EURHISFIRM - M7.2: Final version of the data extraction system

Sébastien Adam, Simon Bouvier, Bertrand B. Coüasnon, Nathalie Girard, Camille Guerry, Aurélie Lemaitre, Iwan Le Floch, Thierry Paquet, Charles Quéguiner, Yann Ricquebourg, et al.

Long-term data for Europe

EURHISFIRM

M7.2: Final 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	M7.2: Final version of the data extraction system
Due Date of Deliverable	Month 39, 30/06/2021
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)
Nathalie GIRARD (UNIVERSITÉ DE RENNES 1, 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)
Charles QUÉGUINER (UNIVERSITÉ DE RENNES 1, INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Yann RICQUEBOURG (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Andres ROJAS CAMACHO (UNIVERSITÉ DE ROUEN NORMANDIE)
Achille FEDIOUN (UNIVERSITÉ DE ROUEN NORMANDIE)
Wassim SWAILEH (UNIVERSITÉ DE ROUEN NORMANDIE)

APPROVED IN 2021 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

1 Introduction 5

2 Yearbook Information Extraction system 5

2.1 Text Block Selection by Document Structure Recognition (Task 7.2) 6

2.1.1 System 6

2.1.1.1 Evolution in structural analysis of yearbooks 6

2.1.1.2 New yearbook: Madrid 1931 8

2.1.1.3 Genericity 12

2.1.2 Evaluation results 12

2.1.2.1 Evaluation of bounding boxes (ZoneMap) 12

2.1.2.2 Evaluation of table analysis 14

2.2 Information Extraction System (Task 7.3) 15

2.2.1 System overview 15

2.2.2 Processing the Spanish 1931 Madrid yearbook 16

2.2.3 Evaluation on all corpuses 19

2.2.3.1 Annotated data for the French Yearbook 19

2.2.3.2 French Yearbook Desfossés 1962, per rubric Tag evaluation 19

2.2.3.3 Annotated data for the German Yearbook 22

2.2.3.4 German Yearbook Handuch 1914-1915, per rubric Tag evaluation 22

2.2.3.5 Annotated data for the Spanish Yearbook 24

2.2.3.6 Spanish 1931 Madrid Yearbook, per rubric Tag evaluation 25

2.2.3.7 Global results in yearbooks 26

2.3 Application of the Yearbook Information Extraction System to three yearbooks 28

3 Price List Extraction system 33

3.1 Strategy and transversal analysis 33

3.2 Global Meta Table Structure: tables of tables 34

3.2.1 System 34

3.2.2 Evaluation results 36

3.3 Generic description of table content 38

3.3.1 System 38

This project has received funding from

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

3.3.1.1 Genericity 38
3.3.1.2 Evolution in structural analysis 39
3.3.2 Evaluation result 41
3.3.2.1 Columns & Headers recognition and validation 41
3.5 General-purpose text recognizer (OCR) 43
3.5.1 System overview 43
3.5.2 Data augmentation 44
3.5.2.1 Erosion/Dilatation 45
3.5.2.2 Lightening / Contrast modification 45
3.5.2.3 Change of Resolution (DPI change) 45
3.5.2.4 Elastic Distortion 45
3.5.2.5 Gaussian noise 45
3.5.2.6 Bounding Box modification 45
3.5.2.7 Image sharpness modification 45
3.5.2.8 Evaluation 45
3.5.3 Final experiment results 46
3.5.3.1 Datasets and Evaluation 46
3.6 Evaluation and Application of the Price List Information Extraction System to four price lists 49
3.6.1 Evaluation of Stock identification on Paris “La Coulisse” 1899, 1696 stock lines 49
3.6.2 Data Extraction on 6 months of Paris “La Coulisse” 1899 price lists, 536 pages 50
3.6.3 Application of the Price Lists Information Extraction System on EurHisFirm dataset 51
4 WP7 Information Extraction System demonstration 56
5 References 57

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 sources related to listed companies such as yearbooks and exchange lists.

In the previous 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. Deliverable D.7.2 was composed of the first version of two recognition systems: one for yearbook information extraction, and one for price lists data extraction. This deliverable D.7.4 contains the final versions of the two recognition systems with the final evaluation of those systems. The yearbook information extraction system is presented in section 2 and the price list extraction system is presented in section 3. Those systems have been applied to several corpora: the German Yearbook 1913-1914 (Handbuch der deutschen Aktiengesellschaften), the Spanish Yearbook Madrid 1931, the French Desfossés Yearbook 1962, the official price lists Brussels 1912, Madrid 1931 and Paris 1961-1962, which are the document samples dataset validated by the Steering Committee. We also present an evaluation of the complete Information Extraction System on price-lists from Paris "La Coulisse" 1899, made in the French National project ANR HBDEX. All the data extracted has been also inserted in the DFIH database. This experimentation shows the strong interest of the developed system to drastically reduce the need of human experts manual validation while generating high quality of data.

