaining corpora are the Spanish yearbook 1929-1930 and the official price lists for Madrid 1934.

Details on the evaluation of those two systems have been presented in the deliverable D7.2.

2 Yearbook Information Extraction system

A generic pipeline of processes that can run similarly on the various Yearbooks that are considered within the consortium has been implemented (see *Figure 1* below). Inputs of the pipeline are images of documents and outputs are information attached to each company the Yearbook is reporting on. The information that is to be extracted from the yearbooks is structured in rubrics composed of lists of named entities (i.e. list of person names, as is the case of the "governing board" rubric), or lists of linked named entities (i.e. [date, amount, currency] as is the case of the "capital" rubric). This pipeline (see Figure 1) is composed of optical character recognition (OCR) followed by layout analysis including table detection and recognition (IRISA), followed by text analysis for named entities extraction (LITIS). For the experiments conducted so far, a general-purpose industrial OCR was used and proved to give sufficiently good results so that LITIS and IRISA mostly concentrated on the extraction process of rubrics and tables (IRISA), and linked named entities extraction in yearbooks (LITIS).

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

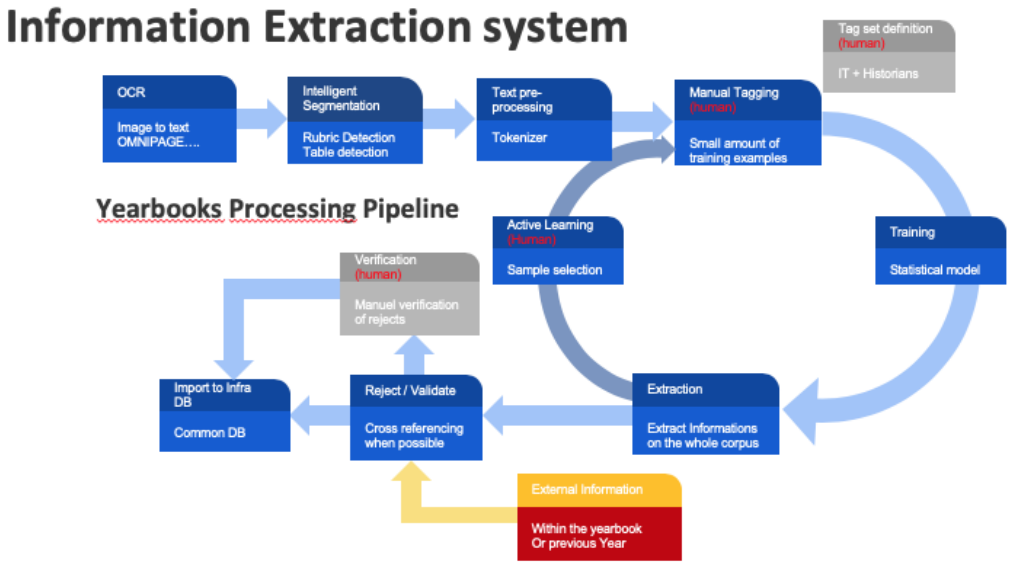


Figure 1: Overview of the generic Information extraction pipeline in yearbooks.

2.1 Text Blocks Selection by Document Structure Recognition (Task 7.2)

2.1.1 System

In this section, we first present the improvement of the component detectors provided in D7.1, then we detail the system for the structural analysis of yearbooks. We focus on the specific cases of tables and balance sheets.

A. Component detectors and tools

In D7.1 we introduced a library containing tools to extract the structure of different documents found in financial yearbooks and price lists.

This library is shared with a national French economics project (ANR HBDEX - Exploitation of Big Historical Data for the Digital Humanities: application to financial data) and includes tools such as:

- recognition of table rulings;
- recognition of table separators without rulings: logical separators of columns and rows in tables which do not contain physical rulings;
- localization of text lines;
- contextual segmentation of text lines: reconstruction of fragmented text lines inside a same column and segmentation of these lines depending on the tabular structure;
- connection with a commercial OCR.

Since D7.1, these tools have been enriched and improved to be more efficient and more flexible. One notable addition is a new grammar dedicated to line alignments detection, which is particularly useful to modelize columns in tables with no rulings (Figure 2).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

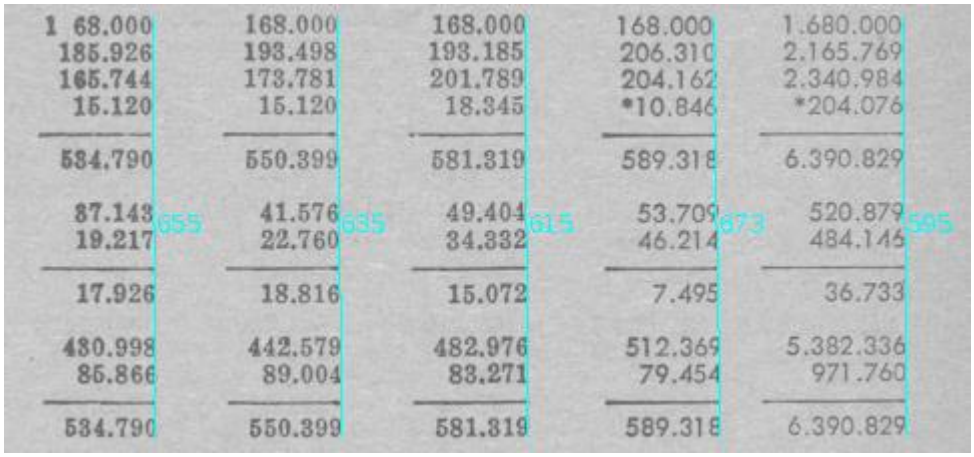


Figure 2: Alignments can be used as separators between columns

Another major change is the use of ARU-Net [T. Grüning 2018] instead of dhSegment for text-lines detection. ARU-Net gives slightly better results on cBAD and on our corpuses. We trained this new deep learning-based system on documents from French and German yearbooks to further improve its efficiency on printed sources. Detected text-lines are now more precise, which gives better results when detecting alignments.

The same neural network was trained to detect rubric titles in the German yearbook and gives really good results (Figure 3).

Bilanz am 31. Dez. 1913: Aktiva: Kto der Aktionäre 100 000, Kassa 40 951, Debit. 532 995, Bank- u. Inkasso 54 758, Wechsel 177 740, Hypoth. 47 833, Effekten 104 909, Mobil. 283. — Passiva: A.-K. 500 000, Mehrzahl. auf Aktien 616, Spar- u. Depositenkto 210 797, Kredit. 196 782, Div. 24 000, Tant. 2560, Rückl. für Talonsteuer 2000, Delkr.-Kto 9038, R.-F. 52 847, Vortrag 829. Sa. M. 999 472.

Gewinn- u. Verlust-Konto: Debet: Kursverlust auf Effekten 1455, Abschreib. 31, Handl.-Unk. 9096, Div. 24 000, Tant. 2560, Talonsteuer 912, Delkr.-Kto 2000, R.-F. 2000, Vortrag 829. — Kredit: Zs. 36 323, Provis. 6164, Verschiedenes 387, Kursgewinn 9. Sa. M. 42 884.

Dividenden: 1897: M. 17 pro Aktie p. r. t.; 1898—1913: 4, 4½, 5, 5½, 5½, 6, 6, 6, 6, 6, 6½, 6½, 6, 5½, 5½, 6% Coup.-Verj.: 4 J. (F.)

Vorstand: Franz Kügelgen, Joh. Geusgen.

Aufsichtsrat: (6) Vors. Jos. Berk, Neu-Hemmerich; Gottfr. Hendrickx, Frechen; Jos. Felten, Bachem; Carl Baumann, Haus Vorst; Fabrikant G. Dorn, Frechen.

Figure 3: Rubric titles detection with ARU-Net (Handbuch 1914-15, page 105)

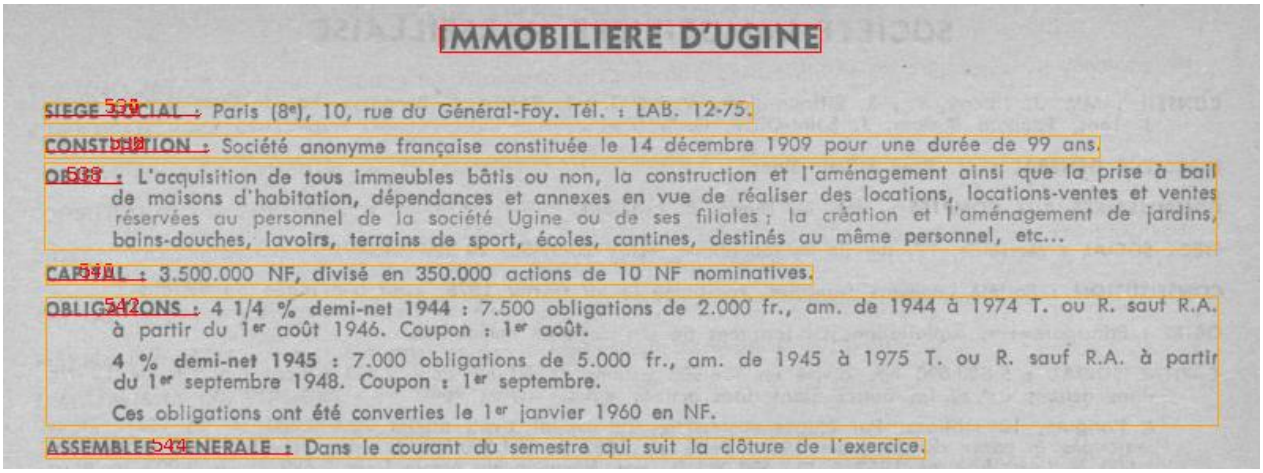
B. Structural analysis of yearbooks

These tools are the bricks of more complex systems describing entire pages of the yearbooks or price lists.

Yearbooks from different origins are built following the same structure: a main title containing the name of the issuer, followed by several rubrics composed of a title and a content.
In the French Desfossés yearbook (year 1962) for instance, we can identify rubric titles with capitals and semicolon localization. The other lines of the same rubric are always indented to the right and all part of the same alignment (Figure 4).

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489



IMMOBILIERE D'UGINE

SIEGE SOCIAL : Paris (8e), 10, rue du Général-Foy. Tél. : LAB. 12-75.

CONSTITUTION : Société anonyme française constituée le 14 décembre 1909 pour une durée de 99 ans.

OBJET : L'acquisition de tous immeubles bâtis ou non, la construction et l'aménagement ainsi que la prise à bail de maisons d'habitation, dépendances et annexes en vue de réaliser des locations, locations-ventes et ventes réservées au personnel de la société Ugine ou de ses filiales ; la création et l'aménagement de jardins, bains-douches, lavoirs, terrains de sport, écoles, cantines, destinés au même personnel, etc...

CAPITAL : 3.500.000 NF, divisé en 350.000 actions de 10 NF nominatives.

OBLIGATIONS : 4 1/4 % demi-net 1944 : 7.500 obligations de 2.000 fr., am. de 1944 à 1974 T. ou R. sauf R.A. à partir du 1er août 1946. Coupon : 1er août.

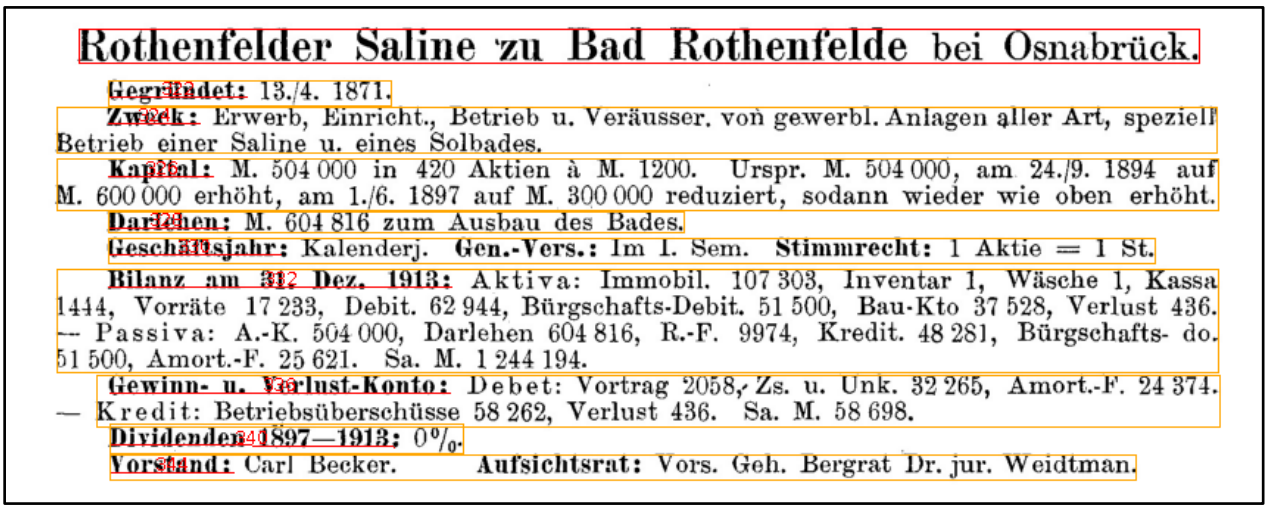
4 % demi-net 1945 : 7.000 obligations de 5.000 fr., am. de 1945 à 1975 T. ou R. sauf R.A. à partir du 1er septembre 1948. Coupon : 1er septembre.

Ces obligations ont été converties le 1er janvier 1960 en NF.

ASSEMBLEE GENERALE : Dans le courant du semestre qui suit la clôture de l'exercice.

Figure 4: Rubrics detection example on the Desfossés yearbook (1962, tome 2, page 267)

The same structure is found in the German Handbuch yearbook (1914-1915) with minor differences: rubrics titles are not in capital letters, indentation is not necessarily to the right or left, etc. With few changes and adaptations to these specificities, the same grammar can be applied to both yearbooks, and thus have the same source code and executable for all documents. As indentation is not reliable on this yearbook to delimit the beginning of a new rubric, we use ARU-Net to detect rubric titles without introducing major differences in the grammar (Figure 5).



Rothenfelder Saline zu Bad Rothenfelde bei Osnabrück.

Gegründet: 13./4. 1871.

Zweck: Erwerb, Einricht., Betrieb u. Veräusser. von gewerbl. Anlagen aller Art, speziell Betrieb einer Saline u. eines Solbades.

Kapital: M. 504 000 in 420 Aktien à M. 1200. Urspr. M. 504 000, am 24./9. 1894 auf M. 600 000 erhöht, am 1./6. 1897 auf M. 300 000 reduziert, sodann wieder wie oben erhöht.

Darlehen: M. 604 816 zum Ausbau des Bades.

Geschäftsjahr: Kalenderj. **Gen.-Vers.:** Im 1. Sem. **Stimmrecht:** 1 Aktie = 1 St.

Bilanz am 31. Dez. 1913: Aktiva: Immobil. 107 303, Inventar 1, Wäsche 1, Kassa 1444, Vorräte 17 233, Debit. 62 944, Bürgschafts-Debit. 51 500, Bau-Kto 37 528, Verlust 436. — Passiva: A.-K. 504 000, Darlehen 604 816, R.-F. 9974, Kredit. 48 281, Bürgschafts- do. 51 500, Amort.-F. 25 621. Sa. M. 1 244 194.

Gewinn- u. Verlust-Konto: Debet: Vortrag 2058, Zs. u. Unk. 32 265, Amort.-F. 24 374. — Kredit: Betriebsüberschüsse 58 262, Verlust 436. Sa. M. 58 698.

Dividenden 1897—1913: 0 %.

Vorstand: Carl Becker. **Aufsichtsrat:** Vors. Geh. Bergrat Dr. jur. Weidtman.