2 Yearbook Information Extraction system

We implemented a generic pipeline of processes that can run similarly on the various yearbooks considered within the consortium (see *Figure 1* below). Inputs of the pipeline are images of documents and outputs are information attached to each company reported in the Yearbook. Information extracted are 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 list 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

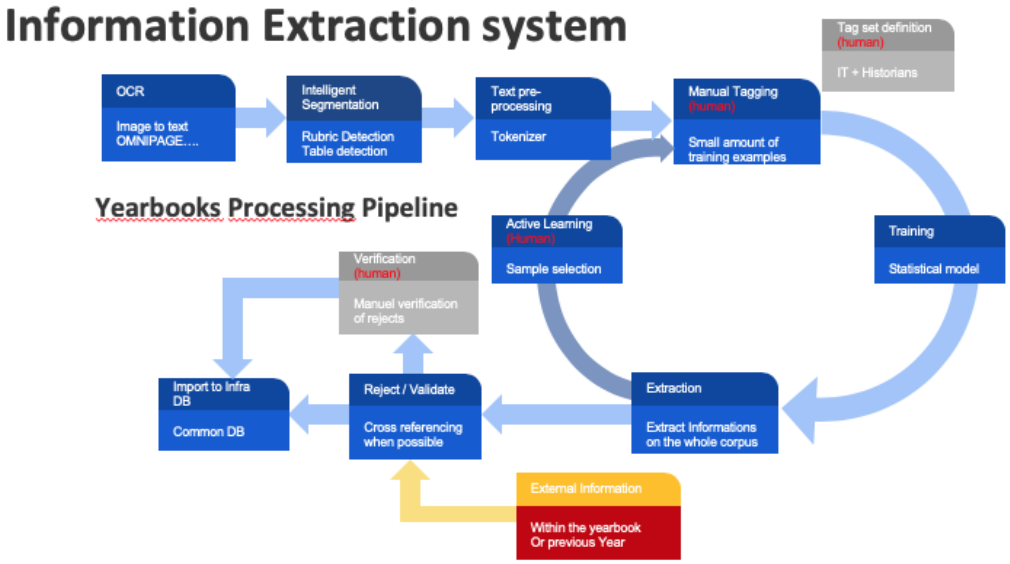


Figure 1: Overview of the generic Information extraction pipeline in yearbooks.

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

2.1.1 System

We built a generic structure recognition system for yearbooks. It is based on the extraction of graphical components such as table separator recognition, with or without existing borders, contextual segmentation of text lines, rubric header detectors, text alignments detectors… These detected elements are then used in a generic way by a bi-dimensional grammatical description of the yearbook layout structure. They are combined to produce a structural analysis of the yearbook organization into text paragraphs (title, subtitle, rubrics) and complex table structures.

In this section, we will show the evolution of the libraries and methods of analysis of yearbooks we presented in D7.2. Then we will talk about the integration of a new yearbook and what we did to facilitate adding even more in the future.

2.1.1.1 Evolution in structural analysis of yearbooks

In D7.2, we briefly exposed the structure of an issuer, which stays the same across different yearbooks: a main title contains the name of the issuer, followed by several rubrics composed of a title and a content.

To make it more flexible, and easier to switch between yearbooks, we can now define the specificities of the rubrics titles in the different yearbooks. We already saw in D7.2 that rubrics titles in the German yearbook (Handbuch 1914-1915) were extracted by a neural network. We can now have a totally different way of extracting titles in the French yearbook (Desfossés 1962) and still have the exact same structure for all rubrics, as the content is the same format (only text lines).

This project has received funding from

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

In the Desfossés yearbook, all rubrics titles are aligned on the left (see Figure 2), and we can detect this alignment thanks to the detector we presented in D7.2.

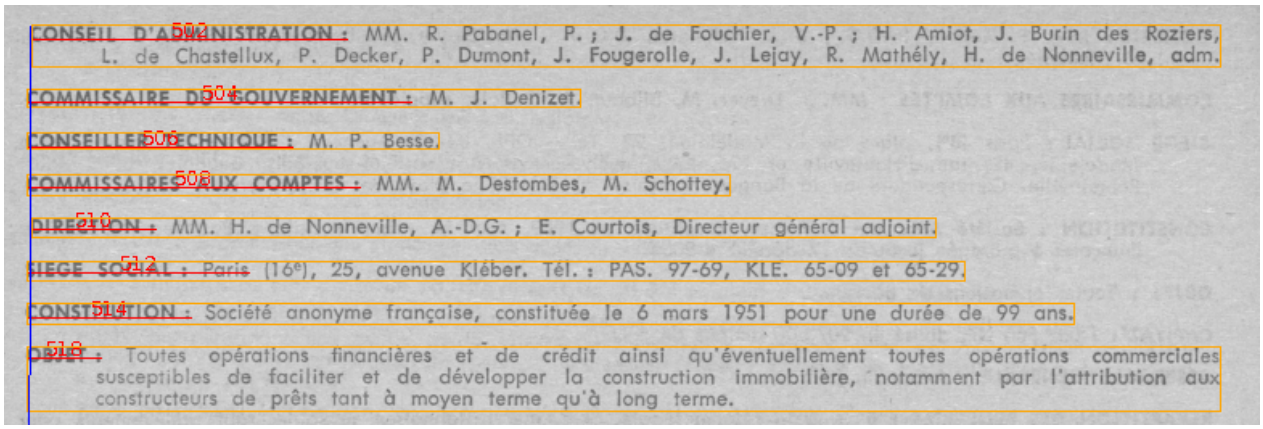


Figure 2: In blue, the alignment on the left used to define rubrics titles (in red)

Based on the analysis of balance sheets (see D7.2), we added the analysis of other more generic tables in the French yearbook.

These tables can have different columns of different sizes, but still have a lot in common with the balance sheet regarding their structure, especially that they have no borders.

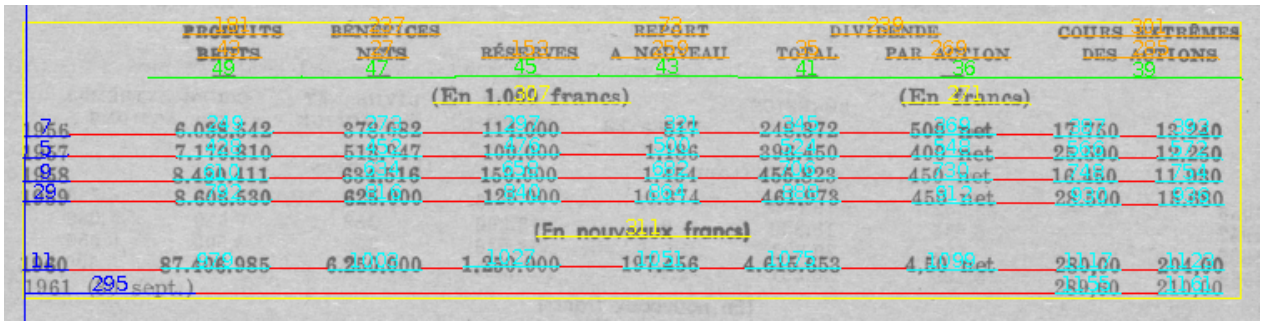


Figure 3a: Generic table in the Desfossés yearbook

One of their particularities is the possibility to find intermediate rows, often to specify a new currency for the different columns.

In that case, the currency found in the original header is replaced by this new one, still based on the nearest for each column (see Figure 3a).

To improve the detection and sizing of table columns and cells, we added a new separator to each column that indicates its width and the local slope of the text lines in the column (see Figure 3b).

Figure 3b: In green, the separators of the columns in the header

This project has received funding from

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

One of the direct applications of this feature is to be able to correct the slope of the text lines inside cells, based on the slope of the separator of their respective column (see Figure 4).

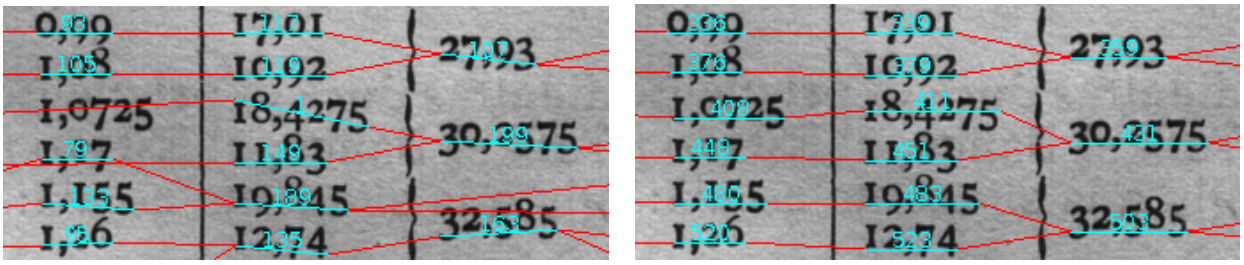
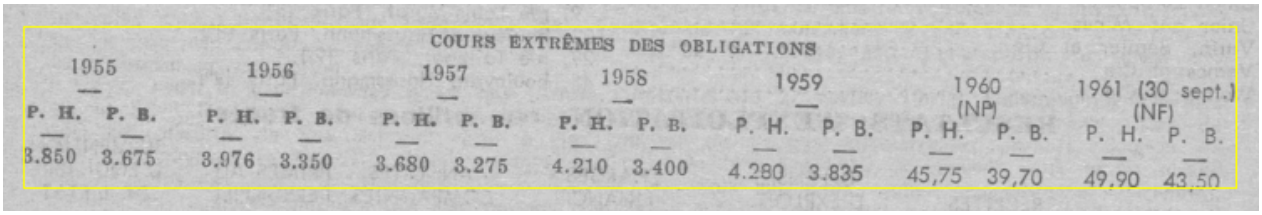