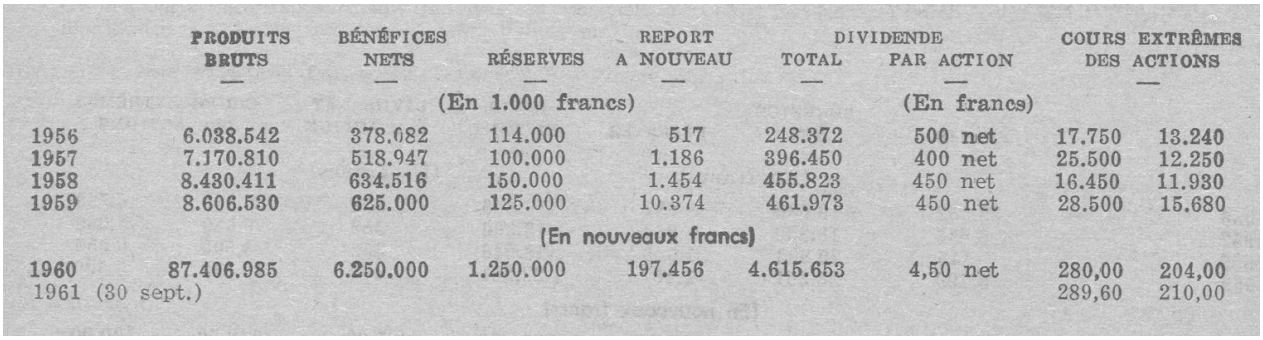
Figure 5: Rubrics detection example on the Handbuch yearbook (1914-15, page 951)

C. The specific case of tables and balance sheets

Alongside the rubrics we can find tables such as the one in Figure 6.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489



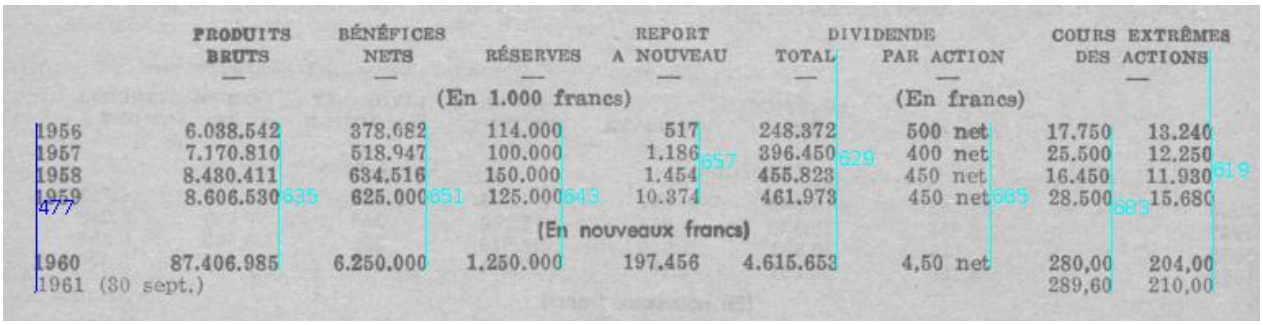
	PRODUITS BRUTS —	BÉNÉFICES NETS —	RÉSERVES —	REPORT A NOUVEAU —	DIVIDENDE TOTAL —	DIVIDENDE PAR ACTION —	COURS EXTRÊMES DES ACTIONS —	
			(En 1.000 francs)			(En francs)		
1956	6.038.542	378.682	114.000	517	248.372	500 net	17.750	13.240
1957	7.170.810	518.947	100.000	1.186	396.450	400 net	25.500	12.250
1958	8.480.411	634.516	150.000	1.454	455.823	450 net	16.450	11.930
1959	8.606.530	625.000	125.000	10.374	461.973	450 net	28.500	15.680
			(En nouveaux francs)					
1960	87.406.985	6.250.000	1.250.000	197.456	4.615.653	4,50 net	280,00	204,00
1961 (30 sept.)							289,60	210,00

Figure 6: Example of table found in the Desfossés yearbook (1962, page 150)

These tables take different forms depending on the content and the document they are from, but we can break them down to their most basic elements: columns and rows.

During the extraction task, each table cell should be linked to its row and column, and the information they convey. Some tables, such as the ones in the Desfossés, do not have any rulings so we have to rely on the text lines to deduct the structure.

In the case of the Desfossés yearbook, to each column is attached a currency that is usually written just below the column title. Columns are vertical alignments, and rows are the concatenation of horizontally aligned lines (Figure 7).



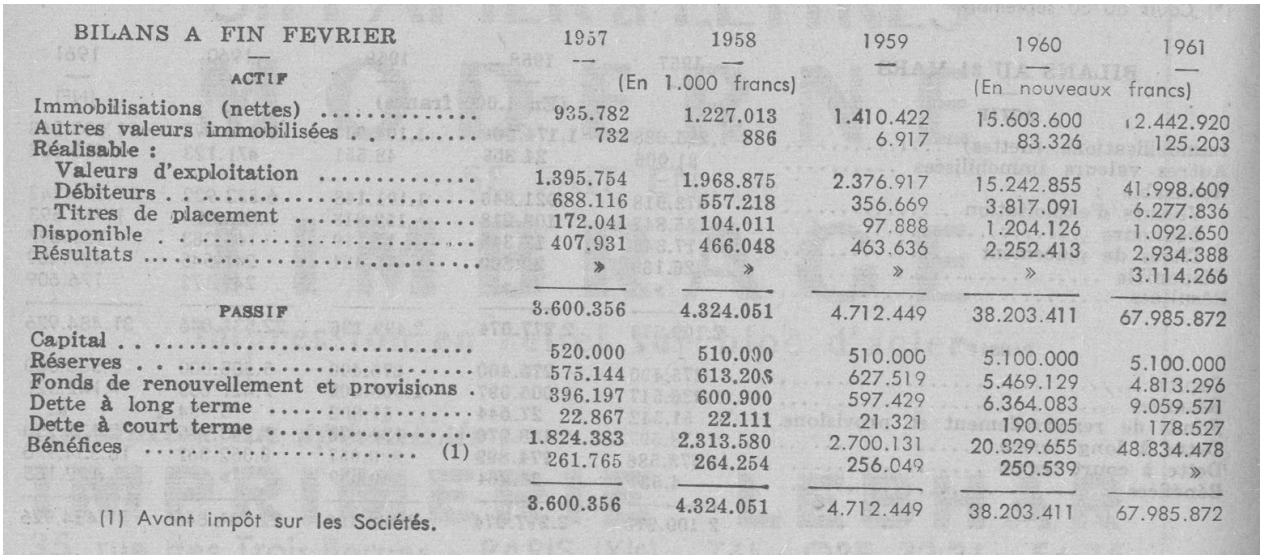
	PRODUITS BRUTS —	BÉNÉFICES NETS —	RÉSERVES —	REPORT A NOUVEAU —	DIVIDENDE TOTAL —	DIVIDENDE PAR ACTION —	COURS EXTRÊMES DES ACTIONS —	
			(En 1.000 francs)			(En francs)		
1956	6.038.542	378.682	114.000	517	248.372	500 net	17.750	13.240
1957	7.170.810	518.947	100.000	1.186	396.450	400 net	25.500	12.250
1958	8.480.411	634.516	150.000	1.454	455.823	450 net	16.450	11.930
1959	8.606.530	625.000	125.000	10.374	461.973	450 net	28.500	15.680
			(En nouveaux francs)					
1960	87.406.985	6.250.000	1.250.000	197.456	4.615.653	4,50 net	280,00	204,00
1961 (30 sept.)							289,60	210,00

Figure 7: Alignments detections in a table

At first we focused on a specific type of tables that are the balance sheets found in the Desfossés (Figure 8).

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489



BILANS A FIN FEVRIER	1957	1958	1959	1960	1961
ACTIF	(En 1.000 francs)			(En nouveaux francs)	
Immobilisations (nettes)	935.782	1.227.013	1.410.422	15.603.600	2.442.920
Autres valeurs immobilisées	732	886	6.917	83.326	125.203
Réalisable :					
Valeurs d'exploitation	1.395.754	1.968.875	2.376.917	15.242.855	41.998.609
Débiteurs	688.116	557.218	356.669	3.817.091	6.277.836
Titres de placement	172.041	104.011	97.888	1.204.126	1.092.650
Disponible	407.931	466.048	463.636	2.252.413	2.934.388
Résultats	»	»	»	»	3.114.266
	3.600.356	4.324.051	4.712.449	38.203.411	67.985.872
PASSIF					
Capital	520.000	510.000	510.000	5.100.000	5.100.000
Réserves	575.144	613.208	627.519	5.469.129	4.813.296
Fonds de renouvellement et provisions	396.197	600.900	597.429	6.364.083	9.059.571
Dette à long terme	22.867	22.111	21.321	190.005	178.527
Dette à court terme	1.824.383	2.313.580	2.700.131	20.929.655	48.834.478
Bénéfices (1)	261.765	264.254	256.049	250.539	»
	3.600.356	4.324.051	4.712.449	38.203.411	67.985.872

(1) Avant impôt sur les Sociétés.

Figure 8: Example of balance sheet found in Desfossés (1962, page 450)

These tables have a stable structure and unique elements (active and passive sections, total rows and sub-rows, see Figure 9), and they make use of the majority of the tools in the library:

- Precise textline detection so they do not get fused between columns;
- Alignment extraction;
- Integration of data collected from a commercial OCR so we can use keywords such as the section names ("actif"," passif") as positional information;
- Line recognition and reconstruction for detecting the whole lines.

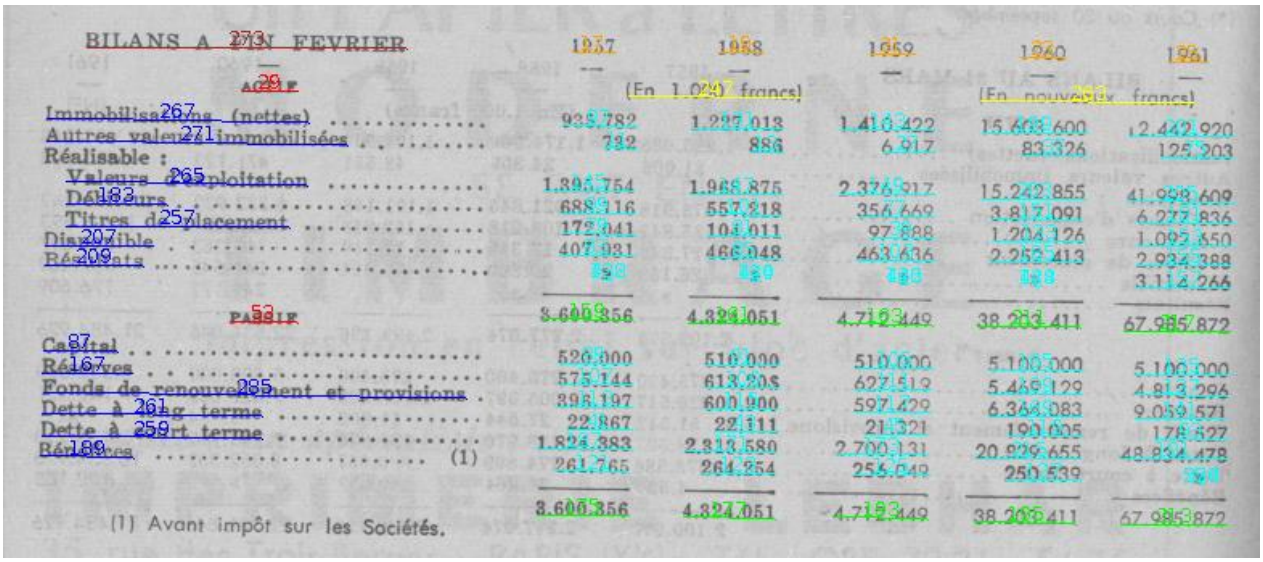
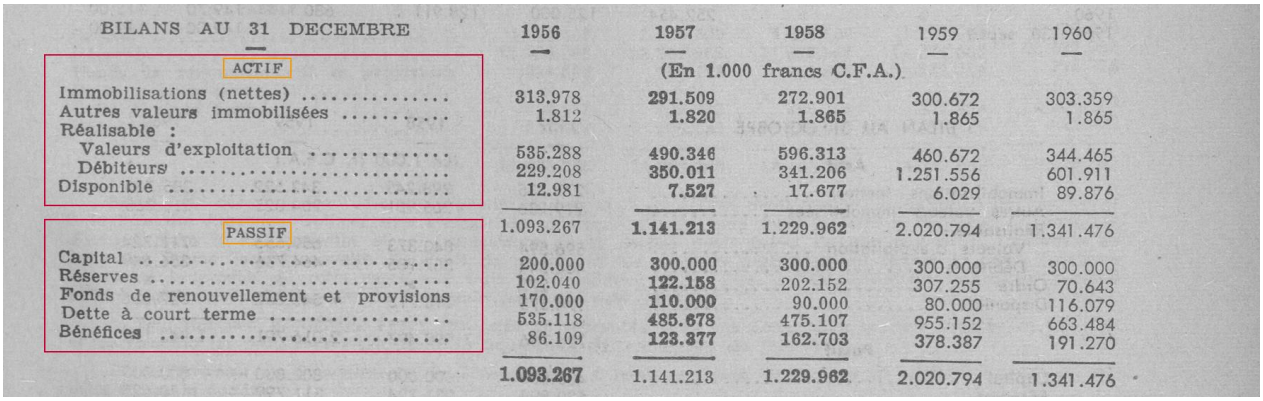


Figure 9: Example of balance sheet analysis: Table title (brown), Columns titles (orange), currency (yellow), sections titles (red), rows title (blue), normal cells (cyan) and total cells (green)

Balance sheets are split in sections containing several items (Figure 10).

This project has received funding from

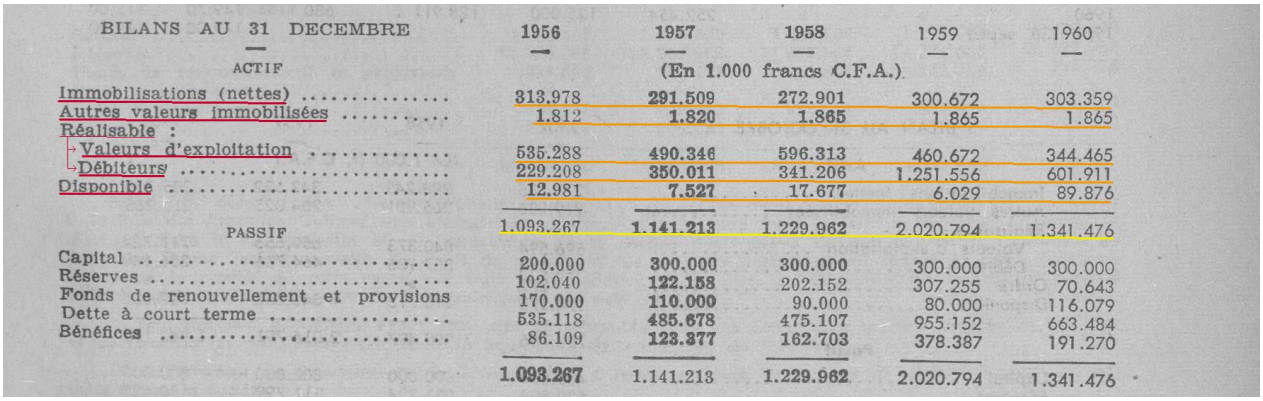
the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489



BILANS AU 31 DECEMBRE	1956	1957	1958	1959	1960
		(En 1.000 francs C.F.A.)			
ACTIF					
Immobilisations (nettes)	313.978	291.509	272.901	300.672	303.359
Autres valeurs immobilisées	1.812	1.820	1.865	1.865	1.865
Réalisable :					
Valeurs d'exploitation	535.288	490.346	596.313	460.672	344.465
Débiteurs	229.208	350.011	341.206	1.251.556	601.911
Disponible	12.981	7.527	17.677	6.029	89.876
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476
PASSIF					
Capital	200.000	300.000	300.000	300.000	300.000
Réserves	102.040	122.158	202.152	307.255	70.643
Fonds de renouvellement et provisions	170.000	110.000	90.000	80.000	116.079
Dette à court terme	535.118	485.678	475.107	955.152	663.484
Bénéfices	86.109	123.377	162.703	378.387	191.270
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476

Figure 10: Two sections of a balance sheet. There may be more than two but these are the most common.