Figure 4: Comparison of cells lines before and after correcting the slope

Another type of tables found in the Desfossés are the extreme courses tables (see Figure 5), which we do not want to extract in detail because they do not present crucial information. However, they are important for the structure of the whole page, so we only detect their bounding boxes.



COURS EXTRÊMES DES OBLIGATIONS

1955		1956		1957		1958		1959		1960 (NF)		1961 (30 sept.) (NF)	
P. H.	P. B.	P. H.	P. B.	P. H.	P. B.	P. H.	P. B.	P. H.	P. B.	P. H.	P. B.	P. H.	P. B.
3.850	3.675	3.976	3.350	3.680	3.275	4.210	3.400	4.280	3.835	45,75	39,70	49,90	43,50

Figure 5: Extreme courses table

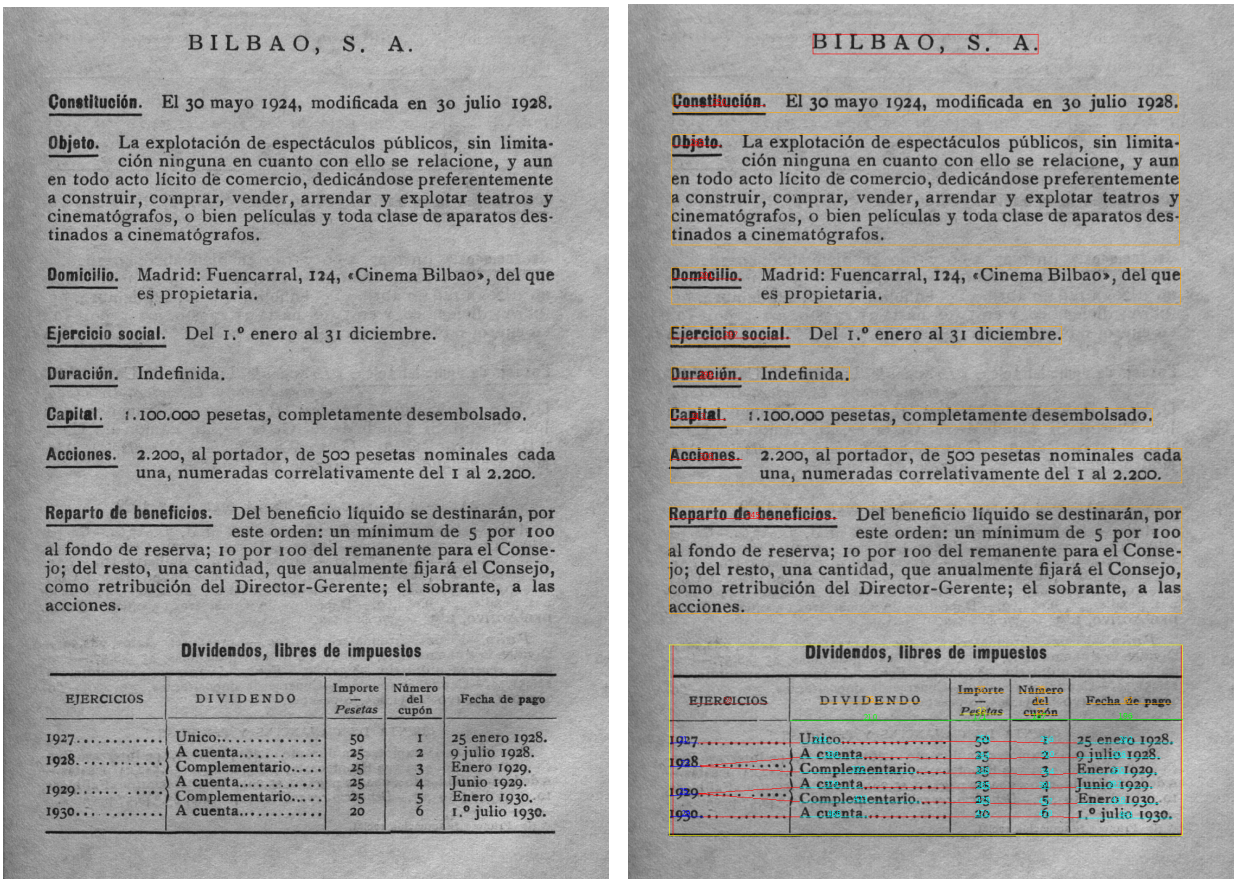
There are some of the improvements we made over the year regarding the analysis of French yearbooks, and most of them also come to extend the library of document structure elements, like table structures without borders, which could be used in other yearbooks than the Desfossés. It is a way to be able to adapt our Information Extraction System by assembling different kinds of elements which will be found in a new yearbook corpus.

2.1.1.2 New yearbook: Madrid 1931

This year, we also worked on a completely new type of yearbook with a Spanish one (Madrid 1931; see Figure 6). We could reuse some of the elements we previously created, and had to develop new ones for other specific elements.

This project has received funding from

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



BILBAO, S. A.

Constitución. El 30 mayo 1924, modificada en 30 julio 1928.

Objeto. La explotación de espectáculos públicos, sin limitación ninguna en cuanto con ello se relacione, y aun en todo acto lícito de comercio, dedicándose preferentemente a construir, comprar, vender, arrendar y explotar teatros y cinematógrafos, o bien películas y toda clase de aparatos destinados a cinematógrafos.

Domicilio. Madrid: Fuencarral, 124, «Cinema Bilbao», del que es propietaria.

Ejercicio social. Del 1.º enero al 31 diciembre.

Duración. Indefinida.

Capital. 1.100.000 pesetas, completamente desembolsado.

Acciones. 2.200, al portador, de 500 pesetas nominales cada una, numeradas correlativamente del 1 al 2.200.

Reparto de beneficios. Del beneficio líquido se destinarán, por este orden: un mínimum de 5 por 100 al fondo de reserva; 10 por 100 del remanente para el Consejo; del resto, una cantidad, que anualmente fijará el Consejo, como retribución del Director-Gerente; el sobrante, a las acciones.

Dividendos, libres de impuestos

EJERCICIOS	DIVIDENDO	Importe — Pesetas	Número del cupón	Fecha de pago
1927	Unico	50	1	25 enero 1928.
1928	A cuenta	25	2	9 julio 1928.
	Complementario	25	3	Enero 1929.
1929	A cuenta	25	4	Junio 1929.
	Complementario	25	5	Enero 1930.
1930	A cuenta	20	6	1.º julio 1930.