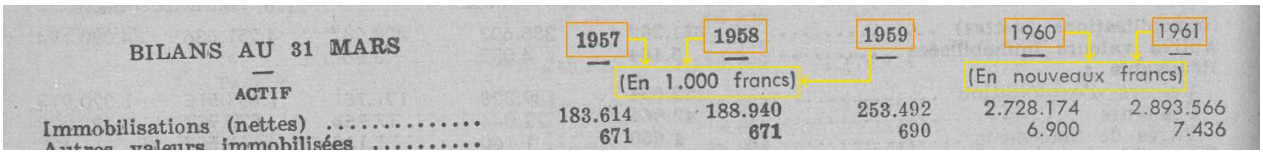
Items either correspond to a row of values or a title for one or more subitems. Each section has a total line where values are the sum of the cells right above in the column (Figure 11).



BILANS AU 31 DECEMBRE	1956	1957	1958	1959	1960
		(En 1.000 francs C.F.A.)			
ACTIF					
Immobilisations (nettes)	313.978	291.509	272.901	300.672	303.359
Autres valeurs immobilisées	1.812	1.820	1.865	1.865	1.865
Réalisable :					
Valeurs d'exploitation	535.288	490.346	596.313	460.672	344.465
Débiteurs	229.208	350.011	341.206	1.251.556	601.911
Disponible	12.981	7.527	17.677	6.029	89.876
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476
PASSIF					
Capital	200.000	300.000	300.000	300.000	300.000
Réserves	102.040	122.158	202.152	307.255	70.643
Fonds de renouvellement et provisions	170.000	110.000	90.000	80.000	116.079
Dette à court terme	535.118	485.678	475.107	955.152	663.484
Bénéfices	86.109	123.377	162.703	378.387	191.270
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476

Figure 11: Items and subitems with their respective rows and section total.

Each column is linked to a currency, based on which one is the nearest (Figure 12).



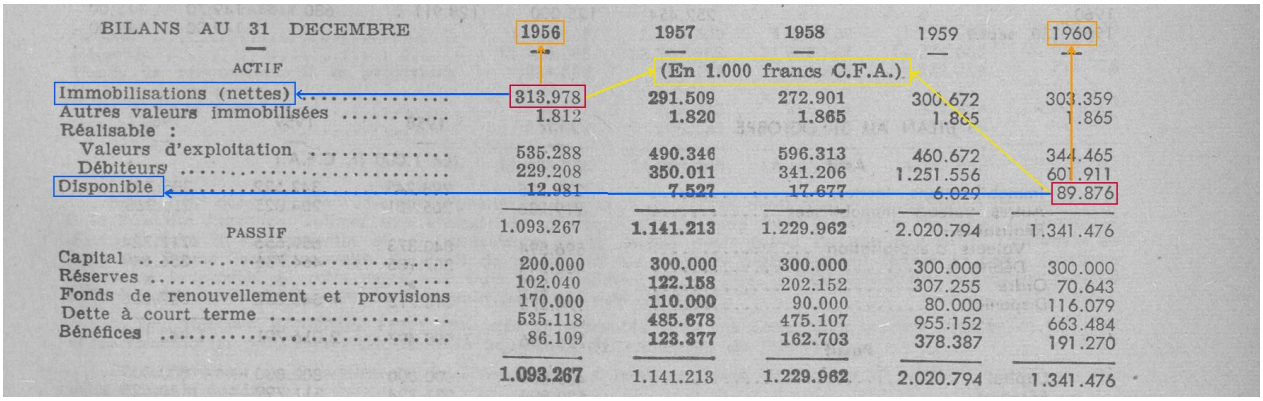
BILANS AU 31 MARS	1957	1958	1959	1960	1961
	(En 1.000 francs)			(En nouveaux francs)	
ACTIF					
Immobilisations (nettes)	183.614	188.940	253.492	2.728.174	2.893.566
Autres valeurs immobilisées	671	671	690	6.900	7.436

Figure 12: Example of columns and currencies.

All in all, each cell contains all this information: its item (row title) and what section it belongs to, its date (column title) and what currency it corresponds to, and its own value (Figure 13).

This project has received funding from

the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489



BILANS AU 31 DECEMBRE	1956	1957	1958	1959	1960
		(En 1.000 francs C.F.A.)			
ACTIF					
Immobilisations (nettes)	313.978	291.509	272.901	300.672	303.359
Autres valeurs immobilisées	1.812	1.820	1.865	1.865	1.865
Réalisable :					
Valeurs d'exploitation	535.288	490.346	596.313	460.672	344.465
Débiteurs	229.208	350.011	341.206	1.251.556	601.911
Disponible	12.981	7.527	17.677	6.029	89.876
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476
PASSIF					
Capital	200.000	300.000	300.000	300.000	300.000
Réserves	102.040	122.158	202.152	307.255	70.643
Fonds de renouvellement et provisions	170.000	110.000	90.000	80.000	116.079
Dette à court terme	535.118	485.678	475.107	955.152	663.484
Bénéfices	86.109	123.377	162.703	378.387	191.270
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476

Figure 13: All the information extracted for each cell of a balance sheet.

2.1.2 Evaluation

A. Evaluation process

We can evaluate detection and classification of bounding boxes on both the Desfossés and the Handbuch. We manually annotated 61 Desfossés pages for 4 classes: 61 titles, 793 rubrics, 52 balance sheets and 70 other tables.

We did the same for 35 Handbuch pages with only 2 classes: 47 titles and 524 rubrics (tables extraction is not necessary). More pages are to be annotated soon to extend the evaluation corpus.

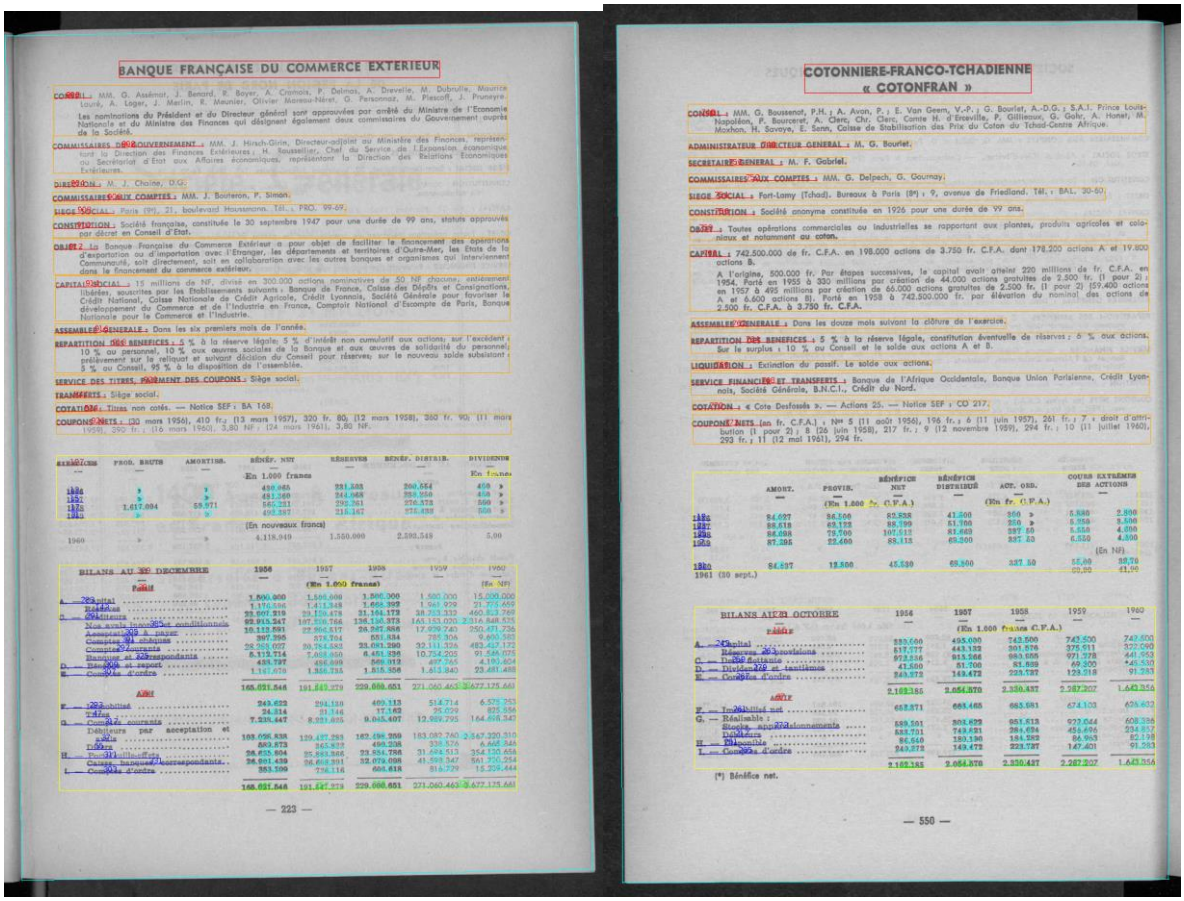
As a metric, ZoneMap [O. Galibert 2014] has been chosen because it takes into account overlaps between different bounding boxes, and merge/split errors. ZoneMap computes an error score: the objective is to get it near zero. It is an interesting score for relative comparison of systems but cannot be interpreted in an absolute way: it is not a percentage as bad scores can rise up with no real fixed limit, 1000 or more can occur on not well recognized documents). This metric has been used for the evaluation of document structure segmentation systems in the international Maurdor competition on the extraction of relevant information in scanned documents. Having such a metric helps us keep track of the improvement made in the grammar, both for Desfossés and Handbuch, and detect side effects of these modifications.

B. Results

Details on the results on the evaluation datasets on the Desfossés and the Handbuch are given in the deliverable D7.2. Figure 14 presents examples of entire Desfossés pages analysis.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489



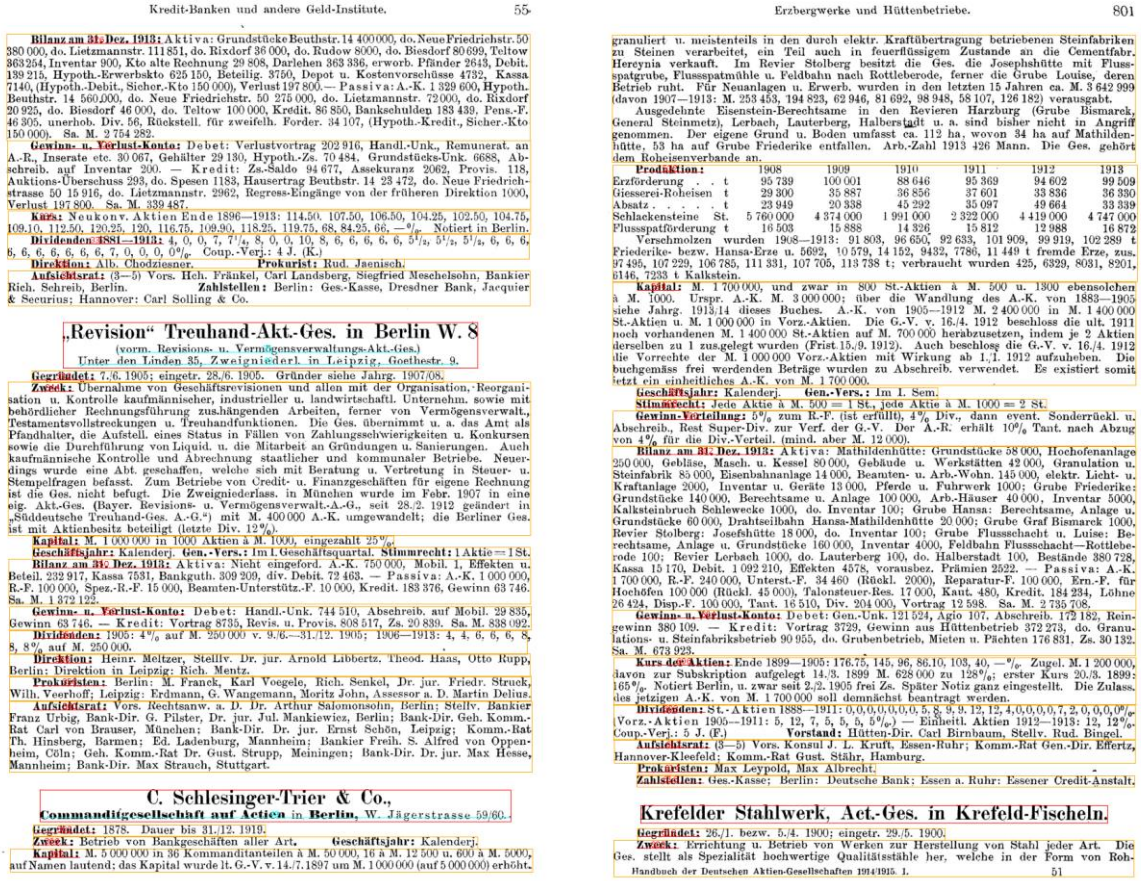
BANQUE FRANÇAISE DU COMMERCE EXTERIEUR

COTONNIERE-FRANCO-TCHADIENNE « COTONFRAN »

Figure 14: Entire Desfossés pages analysis (1962, pages 267 and 550)

Figure 15 presents examples of entire Handbuch page analysis.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489



Kredit-Banken und andere Geld-Institute. 55

Bilanz am 31. Dez. 1913: Aktiva: Grundstücke Beuthstr. 14 400 000, do. Neue Friedrichstr. 50 380 000, do. Lietzmannstr. 111 851, do. Rixdorf 36 000, do. Rudow 8000, do. Biesdorf 80 699, Teltow 363 254, Inventar 900, Kto alte Rechnung 29 808, Darlehen 363 356, erworb. Pfänder 2643, Debit. 139 215, Hypoth.-Erwerbskto 625 150, Beteiltg. 3750, Depot u. Kostenvorschüsse 4732, Kassa 7140, (Hypoth.-Debit., Sicher.-Kto 150 000), Verlust 197 800. — Passiva: A.-K. 1 329 600, Hypoth. Beuthstr. 14 560 000, do. Neue Friedrichstr. 50 275 000, do. Lietzmannstr. 72 000, do. Rixdorf 20 925, do. Biesdorf 46 000, do. Teltow 100 000, Kredit. 86 850, Bankschulden 183 439, Pens.-F. 46 305, unerhob. Div. 56, Rückstell. für zweifelh. Forder. 34 107, (Hypoth.-Kredit., Sicher.-Kto 150 000). Sa. M. 2 754 282.

Gewinn- u. Verlust-Konto: Debet: Verlustvortrag 202 916, Handl.-Unk., Remunerat. an A.-R., Inserate etc. 30 067, Gehälter 29 130, Hypoth.-Zs. 70 484, Grundstücks-Unk. 6688, Abschreib. auf Inventar 200. — Kredit: Zs.-Saldo 94 677, Assekuranz 2062, Provis. 118, Auktions-Überschuss 293, do. Spesen 1183, Hausertrag Beutbstr. 14 23 472, do. Neue Friedrichstrasse 50 15 916, do. Lietzmannstr. 2962, Regress-Eingänge von der früheren Direktion 1000, Verlust 197 800. Sa. M. 339 487.