Figure 6: Analysis of a page of the Madrid yearbook

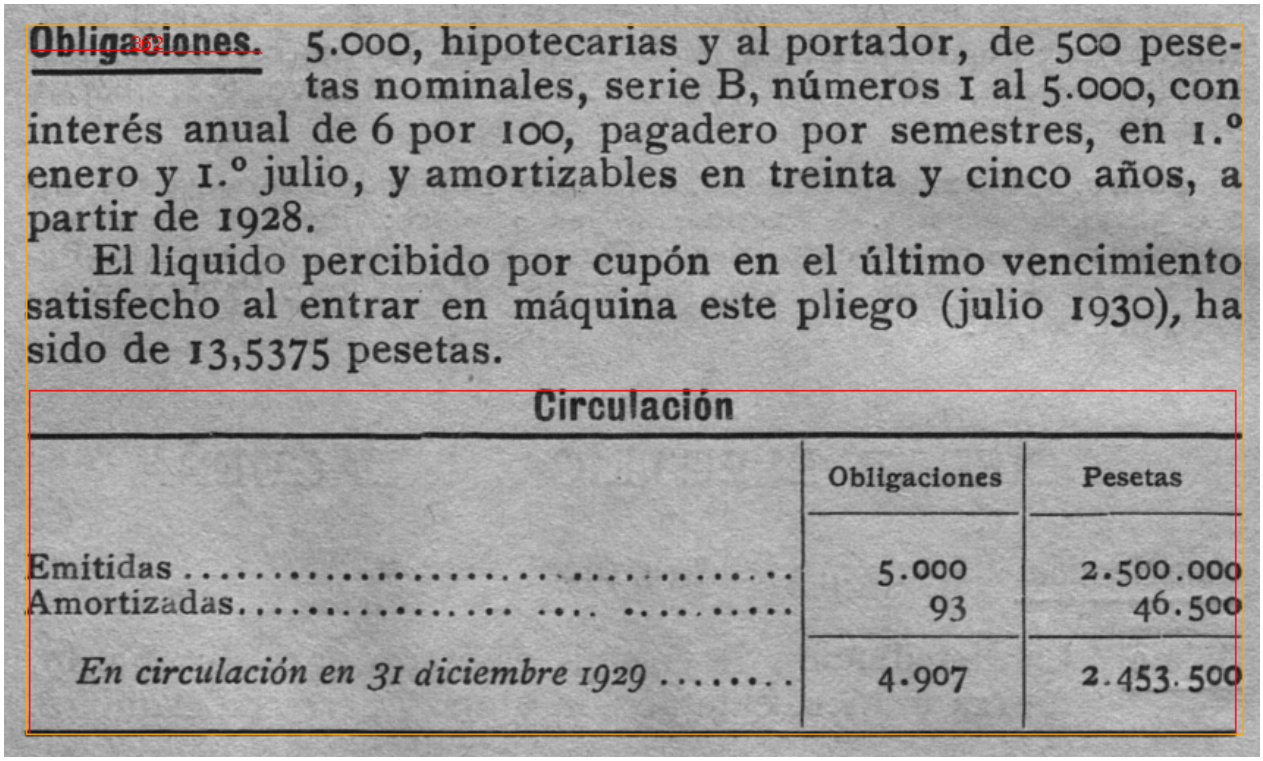
In D7.2 we talked about how we used the neural network ARU-Net [Grüning 2018] for text-lines detection. Due to the difference of scanning quality and luminosity between this new yearbook and the two others, we started by training a new model. By adding Spanish images to the training corpus, we greatly improved the network's performance on this yearbook, especially reducing the number of false-positives due to the transparency of the page.

In the Madrid yearbook, pages follow the same global structure for issuer titles and textual rubrics. As exposed in section 2.1.1.1, by specifying a new way of detecting rubrics titles, we can already extract a simple list of rubrics.

However, a difference here is the possibility to have tables between or inside rubrics (see Figure 7), so we had to adapt and improve the rubrics detector, so they can contain tables and not only text-lines.

This project has received funding from

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



Obligaciones. 5.000, hipotecarias y al portador, de 500 pesetas nominales, serie B, números 1 al 5.000, con interés anual de 6 por 100, pagadero por semestres, en 1.º enero y 1.º julio, y amortizables en treinta y cinco años, a partir de 1928.

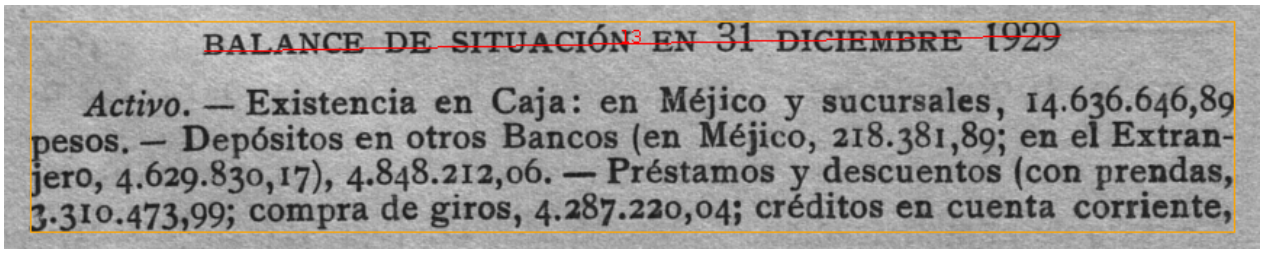
El líquido percibido por cupón en el último vencimiento satisfecho al entrar en máquina este pliego (julio 1930), ha sido de 13,5375 pesetas.

Circulación

	Obligaciones	Pesetas
Emitidas	5.000	2.500.000
Amortizadas	93	46.500
En circulación en 31 diciembre 1929	4.907	2.453.500

Figure 7: Example of a table inside a rubric

Some rubrics don't have the same format of title, and we have the possibility to define more than one description for rubrics titles in each yearbook. Here (in Figure 8) we have a different title for the balance rubrics, which are equivalent to balance sheets in the Desfossés, but here are rubrics instead of tables.