Kurs: Neukonv. Aktien Ende 1896—1913: 114.50, 107.50, 106.50, 104.25, 102.50, 104.75, 109.10, 112.50, 120.25, 120, 116.75, 108.90, 118.25, 119.75, 68, 84.25, 66, —%. Notiert in Berlin.

Dividenden 1881—1913: 4, 0, 0, 7, 7½, 8, 0, 0, 10, 8, 6, 6, 6, 6, 6, 5½, 5½, 5½, 6, 6, 6, 6, 6, 6, 6, 6, 7, 0, 0, 0, 0%. Coup.-Verj.: 4 J. (K.)

Direktion: Alb. Chodziesner. **Prokurist:** Rud. Jaenisch.

Aufsichtsrat: (3—5) Vors. Hch. Fränkel, Carl Landsberg, Siegfried Meschelsohn, Bankier Rich. Schreib, Berlin. **Zahlstellen:** Berlin: Ges.-Kasse, Dresdner Bank, Jacquier & Securius; Hannover: Carl Solling & Co.

„Revision" Treuhand-Akt.-Ges. in Berlin W. 8

(vorm. Revisions- u. Vermögensverwaltungs-Akt.-Ges.)
Unter den Linden 35, Zweigniederl. in Leipzig, Goethestr. 9.

Gegründet: 7./6. 1905; eingetr. 28./6. 1905. Gründer siehe Jahrg. 1907/08.

Zweck: Übernahme von Geschäftsrevisionen und allen mit der Organisation, Reorganisation u. Kontrolle kaufmännischer, industrieller u. landwirtschaftl. Unternehm. sowie mit behördlicher Rechnungsführung zusammenhängenden Arbeiten, ferner von Vermögensverwalt., Testamentsvollstreckungen u. Treuhandfunktionen. Die Ges. übernimmt u. a. das Amt als Pfandhalter, die Aufstell. eines Status in Fällen von Zahlungsschwierigkeiten u. Konkursen sowie die Durchführung von Liquid. u. die Mitarbeit an Gründungen u. Sanierungen. Auch kaufmännische Kontrolle und Abrechnung staatlicher und kommunaler Betriebe. Neuerdings wurde eine Abt. geschaffen, welche sich mit Beratung u. Vertretung in Steuer- u. Stempelfragen befasst. Zum Betriebe von Credit- u. Finanzgeschäften für eigene Rechnung ist die Ges. nicht befugt. Die Zweigniederlass. in München wurde im Febr. 1907 in eine eig. Akt.-Ges. (Bayer. Revisions- u. Vermögensverwalt.-A.-G., seit 28./2. 1912 geändert in „Süddeutsche Treuhand-Ges. A.-G.") mit M. 400 000 A.-K. umgewandelt; die Berliner Ges. ist mit Aktienbesitz beteiligt (letzte Div. 12%).

Kapital: M. 1 000 000 in 1000 Aktien à M. 1000, eingezahlt 25%.

Geschäftsjahr: Kalenderj. **Gen.-Vers.:** Im I. Geschäftsquartal. **Stimmrecht:** 1 Aktie = 1 St.

Bilanz am 31. Dez. 1913: Aktiva: Nicht eingeford. A.-K. 750 000, Mobil. 1, Effekten u. Beteil. 232 917, Kassa 7531, Bankguth. 309 209, div. Debit. 72 463. — Passiva: A.-K. 1 000 000, R.-F. 100 000, Spez.-R.-F. 15 000, Beamten-Unterstütz.-F. 10 000, Kredit. 183 376, Gewinn 63 746. Sa. M. 1 372 122.

Gewinn- u. Verlust-Konto: Debet: Handl.-Unk. 744 510, Abschreib. auf Mobil. 29 836, Gewinn 63 746. — Kredit: Vortrag 8735, Revis. u. Provis. 808 517, Zs. 20 830. Sa. M. 838 092.

Dividenden: 1905: 4% auf M. 250 000 v. 9./6.—31./12. 1905; 1906—1913: 4, 4, 6, 6, 6, 6, 8, 8, 8% auf M. 250 000.

Direktion: Heinr. Meltzer, Stellv. Dr. jur. Arnold Libbertz, Theod. Haas, Otto Rupp, Berlin; Direktion in Leipzig: Rich. Mentz.

Prokuristen: Berlin: M. Franck, Karl Voegele, Rich. Senkel, Dr. jur. Friedr. Struck, Wilh. Veerhoff; Leipzig: Erdmann, G. Wangemann, Moritz John, Assessor a. D. Martin Delius.

Aufsichtsrat: Vors. Rechtsanw. u. D. Dr. Arthur Salomonsohn, Berlin; Stellv. Bankier Franz Urbig, Bank-Dir. G. Pilster, Dr. jur. Jul. Mankiewicz, Berlin; Bank-Dir. Geh. Komm.-Rat Carl von Brauser, München; Bank-Dir. Dr. jur. Ernst Schön, Leipzig; Komm.-Rat Th. Hinsberg, Barmen; Ed. Ladenburg, Mannheim; Bankier Freih. S. Alfred von Oppenheim, Cöln; Geh. Komm.-Rat Dr. Gust. Strupp, Meiningen; Bank-Dir. Dr. jur. Max Hesse, Mannheim; Bank-Dir. Max Strauch, Stuttgart.

C. Schlesinger-Trier & Co.,

Commanditgesellschaft auf Actien in **Berlin**, W. Jägerstrasse 59/60.

Gegründet: 1878. Dauer bis 31./12. 1919.

Zweck: Betrieb von Bankgeschäften aller Art. **Geschäftsjahr:** Kalenderj.

Kapital: M. 5 000 000 in 36 Kommanditanteilen à M. 50 000, 16 à M. 12 500 u. 600 à M. 5000, auf Namen lautend; das Kapital wurde lt. G.-V. v. 14./7. 1897 um M. 1 000 000 (auf 5 000 000) erhöht.

Erzbergwerke und Hüttenbetriebe. 801

granuliert u. meistenteils in den durch elektr. Kraftübertragung betriebenen Steinfabriken zu Steinen verarbeitet, ein Teil auch in feuerflüssigem Zustande an die Cementfabr. Herzynia verkauft. Im Revier Stolberg besitzt die Ges. die Josephshütte mit Flussspatgrube, Flussspatmühle u. Feldbahn nach Rottleberode, ferner die Grube Louise, deren Betrieb ruht. Für Neuanlagen u. Erwerb. wurden in den letzten 15 Jahren ca. M. 3 642 999 (davon 1907—1913: M. 253 453, 194 823, 62 946, 81 692, 98 948, 58 107, 126 182) verausgabt.

Ausgedehnte Eisenstein-Berechtsame in den Revieren Harzburg (Grube Bismarck, General Steinmetz), Lerbach, Lauterberg, Halberstadt u. a. sind bisher nicht in Angriff genommen. Der eigene Grund u. Boden umfasst ca. 112 ha, wovon 34 ha auf Mathildenhütte, 53 ha auf Grube Friederike entfallen. Arb.-Zahl 1913 426 Mann. Die Ges. gehört dem Roheisenverbande an.

Produktion:	1908	1909	1910	1911	1912	1913
Erzförderung . . t	95 739	100 001	88 646	95 369	94 602	99 509
Giesserei-Roheisen t	29 300	35 887	36 856	37 601	33 836	36 330
Absatz . . . . . t	23 949	20 338	45 292	35 097	49 664	33 339
Schlackensteine St.	5 760 000	4 374 000	1 991 000	2 322 000	4 419 000	4 747 000
Flusspatförderung t	16 503	15 888	14 326	15 812	12 988	16 872

Verschmolzen wurden 1908—1913: 91 803, 96 650, 92 633, 101 909, 99 919, 102 289 t Friederike- bezw. Hansa-Erze u. 5692, 10 579, 14 152, 9432, 7786, 11 449 t fremde Erze, zus. 97 495, 107 229, 106 785, 111 331, 107 705, 113 738 t; verbraucht wurden 425, 6329, 8031, 8201, 6146, 7233 t Kalkstein.

Kapital: M. 1 700 000, und zwar in 800 St.-Aktien à M. 500 u. 1300 ebensolchen à M. 1000. Urspr. A.-K. M. 3 000 000; über die Wandlung des A.-K. von 1883—1905 siehe Jahrg. 1913/14 dieses Buches. A.-K. von 1905—1912 M. 2 400 000 in M. 1 400 000 St.-Aktien u. M. 1 000 000 in Vorz.-Aktien. Die G.-V. v. 16./4. 1912 beschloss die ult. 1911 noch vorhandenen M. 1 400 000 St.-Aktien auf M. 700 000 herabzusetzen, indem je 2 Aktien derselben zu 1 zus.gelegt wurden (Frist 15./9. 1912). Auch beschloss die G.-V. v. 16./4. 1912 die Vorrechte der M. 1 000 000 Vorz.-Aktien mit Wirkung ab 1./1. 1912 aufzuheben. Die buchgemäss frei werdenden Beträge wurden zu Abschreib. verwendet. Es existiert somit jetzt ein einheitliches A.-K. von M. 1 700 000.

Geschäftsjahr: Kalenderj. **Gen.-Vers.:** Im I. Sem.

Stimmrecht: Jede Aktie à M. 500 = 1 St., jede Aktie à M. 1000 = 2 St.

Gewinn-Verteilung: 5% zum R.-F. (ist erfüllt), 4% Div., dann event. Sonderrückl. u. Abschreib., Rest Super-Div. zur Verf. der G.-V. Der A.-R. erhält 10% Tant. nach Abzug von 4% für die Div.-Verteil. (mind. aber M. 12 000).

Bilanz am 31. Dez. 1913: Aktiva: Mathildenhütte: Grundstücke 58 000, Hochofenanlage 250 000, Gebläse, Masch. u. Kessel 80 000, Gebäude u. Werkstätten 42 000, Granulation u. Steinfabrik 85 000, Eisenbahnanlage 14 000, Beamten- u. Arb.-Wohn. 145 000, elektr. Licht- u. Kraftanlage 2000, Inventar u. Geräte 13 000, Pferde u. Fuhrwerk 1000; Grube Friederike: Grundstücke 140 000, Berechtsame u. Anlage 100 000, Arb.-Häuser 40 000, Inventar 5000, Kalksteinbruch Schlewecke 1000, do. Inventar 100; Grube Hansa: Berechtsame, Anlage u. Grundstücke 60 000, Drahtseilbahn Hansa-Mathildenhütte 20 000; Grube Graf Bismarck 1000, Revier Stolberg: Josefshütte 18 000, do. Inventar 100; Grube Flussschacht u. Luise: Berechtsame, Anlage u. Grundstücke 160 000, Inventar 4000, Feldbahn Flussschacht—Rottleberode 100; Revier Lerbach 1000, do. Lauterberg 100, do. Halberstadt 100, Bestände 380 728, Kassa 15 170, Debit. 1 092 210, Effekten 4578, vorausbez. Prämien 2522. — Passiva: A.-K. 1 700 000, R.-F. 240 000, Unterst.-F. 34 480 (Rückl. 2000), Reparatur-F. 100 000, Ern.-F. für Hochöfen 100 000 (Rückl. 45 000), Talonsteuer-Res. 17 000, Kaut. 480, Kredit. 184 234, Löhne 26 424, Disp.-F. 100 000, Tant. 16 510, Div. 204 000, Vortrag 12 598. Sa. M. 2 735 708.

Gewinn- u. Verlust-Konto: Debet: Gen.-Unk. 121 524, Agio 107, Abschreib. 172 182, Reingewinn 380 109. — Kredit: Vortrag 3729, Gewinn aus Hüttenbetrieb 372 273, do. Granulations- u. Steinfabriksbetrieb 90 955, do. Grubenbetrieb, Mieten u. Pächten 176 831, Zs. 30 132. Sa. M. 673 923.

Kurs der Aktien: Ende 1899—1905: 176.75, 145, 96, 86.10, 103, 40, —%. Zugel. M. 1 200 000, davon zur Subskription aufgelegt 14./3. 1899 M. 628 000 zu 128%; erster Kurs 20./3. 1899: 165%. Notiert Berlin, u. zwar seit 2./2. 1905 frei Zs. Später Notiz ganz eingestellt. Die Zulass. des jetzigen A.-K. von M. 1 700 000 soll demnächst beantragt werden.

Dividenden: St.-Aktien 1888—1911: 0, 0, 0, 0, 0, 0, 0, 5, 8, 9, 9, 12, 12, 4, 0, 0, 0, 0, 7, 2, 0, 0, 0, 0%. Vorz.-Aktien 1905—1911: 5, 12, 7, 5, 5, 5, 5%. — Einheitl. Aktien 1912—1913: 12, 12%. Coup.-Verj.: 5 J. (F.) **Vorstand:** Hütten-Dir. Carl Birnbaum, Stellv. Rud. Bingel.

Aufsichtsrat: (3—5) Vors. Konsul J. L. Krupp, Essen-Ruhr; Komm.-Rat Gen.-Dir. Effertz, Hannover-Kleefeld; Komm.-Rat Gust. Stähr, Hamburg.

Prokuristen: Max Leypold, Max Albrecht.

Zahlstellen: Ges.-Kasse; Berlin: Deutsche Bank; Essen a. Ruhr: Essener Credit-Anstalt.

Krefelder Stahlwerk, Act.-Ges. in Krefeld-Fischeln.

Gegründet: 26./1. bezw. 5./4. 1900; eingetr. 29./5. 1900.

Zweck: Errichtung u. Betrieb von Werken zur Herstellung von Stahl jeder Art. Die Ges. stellt als Spezialität hochwertige Qualitätsstähle her, welche in der Form von Roh-

Handbuch der Deutschen Aktien-Gesellschaften 1914/1915. I. 51

Figure 15: Entire Handbuch pages analysis (1914-15, pages 55 and 801)

2.2 Information Extraction System (Task 7.3)

2.2.1 System

A. Task description

Financial yearbooks contain a lot of information organised in rubrics. Each rubric is reporting about specific financial information and requires a dedicated extraction module. Some rubrics are simply lists of items such as lists of persons for which simple rules encoded with regular expressions may be enough to get the information extracted. Some other rubrics are much more difficult to analyse as shown on Figure 16 which gives two examples of the rubric "Capital" for both the French and German yearbooks. The two rubrics shown on this figure highlight the typical content of financial yearbooks whatever the stock exchange and the language used.

The rubric "Capital" reports the current capital amount of the company, its currency, the number and amount of shares that build the total capital amount. Moreover, the rubric "capital" is also reporting the

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

history of the capital by showing capital changes from the creation date of the company. Each capital change should occur with the date of change, the new capital value, the currency, the new number of shares with their amount. As these named entities are reported through a textual description and not placed into a table, a certain variability was introduced in phrasing the text at the time of publication. In addition, some information is sometimes partially missing. Figure 16 shows a case where the same tag set can be used for French and German. Notice that irrelevant words in the text are labelled with the tag "Other", as is the standard convention adopted for information extraction.