BALANCE DE SITUACIÓN EN 31 DICIEMBRE 1929

Activo. — Existencia en Caja: en Méjico y sucursales, 14.636.646,89 pesos. — Depósitos en otros Bancos (en Méjico, 218.381,89; en el Extranjero, 4.629.830,17), 4.848.212,06. — Préstamos y descuentos (con prendas, 3.310.473,99; compra de giros, 4.287.220,04; créditos en cuenta corriente,

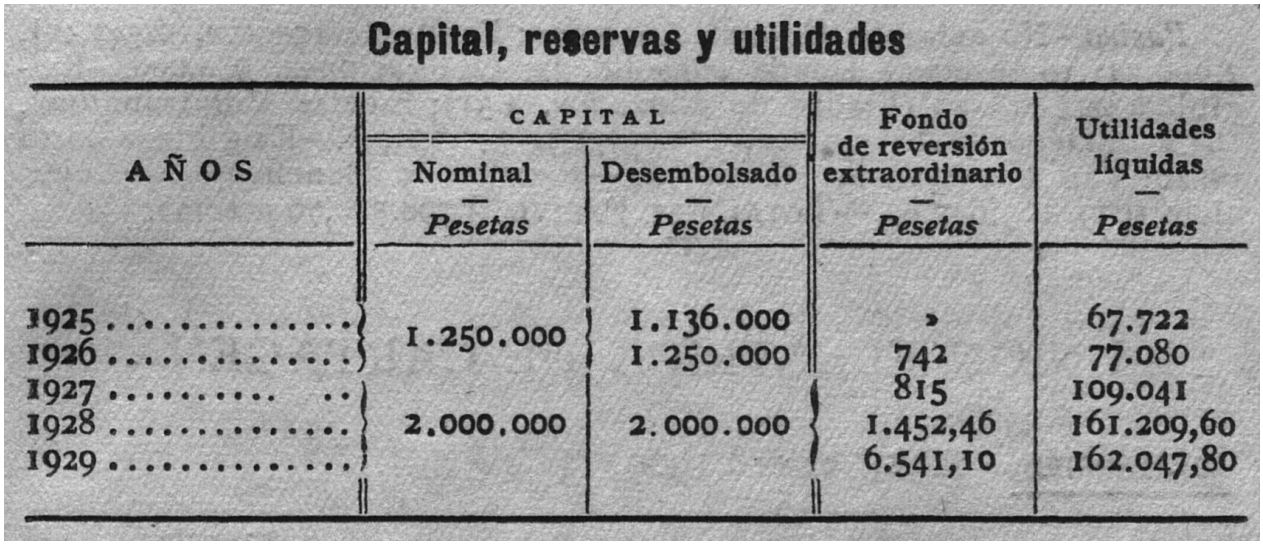
Figure 8: Beginning of a balance rubric

Even if their format is different, these have the same structure as other rubrics with a title and several text lines. They can be integrated to the global list of rubrics by simply defining their specific title.

The first major particularity of Spanish yearbooks is the format of the tables. Here we have explicitly defined columns thanks to borders, so we can build the structure of the table more accurately.

This project has received funding from

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



Capital, reservas y utilidades

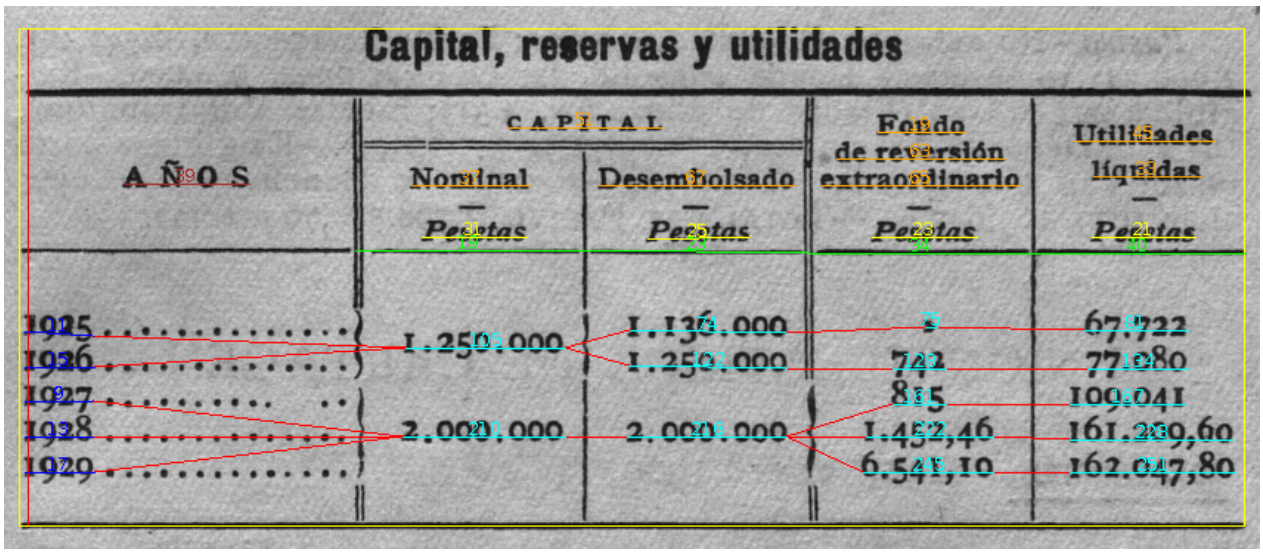
AÑOS	CAPITAL		Fondo de reversión extraordinario	Utilidades líquidas
	Nominal — Pesetas	Desembolsado — Pesetas	— Pesetas	— Pesetas
1925	1.250.000	1.136.000	»	67.722
1926		1.250.000	742	77.080
1927			815	109.041
1928	2.000.000	2.000.000	1.452,46	161.209,60
1929			6.541,10	162.047,80