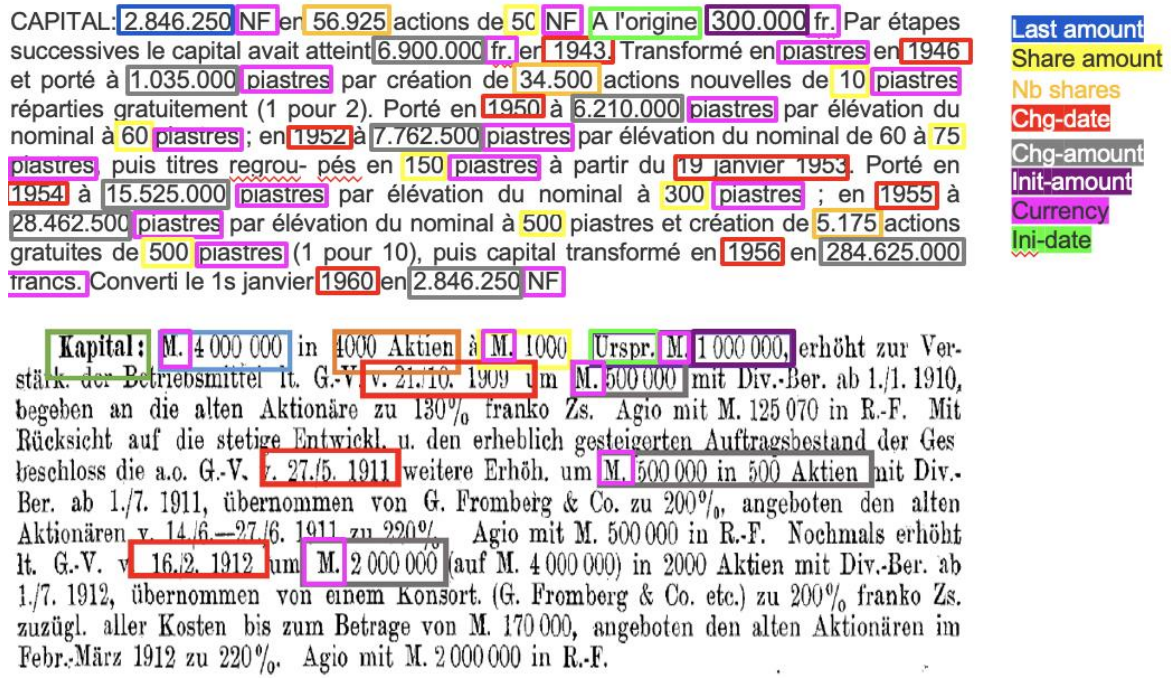


Figure 16: Information to be extracted from the rubric Capital for both the French and German yearbooks with the tagging conventions shown.

One other important aspect is related to how these various information should be linked together to provide timely coherent n-tuples of information in a tabular form as follows:

[date - capital amount - currency - number of shares - amount of share]

Such a 5-tuple is made of linked named entities and we wish the extraction process to extract not only each individual entity but also its linking attributes with the other entities they relate to. In this purpose we have introduced a specific "Link" tag that serves for tagging every non informative word within a single n-tuple, so that a n-tuple is any sequence of tags between two "Other" tags, see Figure 17 below.

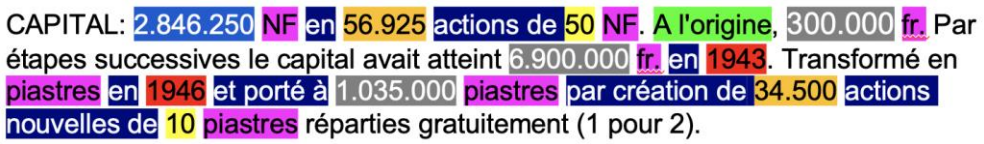


Figure 17: Tagging the linked named entities with tag "Link" in blue.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

B. Statistical models for information extraction

The proposed models are described in [Swaileh 2020]. A Conditional Random Field (CRF) [Lafferty 2001] allows to compute the conditional probability of a sequence of labels $Y = \{y_1, \quad y_2 \quad, \ldots, \quad y_T\}$ given a sequence of input features $X = \{x_1, \quad x_2 \quad, \ldots, \quad x_T\}$with the following equation:

$$P(Y|X) = \frac{1}{Z_0} exp\left(\sum_{t=1}^{T} \sum_{k} w_k \times \phi(y_{t-1}, y_t, x, t)\right)$$

where $\phi_k(y_{t-1}, y_t, x, t)$ is a feature function that maps the entire input sequence of features X paired with the entire output sequence of labels Y to some d-dimensional feature vector. Weight parameters w_k are optimised during training. The normalisation factor Z_0 is introduced to make sure that the sum of all conditional probabilities is equal to 1. Once the optimal weights $\hat{w}$ are estimated, the most likely sequence of output labels $\hat{Y}$for a given sequence of input features X is estimated as follows:

$$\hat{Y} = arg \max_{Y} P(Y|X)$$

CRF have been introduced for Natural Language Processing by considering binary features. $\phi(y_{t-1}, y_t, x, t)$ is a binary feature function that is set to 1 when labels and input tokens match a certain property. We use a 5 tokens width sliding window and a set of 33×5handcrafted templates that describe the text information.

In the literature, Recurrent Neural Network (RNN) architectures have been introduced so as to learn token embeddings. These embeddings are then used in place of the handcrafted features in a CRF model. Most of the state of the art NER systems use pre-trained word embeddings with a standard BLSTM-CRF setup [Akbik 2018, Grover 2008, Lample 2016, Huang 2015]. In addition to pre-trained word embeddings Lample et al. have introduced character-level word embeddings so as to circumvent possible out of vocabulary words. Similarly Peter et al. introduced contextual word embeddings extracted from a multi-layer bidirectional language model of tokens (biLM). Recently, Akbik et al. have used the internal states of two LSTM character language models to build contextual word embeddings, namely contextual string embeddings. Compared to other state of the art systems, this model is able to provide embeddings to any word and not only the known vocabulary words of the training set. Each language model consists of a single layer of 2048 Long Short Term Memory (LSTM) cells. A language model estimates the probability $P(x_{0:T})$ of a sequence of characters $(x_0; : : : ; x_T)$, $x_{0:T}$) with the following equation.

$$P(x_{0:T}) = \prod_{t=0}^{T} P(x_t|x_{0:t-1})$$

where $P(x_t/x_{0:t-1})$is the probability of observing a character given its past. A forward language model $(\overrightarrow{LM})$ computes the conditional probability using the LSTM hidden states as follows:

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

$$P(x_t|x_{0:t-1}) \approx \prod_{t=0}^{T} P(x_t|\overrightarrow{h_t};\theta)$$

where represents a view of the LSTM of the past sequence of characters of character xt while represents the model parameters. Similarly, a backward language model computes the probability in the reverse direction as follows:

$$P(x_t|x_{t+1:T}) \approx \prod_{t=0}^{T} P(x_t|\overleftarrow{h_t};\theta)$$

The word embedding w_i of word i that starts at character x_b and ends at character xe in the sentence is obtained by the concatenation of the hidden states of the forward and the backward LM as follows:

$$w_i = [\overleftarrow{h}_{b-1}, \overrightarrow{h}_{e+1}]$$

Notice that the two character language models can be trained on un-annotated large corpora as they are trained to predict the next/previous character. Then, following the architecture proposed in [Akbik 2018], we use a hybrid BILSTM/CRF model for named entity recognition. A word level BLSTM captures word context in the sentence and its internal state feeds a CRF in place of handcrafted features. The word BLSTM is fed by the string embedding representation. This BILSTM/CRF architecture is trained on for each specific Named Entity Recognition task, while it is fed by the string embedding representation that is pre-trained on a large corpus of the language chosen. In the following experiments we used pre-trained string embeddings proposed by the authors for French and German.

C. Active learning scheme

Due to the lack of annotated data, we have introduced an active learning scheme [Settles 2009]. First, we start by training the extraction model with a few annotated examples. The trained model is then used to predict the annotation of all the unseen examples of the test data set. These automatically annotated examples are sorted according to their labelling score. The examples with a labelling score higher than 0.9 are used as additional training examples to the first training data set for a new training iteration. The examples with a labelling score less than 0.5 are considered as bad examples. Then a small set of those bad examples are annotated manually for enhancing the capacity of the extraction model towards these bad examples. Figure 18 below describes the active learning system architecture.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

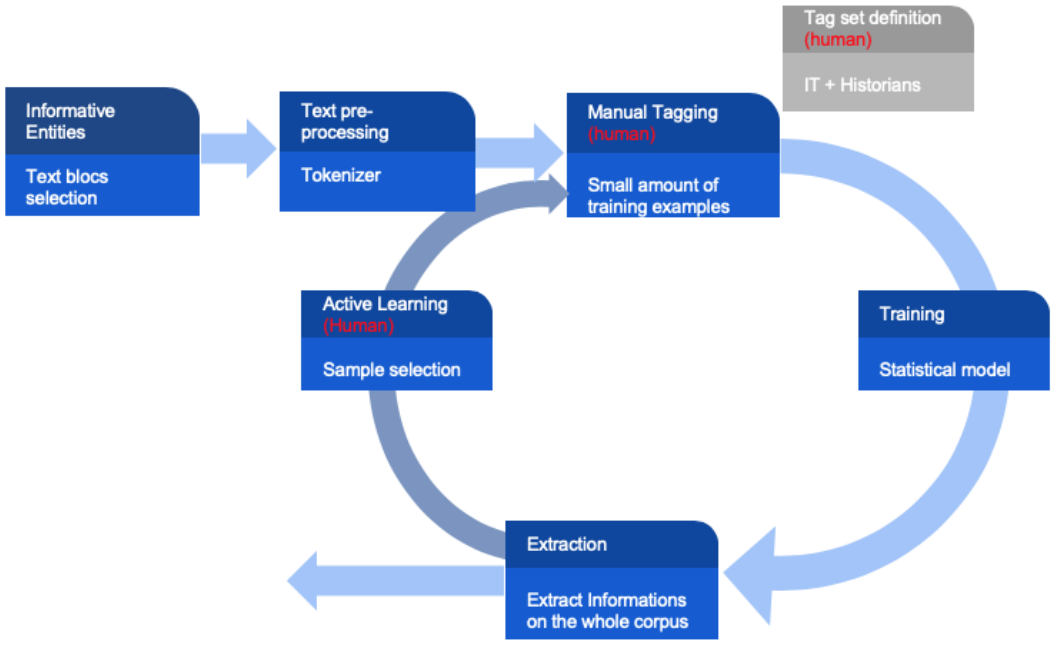


Figure 18: General architecture of the named entity extraction system, with human interaction in an active learning loop.

2.2.2 Evaluation

For the time being, evaluations have been conducted on a limited amount of annotated data. More experiments should be conducted when more annotated data will be available. This will be possible when the web annotation interface will be made accessible to the final users so as to allow an easy consultation / validation / correction experience of the extraction results. As the annotation interface was under development during this period (see next section below), we came to a less flexible way of annotating the extraction results through exchanging spreadsheets.

In order to have an overview, although limited for now, of the information extraction system performance experiments have been conducted on two selected corpora, namely the French Desfossé 1962 and the German Handbuch 1914-15 yearbooks (see deliverable D7.1 for a detailed description of the corpora selected). The illustrations below come from these two yearbooks. It is interesting to notice that despite their different layout, they are organized very similarly in terms of rubrics and the type of information one can find in each rubric.

From these considerations, the system performance have been evaluated on these two yearbooks on two particular rubrics that exhibit some difficulties in terms of specification, namely the capital rubrics (Capital and Kapital) and the constitution rubrics (Constitution and Gegründet). The results illustrate perfectly the strength of the machine learning approach that we propose. These evaluations highlight different aspects of the system: 1- strength of the BLSTM-CRF hybrid algorithm; 2- capacity for extracting linked entities and not only isolated entities; 3- interest of the active learning scheme and user interaction; and 4- capacity to deal with different languages in a transparent manner.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

CAPITAL rubric named entity extraction: The information to be extracted from this rubric are every capital amount, currencies and change dates. The tag set was derived from the examples in Figure 16.

Kapital rubric named entity extraction: Kapital rubric contains the same set of named entities to be extracted as for the CAPITAL rubric. In addition, two new named entities have been considered; Cap-decr and Cap-incr. These two labels refer to an increase or a decrease of the capital. In deliverable D7.2 - Table 3 we show the extraction results on the CAPITAL and Kapital rubrics and using the CRF and the BLSTM-CRF extraction models. We observe very good performance on the CAPITAL rubric with both the CRF and the BLSTM-CRF model with small differences. For every entity we obtain precision and recall higher than 95% while the average F1-score is higher than 96%. We also observe similar excellent performance on the Kapital rubric. However, the BLSTM-CRF model performs better than the CRF model. Both the CRF and the BLSTM.

CONSTITUTION rubric names entity extraction: One example is illustrated on Figure 19 below. From this rubric, we want to extract information about the company legal status, the date of creation, the period of activity and expiration date if applicable. We have introduced nine tags for this rubric, defined as follows: 1) ini-status: initial legal status of the company once created. 2) ini-startdate: the company creation date. 3) ini-enddate: the company expiration date. 4) ini-period: the company activity period. 5) chg-status: the changed legal status of the company. 6) chg-startdate: the start date of the changed legal status of the company. 7) chg-enddate: the end date of the changed legal status of the company. 9) link: the linking tag. In deliverable D7.2 - table 4 we report the results on the CONSTITUTION rubric using the CRF and BLSTM-CRF models. Due to the small size of the training data set, the results show lower performance compared to those reported on the CAPITAL and Kapital rubrics.

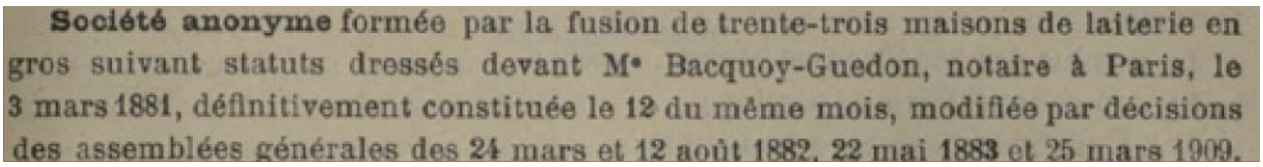

Société anonyme formée par la fusion de trente-trois maisons de laiterie en gros suivant statuts dressés devant Me Bacquoy-Guedon, notaire à Paris, le 3 mars 1881, définitivement constituée le 12 du même mois, modifiée par décisions des assemblées générales des 24 mars et 12 août 1882, 22 mai 1883 et 25 mars 1909.

Figure 19: One example of the Constitution rubric found on the Desfossés 1962 yearbook.

Entity linking

Once the entities have been extracted, we link them into tuples called chunks. We consider three different chunks on the CAPITAL and Kapital rubrics; 1) the ini-chunk consists of the ini-date, ini-amount and currency labelled tokens. 2) the chg-chunk includes the chg-date, chg-amount and currency labelled tokens. 3) the last-chunk enclose last-amount and currency labelled tokens. Notice that the date associated with the last-amount entity is the date of the yearbook (1962), and for this reason we don't consider extracting this information.

We experimented two methods for linking the entities into chunks. The minimum distance method regroups entities with their closest neighbour entity. Using the link tag, we are able to learn how to link the entities. We then consider a sequence of linked entities, as entities of the same chunk. The two

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

methods have been evaluated on the CAPITAL rubric using the CRF model, see deliverable D7.2 - table 5. The results show better performance when using the learned tag link.

Active learning

To show the effectiveness of the active learning scheme, we conducted three experiments on the CAPITAL rubric of the French Desfossés 1962 Yearbook. During these experiments, we used 200 manually annotated examples for evaluating the performance of the trained models.

From the first experiment we have evaluated how increasing the size of the training dataset on the performance. The more data the better the performance.

In the second experiment, we studied the effect of increasing the training dataset with the examples labelled by the model it-self whose labelling score is higher than 0.9. From this second experiment, we can say that the model learns better from the same examples by specialising to almost similar examples, whereas it does not cope with rare examples that the system does not . To tackle this problem, the training dataset must contain more heterogeneous examples.

We introduced this notion in the third experiment during which we not only inject labelled data with high scores but also some poorly labelled examples with a labelling score < 0.5 (C10) which are manually corrected and then introduced in the training dataset for the next training iteration. After five active learning iterations, we observe a quick increase of recall and F1-score with a slight degradation of precision (see table 1). In comparison with the results obtained from the first experiment, we observe that with only 30 automatically selected and manually annotated examples and three training iterations, the performance reaches the performance obtained during the first experiment.

2.2.3 Specification of tags for each rubric

In order to categorise the information inside the text of each section of the yearbooks, the CRF model uses a set of tags in order to labelise the data as instructed. A set of tags is the list of items that we wish to extract from the text in a coherent way. This can be any range of items like for example the *FoundingDate* or the *CapitalAmount* of an enterprise. Every yearbook has its own way of writing and although many rubrics are similar in the sense of reading, some others have their information written in a different way for the same idea.

In order for us to be able to understand the items to be extracted, we make a first meeting and a first draft of the specifications of the corpus. We also decide the priority of the rubrics to treat in order to have an organised list of rubrics to work on. As for now, the two yearbooks that have been treated are the French Yearbook *Desfossés 1962* and the German Yearbook *Handbuch 1914-1915*. For both yearbooks a lot of communication was needed between LITIS and the economist/historian groups of each country. This interaction involved the understanding of the information, creating an optimal list of tags for each rubric of their yearbook and even more understanding the language and specific expressions of certain texts.

This process takes many iterations, mostly because of the size a yearbook and the quantity of pages. This means that even if a specification document was done, there have always been some special cases that

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

were hard to find in the yearbook and so extra consultation was needed in order to know how to proceed with these new cases.

The specification documents for the two Yearbooks have been written:

- Specification Document for the French Yearbook Desfossés 1962
- Specification Document for the German Yearbook Handbuch 1914-1915

It is important to keep in mind that the tag definition is crucial before any data can be labelled or used. There the best strategy to adopt in order to be efficient is for the economist/historic group to understand what they want to extract, and the complexity of each rubric.

2.2.4 Labelled Datasets

Once the list of rubrics is declared, the tags are decided and the understanding of the rubrics is done, we started the first step which is the annotation of the first dataset. This dataset is used to test the system and get a feedback about the performance and understand how much more data is needed. We present in table 1 and table 2 the different rubrics that have been treated for each yearbook as well as the quantity of examples that were needed.

French Yearbook Desfossés 1962	
TREATED RUBRIC	**LABELED RUBRICS FOR THE SYSTEM**
Administrators	189
Headquarters	190
Founding	161
Capital	317
Operations	96
Sales	175
Financial Year	114
Coupons	173

Table 1: Treated rubrics with number of labelled examples on the French Desfossés 1962 yearbook

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

German Yearbook Handuch 1914-1915	
TREATED RUBRIC	**LABELED RUBRICS FOR THE SYSTEM**
Capital	195
Founding	900
Financial Year	598
Balance Sheet	130
Voting Right	671

Table 2: Treated rubrics with number of labelled examples on the German Yearbook Handbuch 1914-1915

2.2.5 Annotation Interface

In order to visualise, correct and validate the numerous pages and data that is treated in the yearbooks, a new interface has been developed as a way to facilitate the different needs towards these documents. This EURHISFIRM WP7 Data Extraction Viewer interface contains different features for navigating as well as visualising the data extracted from the documents, a login account is needed in order to access most of these features.

Search and Filter specific documents

The home page of the interface allows the user to select any collection or yearbook that will be available in the interface as well as filter them by the desired category (ex: Capital, Administrators, etc…). Since the interface is also a way to produce data for the Information Extraction System, it is also possible to filter the documents by non-validated and validated pages. This way the correct data will remain untouched for data extraction and also for learning material for the system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

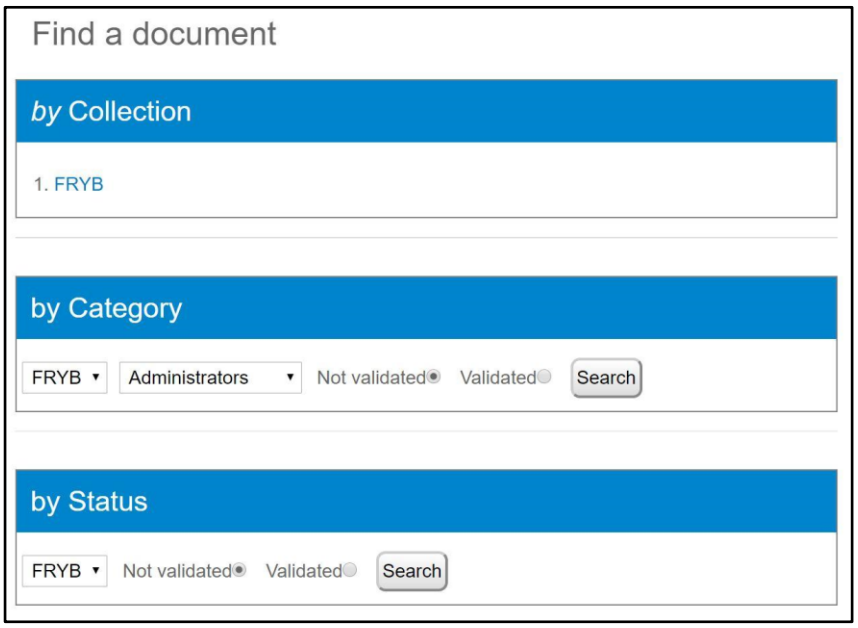


Figure 20: Search and Filter feature of the interface

Visualisation of a document

The visualisation of the documents has been designed according to the structure of the document, respecting its order and segmentation of paragraphs and categories as well as tables that are printed in the pages. Each page will contain different section titles in English and also in its mother language. This will facilitate the lecture of the document by other members of different countries. Once a section has been chosen, its content will be highlighted and pointed in order to show its location in the page and the text will be shown right away.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

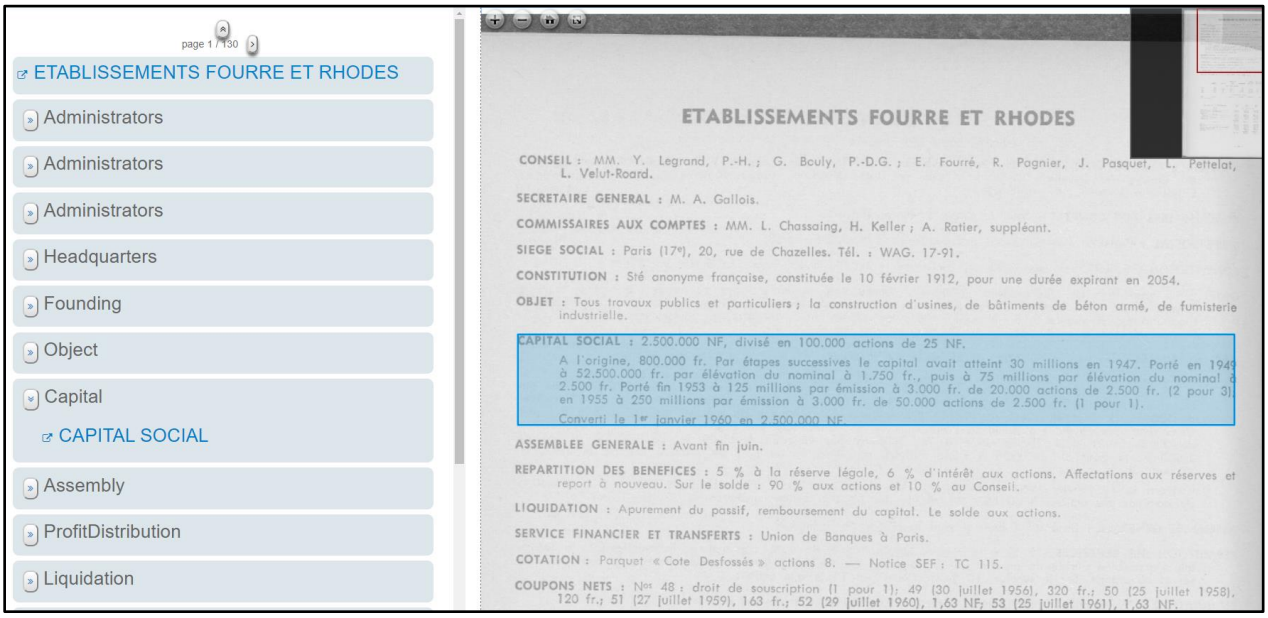


Figure 21: Visualisation of the sections of a document in English

Visualisation of a textual section

Once the text from the section has been clicked, the user will then arrive at the correction interface. This page will show the selected text and the proper tools for correcting the data as well as validating it and getting the result of the tagging in an extraction table. The procedure to correct any data follows a simple coloring tool linked to a certain label. The user will read the text and then proceed to click on the different labels and color the text in order for the Extraction System to digest the information and transform it into a result table.

The result table uses the chunks in order to construct the entries of information. A chunk, as previously mentioned, is the group of tags in a phrase that are different from "O". As the text is corrected, the interface will reconstruct the result table immediately to reflect the new changes that were made and understand the changes that were made. This will give fast feedback to the user about the tagging done and also help him understand how the algorithm works to create this table.

The extraction table is the visual representation of the extracted data that will be produced by the interface which will be then sent to the main database for later analysis by other teams.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

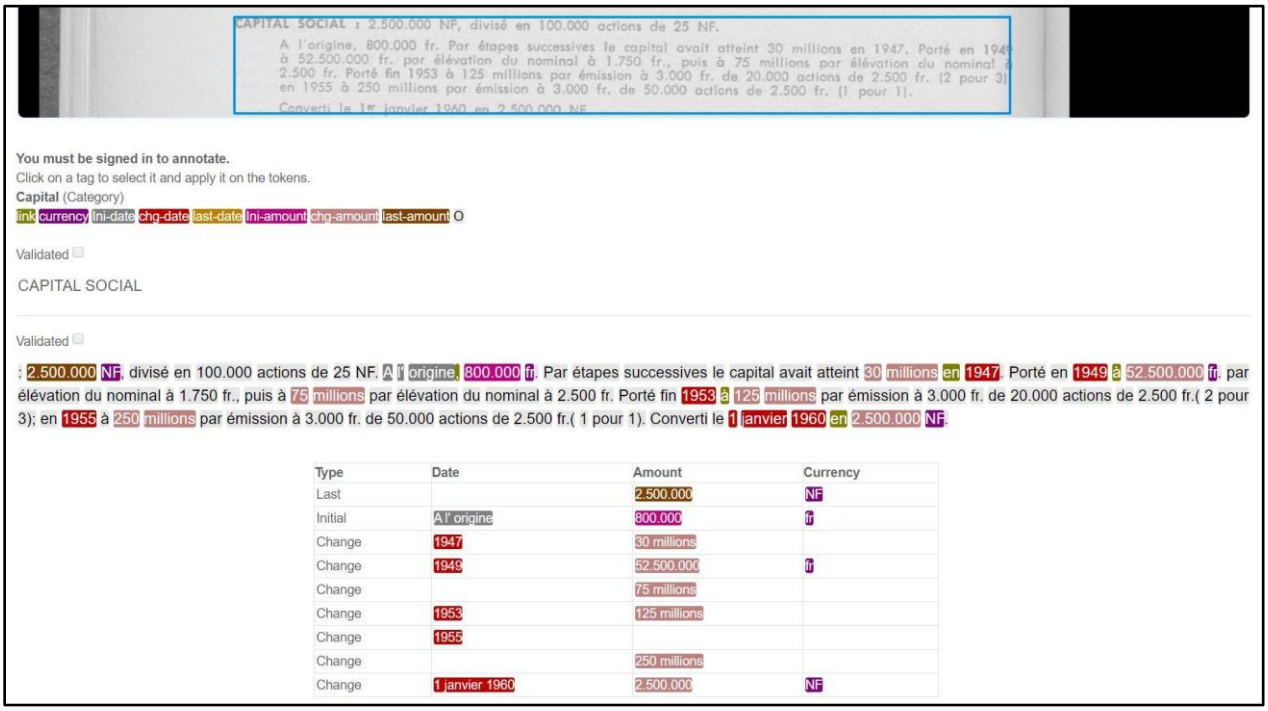


Figure 22: Visualisation of the correction interface for a textual section

The step by step to manipulate the correction interface is the following (together with the next figure):

1. Click on a desired tag that will be used for marking the words.
2. Click on the words to change, the color box will change and also a black thin border will show the words that have been corrected
3. Click on the save icon in the lower part of the page in order to apply the changes
4. Visualise the new correction in the extraction table just below the text

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

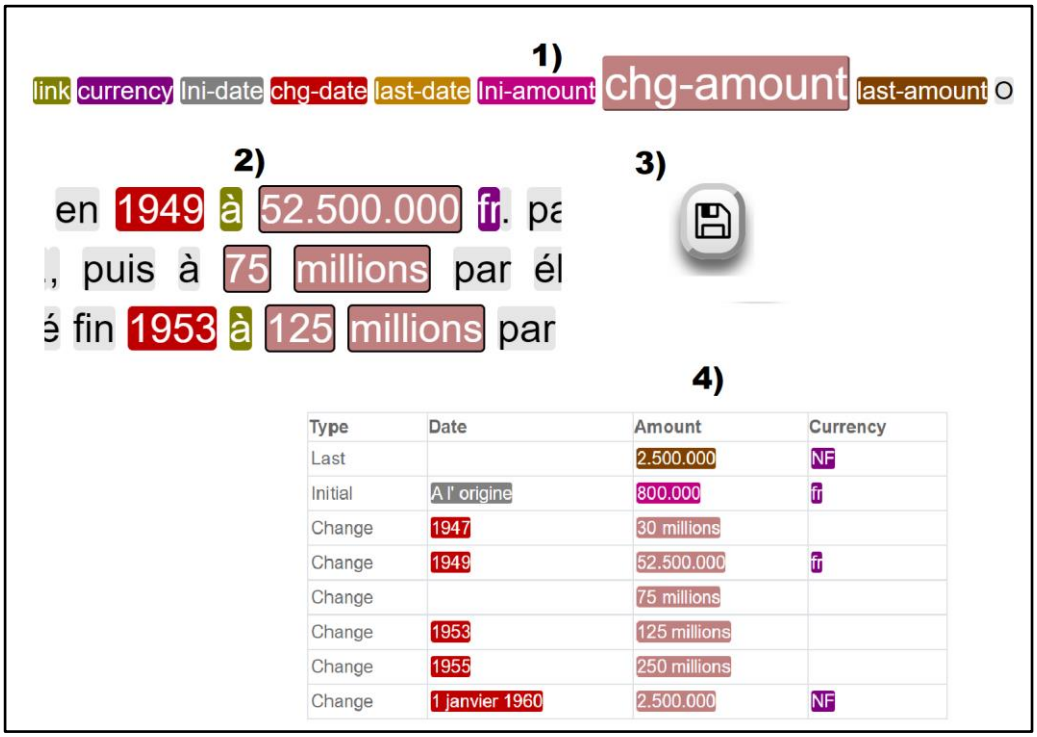


Figure 23: Use of the correction interface

Visualisation of a tabular section

In the document section page, the user will also be able to click on sections containing tabular information. This information is contained in a table and structured in a special way for regrouping the different types of data in it. The information will be grouped depending on the type of table. In the following example, the table in question is the BalanceSheet table from the Desfossés French Yearbook. The table is divided first by Assets and Liabilities, then each part will contain the different items that belong to it. So far now we can have an Asset from the Assets Part, and then this item will hold the different amounts of money for it, in the different years that the table shows. Inside the correction interface, the user will be able to visualise the different entries for that item and also visualise the table of information to be extracted.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

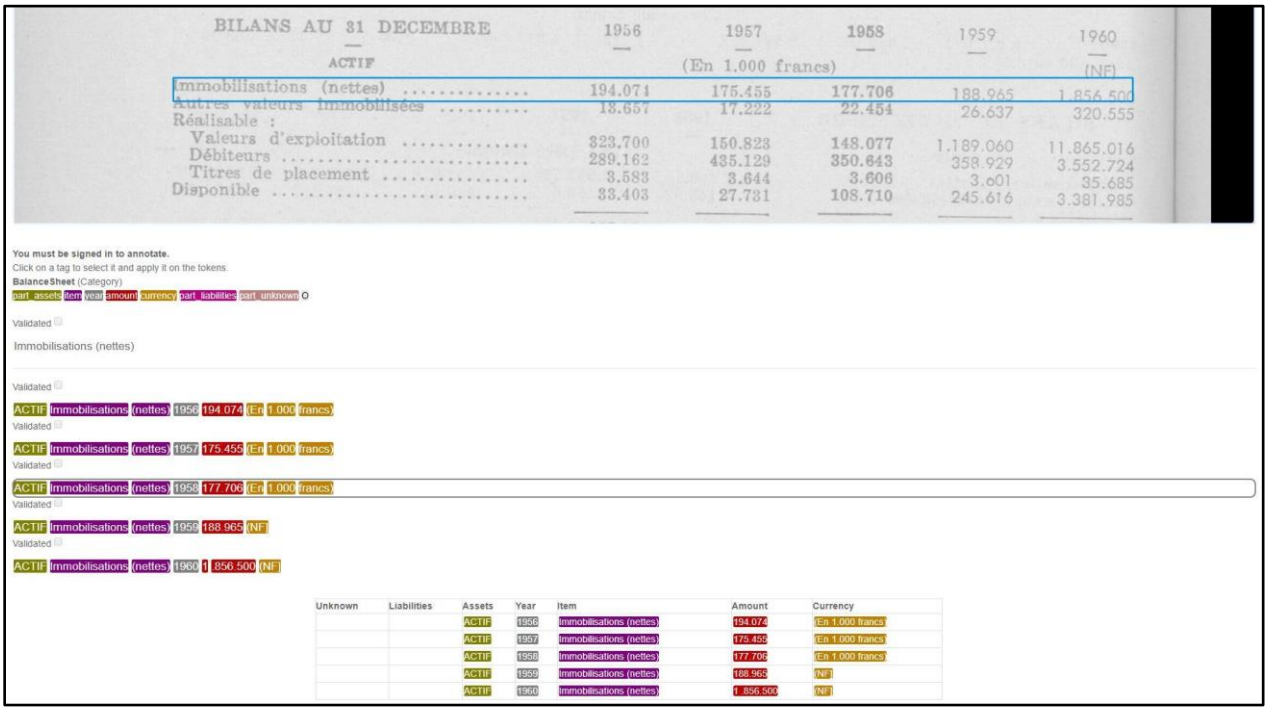


Figure 24: Visualisation of the correction interface for a table section

Towards a more friendly interface

This interface was recently developed and so it solves the biggest problems we had before for navigating through the data, visualising it and correcting it. However, no user group has been asked for feedback or new ideas for this interface, and so the next step is to evolve this interface towards the most friendly possible for the user and so it should be validated by the users who will depend on it to analyse the different data from the yearbooks.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

3 Price list data extraction system

The second system focuses on data extraction into price lists. This system is decomposed on two main tasks: the document structure recognition using a cross validation module, and the definition of a general-purpose text recognizer (OCR). In the context of the ANR project HBDEX (Exploitation of Big Historical Data for the Digital Humanities: application to financial data), we develop a system to extract data in price lists. This system has been first tested on price lists that come from Paris unofficial market: "la Coulisse". In the context of the EurHisFirm project, we generalise this work on official price lists from Paris and Brussels.

3.1 Document Structure Recognition (Task 7.2)

3.1.1 System

The first step of our system is a structural analysis of pages. This structural analysis is done with a combination of deep-learning and syntactic approaches. In order to localize text lines within the page, we use an existing system based on deep learning, called Aru-Net [Grüning 2018]. Aru-Net is a fully convolutional network which follows a U-net architecture with residual blocks. Aru-Net produces images in which each pixel has a probability of belonging to a text-line (Figure 25 (b)). Text-lines are then extracted from the probability maps produced by the network thanks to simple filtering operations (gaussian filter and hysteresis thresholding).

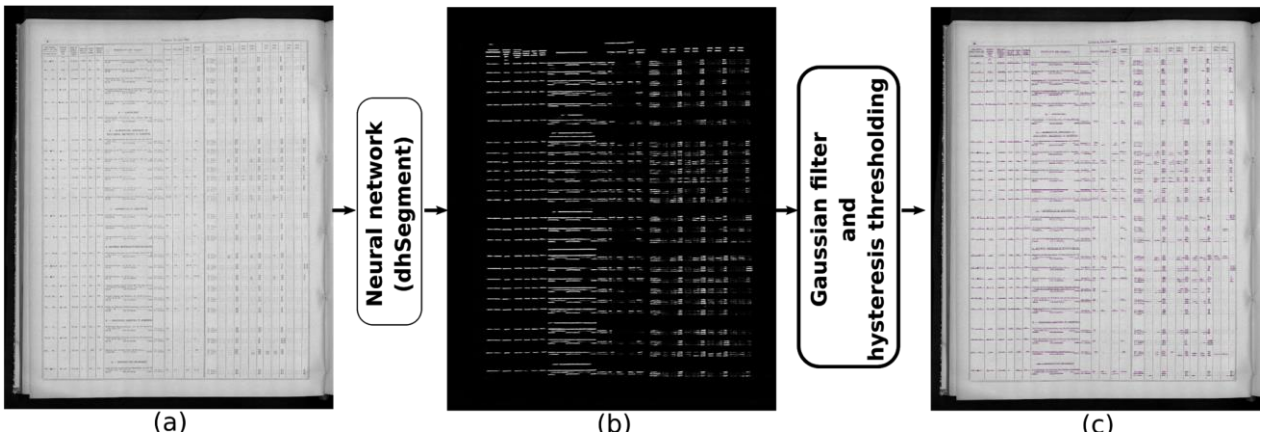


Figure 25: Example of text line extraction in a stock price lists document. (a) Image to be processed - (b) Probability map for text-line in this image - (c) Extracted text-lines

We then use the localised text-lines and vertical rulings (extract with a Kalman filter) as terminals of a grammatical description. We describe the price-list structure in a general way. Our description can be applied on documents of "La Coulisse" (Paris unofficial price-list), Paris official price-list ("Le Parquet") and Brussels price-list. In Figure 26 one can see the results of the structural analysis.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

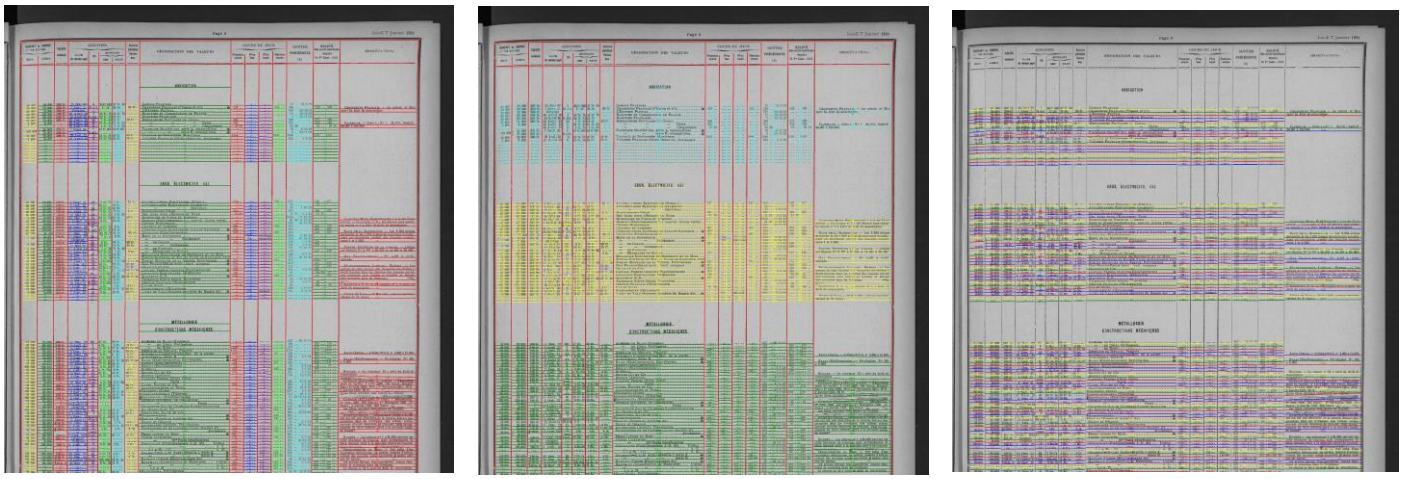


Figure 26: Columns localisation - Section localisation - Table row localisation

In order to be generic enough, we do not precise the number of columns expected or physical indications like the width of each column in our description. However, in noisy documents, our grammatical description makes errors.

On Figure 27, one can see that the error produced on the 29th June 1899 can be corrected if we consider the context of the days before and the days after.

Figure 27: Columns localisation - Section localisation - Table row localisation

To take advantage of the sequentiality of the collection and correct errors in noisy documents, we design a global strategy. Our global strategy is based on an iterative process (see Figure 28). The aim of each iteration is to recognize and validate a structural element of the documents: columns, sections, stock names (table entry), other fields.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

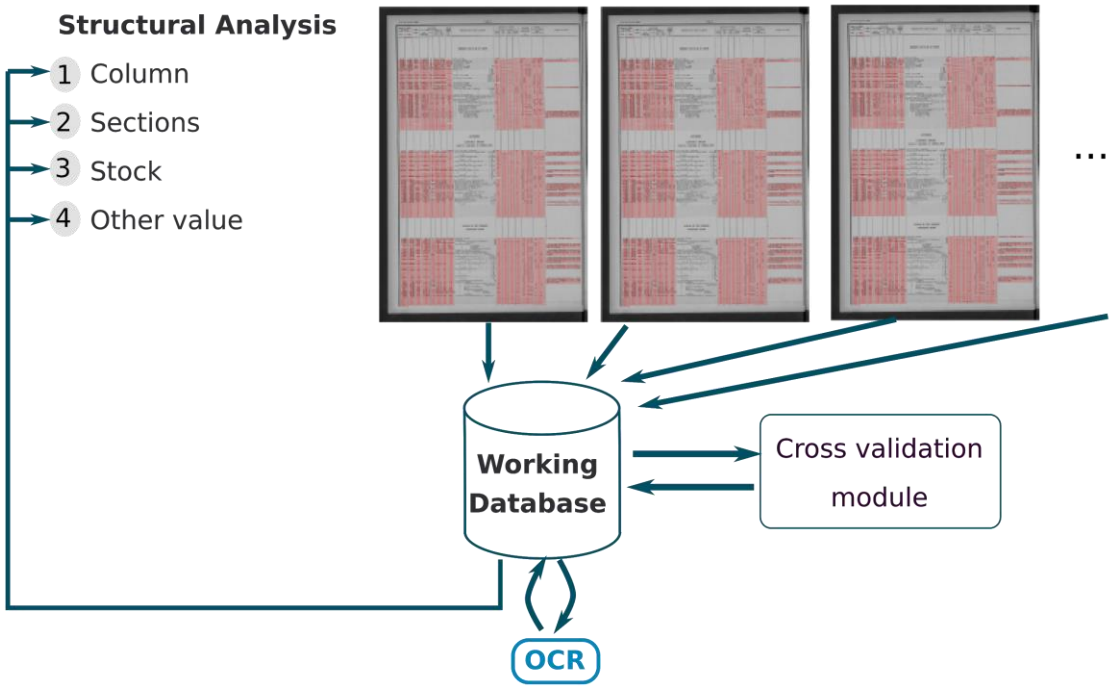


Figure 28: Overview of the global strategy

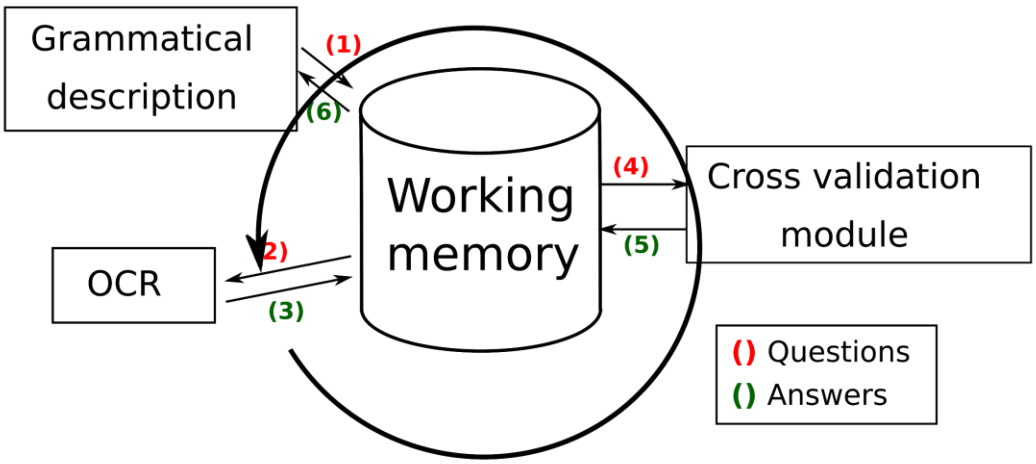


Figure 29: Interaction between the modules of the strategy

An iteration is composed of different steps (see Figure 29):

- **(1)** A first structural analysis of each document. This analysis relies on a combination of deep-learning and syntactical approach (see Figure 26). With this analysis, we extract different knowledge from the documents and produce questions for the other modules (OCR and cross-validation module) with an interaction formalism [Chazalon 2011].

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

- **(2)**/**(3)** The OCR answers questions about the transcription of text-lines localized with the grammatical description.
- **(4)**/**(5)** The cross-validation module takes all structural knowledge obtained with the grammar as input. Then, the module processes the knowledge together in order to determine some properties: the expected width and the expected title of each column.
- **(6)** The answers of the different questions are injected in the grammatical description.

3.1.2 Evaluation

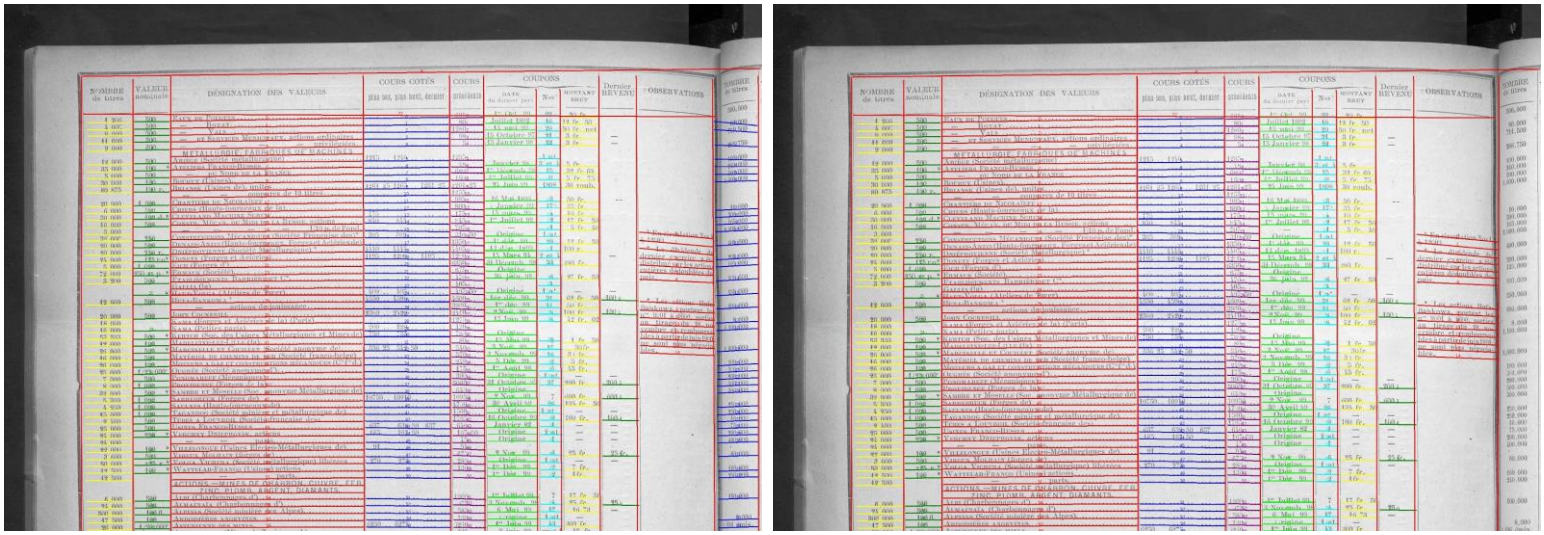
In this section we present some qualitative results of the page analyzer (grammatical description) on Paris official market ("Le Parquet") and on Brussels official lists. Quantitative results of two experiments done on Paris unofficial price-lists in the context of the HBDEX project are presented in the deliverable D7.2. These experiments show the interest of our global strategy.

A. Evaluation on "La Coulisse"

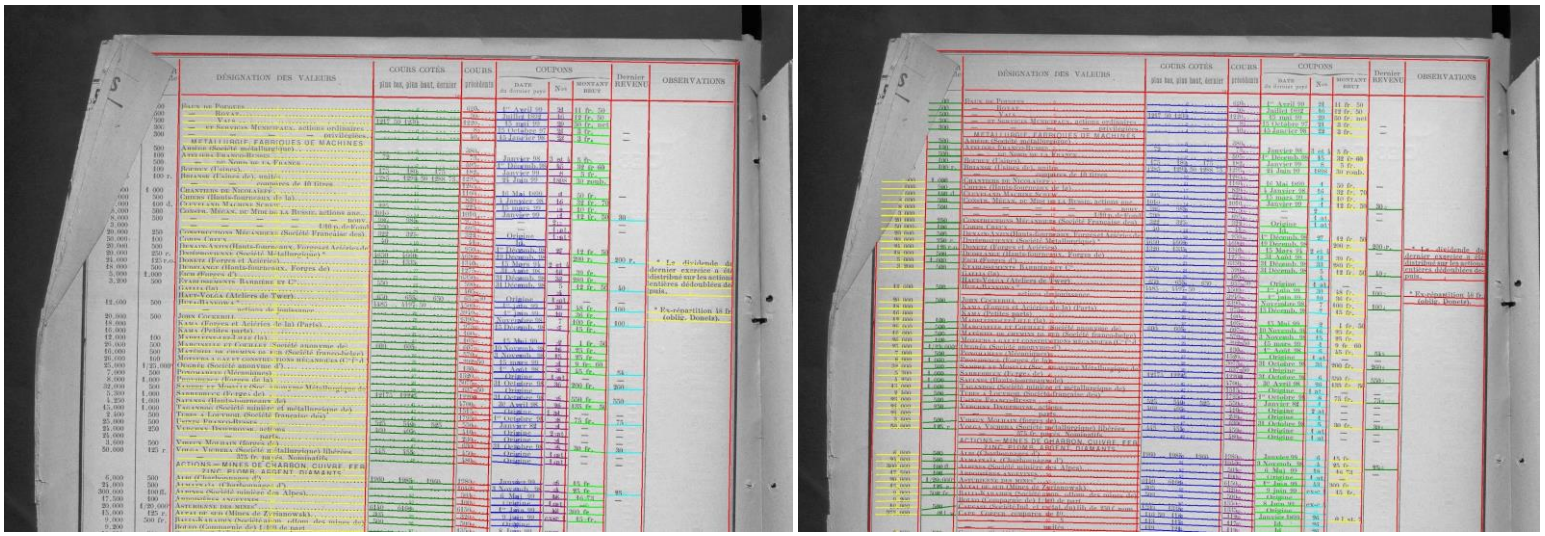
We test our global strategy on 2 subset of the collection of "La Coulisse" (see deliverable D7.2). On Figure 30 some qualitative results are presented. The correction/validation of columns is important to continue the analysis without accumulating errors. Notably the errors done on example of Figure 30(b) can have important consequences on the next steps of the analysis because all columns are shifted. Therefore, without an automatic correction, a wrong treatment (ex: wrong language model for text recognition) will be applied on each column.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489



(a)



(b)

Figure 30: Qualitative results: Improvement obtained with our strategy
(a)Left: one column from the next page is detected - (a)Right: with our strategy, it is not in the table
(b)Left: two first columns not detected due to folding - (b)Right: with our strategy, they are detected

So far, we apply and test the cross-validation mechanism on columns recognition. In further work, we will apply a similar process on the other elements we want to extract: sections, stocks names, other fields.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

B. First evaluation on Paris and Brussels official lists

(a) **(b)**

Figure 31: Results of table structure recognition on Paris (a) and Brussels (b) price-lists

On Figure 31, one can see the first results of the grammatical description on price-lists from Paris on Figure 31(a) and Brussels on Figure 31(b). In further work, we will adapt the grammatical description to take into account the specificity of these collections: recognition of double tables, recognition of sections titles that cross tables rulings, recognition of stocks of several table rows… This adaptation will be done using the same approach as for yearbooks: a general description of price lists including specificities from one corpus as light as possible to guarantee a fast adaptation of the system to a new corpus.

3.2 General-purpose text recognizer (OCR) (Task 7.3)

3.2.1 System

We design our own Deep Learning based-OCR platform. It is built on convolutional neural networks (CNN) combined with bilateral recurrent Neural Networks layers. Training is performed using the CTC (Connectionist Temporal Classification) loss function. Outputs can be parsed (Viterbi Beam Search) using different language models depending on the context of use within the document. Language models can

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

encode lists of possible stock names, or the syntactic rules used to write specific information such as prices, dates, etc.

This description follows the model declared as well as the language model in the *deliverable D7.1: general software libraries*. Since the system uses machine learning to understand the written context of the images, new data had to be given to the OCR as a way for it to see more examples and have a wider understanding of the data. The images treated in this project are a group of different papers and fonts that were used at the time of the creation of the original paper data, and thus making it harder for the OCR to be able to understand certain texts in different scenarios.

3.2.2 First instance of the system

The OCR was first trained on pages from the *French Pricelist La Coulisse*. This was the first data to be available for use since the images used for its training came from the segmentation of the system Task 7.2. These pages are also in use for the French national project ANR "HBDEX: Exploitation of Big Historical Data for the Digital Humanities: application to financial data".

The advantage of this first data is to make our OCR perform very accurately on the different types of papers, fonts and characters of this corpus. This strategy will reward us with a generic OCR which will not be specialized and thus ready for adaptation of other corpuses. If we started from zero for every corpus, we would have undergone the same procedure of annotating a big volume of data for the OCR to start behaving correctly, and only in that corpus. Having a generic OCR from the beginning allows us to shift into any corpus easily with the least amount of data possible. This way we will be able to easily detect any problems in a new corpus, annotate a few pages and quickly adapt the OCR towards the new context.

3.2.3 Towards a more generic system

The first data used to train the OCR was produced from a Commercial OCR that is suited for modern papers and modern fonts. Therefore, any produced data had a small number of mistakes which gave the OCR a better performance but not perfect. In order to tackle this problematic, hand-labelled data was needed to be produced. Using the segmentation from Task 7.2, two years were treated for testing and further learning: 1899 and 1924.

During this exercise, the first year treated was 1924. The first version of the OCR trained from this data and reached a common problem called "over-learning". This causes the OCR to learn a new sort of data so well, that when used in the same context, it gets almost perfect accuracy but then when tested in other types of images, its accuracy decreases.

The next step was to begin the annotation of another year: 1899. The OCR had errors in the recognition of characters in this year and so a new annotation was made for this year and so a new dataset for the OCR to learn and become more general towards the analysis of the images that are given to it.

The hand-labelled data produced for the new learning iteration was:

- 5,000 line images from the Coulisse 1899 papers
- 70,000 line images from the Coulisse 1924

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

This new datasets and models can be visualised in the following figure:

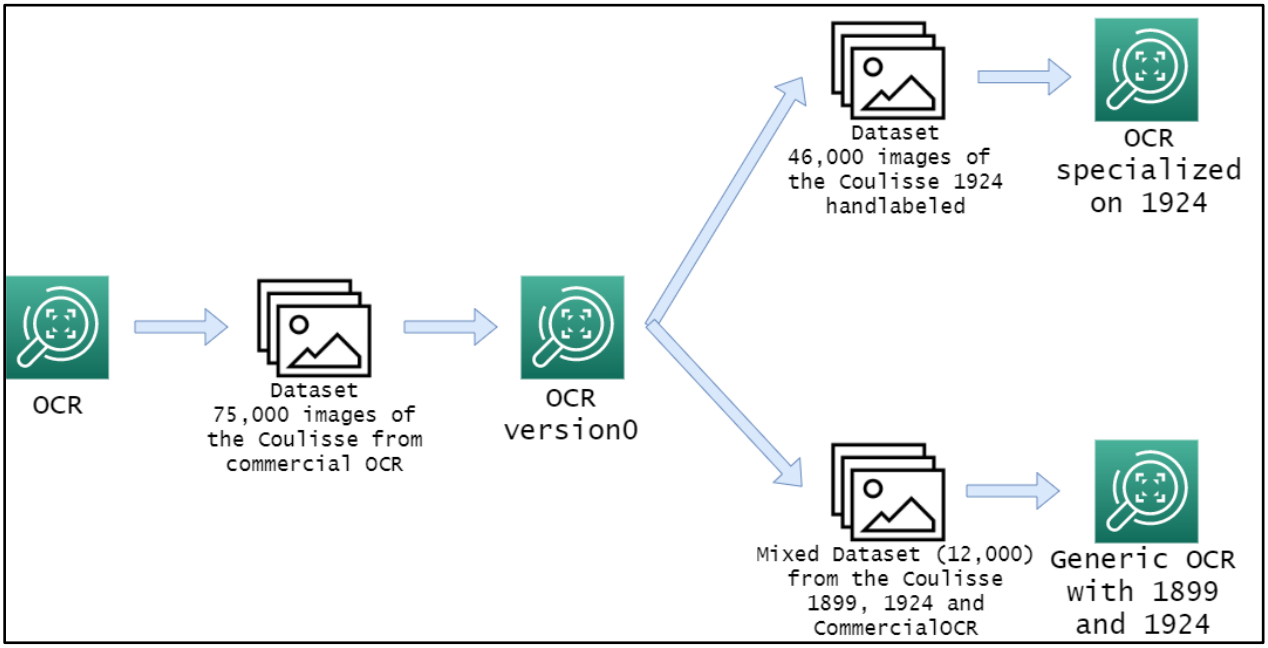


Figure 32: Learning strategy for the OCR with the new datasets

Size of data

For the latest Mixed dataset that contains around 12.000 images, it is important to remember that one image is equal to a single line of text entry in a column of a page, so a Price list page could contain from 700 to 1800 entries (and so images).

Thus, for the creation of a dataset, there will be a need for annotating at least 12 pages.

Conclusion

From these experimentations, we have concluded that the best strategy to adopt in order to have a precise OCR is to show as many different examples as possible to the system for it to learn all the different variations of the pages, characters, fonts and interpretations. This includes the task of exploring different corpuses, checking for ruptures in the type of images and pages and creating a sufficient volume of data to be labelled by hand. Also, in case of further specialisation it should be kept in mind that it will be needed to annotate some more pages in order to adapt the system to a more specific context that is not considered as general as the others.

3.2.4 Annotation Interface

A new interface was deployed in order to facilitate the correction of images. The segmented images are being transformed into a special format of XML files called ALTO and METS. These formats are used for document descriptions which include the positioning of lines and text, columns or rows, paragraphs or sections. Also, the logic point of view which translates to the structuring of the information that is going to be visualised in the interface. The combination of both is what makes the annotating interface have a

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

precise manipulation of data while also giving it a logical point of view required to read the document as it is intended to be.

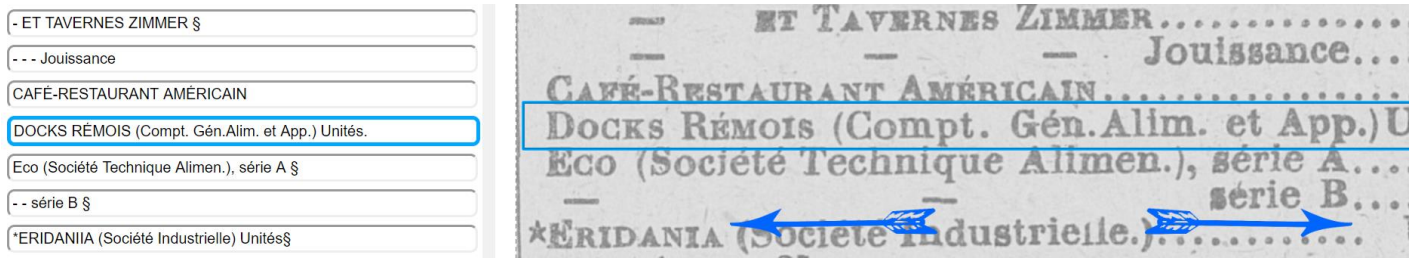


Figure 33: The correction interface for hand labelling data

The interface also comes with a user login feature in order to track changes and also allow more features like error detection in the chosen document and easier correction of data. Anonymous users can also correct the data, however they will have to enter a Captcha in order to prove that they're human.

At any point, the data corrected in the interface can be retrieved in order for it to be used for evaluating purposes, improving the dataset for later learning, or well for direct learning to the OCR model.

Figure 34: Day after and day before feature

In case of doubt in the page, it is always a good idea to compare the same item in the pages of the next day or the day before. In order to allow the user easy access to simultaneous pages, the interface allows the user to click on the desired entry and check the other pages in order to facilitate the correction of the item.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

4 References

[Akbik 2018] Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638{1649 (2018)

[Ares Oliveira 2018] S. Ares Oliveira, B.Seguin, and F. Kaplan, "dhSegment: A generic deep-learning approach for document segmentation," in Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on, pp. 7-12, IEEE, 2018.

[Chazalon 2011] J. Chazalon, B. Couasnon, and A. Lemaitre. Iterative analysis of pages in document collections for efficient user interaction. In International Conference on Document Analysis and Recognition, 2011

[Diem 2017] Diem, M., Kleber, F., Fiel, S., Grüning, T., & Gatos, B. (2017, November). cbad: Icdar2017 competition on baseline detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 1355-1360). IEEE.

[Galibert 2014] O. Galibert, J. Kahn, I. Oparin. (2014, October). The ZoneMap metric for page segmentation and area classification in scanned documents. IEEE.

[Grover 2008] Grover, C., Givon, S., Tobin, R., Ball, J.: Named entity recognition for digitised historical texts. In: LREC (2008)

[Grüning 2018] Grüning, T., Leifert, G., Strauß, T., & Labahn, R. (2018). A two-stage method for text line detection in historical documents. *arXiv preprint arXiv:1802.03345*.

[Huang 2015] Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)

[Laferty 2001] Laferty, J., McCallum, A., Pereira, F.C.: Conditional random elds: Probabilistic models for segmenting and labeling sequence data (2001)

[Lample 2016] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)

[Settles 2009] B. Settles, Active Learning Literature Survey.: A Computer Sciences Technical Report, University of Wisconsin–Madison, 2009.

[Swaileh 2020] W. Swaileh, T. Paquet, Sebastien Adam1, and Andres Rojas Camacho1, A Named Entity Extraction System for Historical Financial Data, Document Analysis Systems (DAS), accepted, 2020.

This project has received funding from

the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489