Figure 9: Spanish table before analysis

The table follows the same structure as the French ones, with columns and different rows, so we can build it the same way. However, the methods to recognise the different parts of the table are mostly different because they are based on borders recognition.

In the example above (see Figure 9), we can see that columns can be subdivided in two or more sub-columns. The same structure can be found in some tables of the French yearbook, and implementing it for the Spanish also helped with the French ones.

The second major difference in these tables is the presence of braces, used to spread a value over several rows in a column. Depending on the structure of the row and alignment between the elements, we can tell if a row belongs to a brace, and spread the values accordingly (see Figure 10).



Capital, reservas y utilidades

AÑOS	CAPITAL		Fondo de reversión extraordinario	Utilidades líquidas
	Nominal — Pesetas	Desembolsado — Pesetas	— Pesetas	— Pesetas
1925	1.250.000	1.136.000	»	67.722
1926		1.250.000	742	77.080
1927			815	109.041
1928	2.000.000	2.000.000	1.452,46	161.209,60
1929			6.541,10	162.047,80

Figure 10: Spanish table after analysis

This project has received funding from

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

2.1.1.3 Genericity

To have an easier adaptation to new types of documents of the information extraction on yearbooks, we built it in a generic way, where the core recognition of the structure is the same from one yearbook to another and we just have to specify the characteristics of each document before starting the generic analysis of document structure.

Examples of these characteristics are: the font size of the titles, the position of tables regarding the rubrics, the type of the tables (with/without borders), the presence of a separator between issuers, etc.

Instead of developing a new system for each yearbook, we simply build the right combination of characteristics and elements found in a new corpus. By doing so, it becomes easier to add new documents by simply specifying its characteristics. We can use some of the Desfossés characteristics and some of the Madrid ones without changing the detectors of the elements composing a page.

Here are examples of how we define characteristics for different yearbooks:

```
pagecarac P ::=
     ``(pageType "Handbuch") &&

     (titleFontSize 26 ==>
     (titleAcrossPages ==>
     (centeredTitle ==>
     (issuerSeparator "Blank" ==>
     (tableType "NoTab" ==>
     (tablePosition "AfterRub" ==>
     (lineSpacing 75 ==>

     (rubTitleDef rubricTitleDE ==>

     page P)))))))))).
```

```
     pagecarac P ::=
          ``(pageType "Madrid") &&

          (titleFontSize 40 ==>
          (issuerSeparator "Blank" ==>
          (titleInCaps ==>
          (titleForbiddenWords
["Obligaciones",
     "Capital", "Dividendos "] ==>
          (tableType "WithRulings" ==>
          (tablePosition "InsideRub" ==>
          (lineSpacing 125 ==>

          (rubTitleDef rubricTitleES ==>

          page P)))))))))).
```

To add a whole new type of document, we can use these characteristics without modifying the detectors, and when needed, we can add new characteristics and improve the related description to recognize it, so it adds up to the library in case we encounter the same needs once again in the future.

2.1.2 Evaluation results

2.1.2.1 Evaluation of bounding boxes (ZoneMap)

To evaluate detection and classification on all three yearbooks, we still use the ZoneMap metric [Galibert 2014] for reasons we covered in D7.2. ZoneMap results are based on bounding boxes similarity, and tend towards zero when hypothesis and ground truth are close. ZoneMap provides a score between 0 (better) and arbitrary high values.

For the Desfossés yearbook, we have the same corpus of 61 images, with 61 titles, 792 rubrics and 123 tables (including 52 generic tables, 53 balance sheets and 18 extreme courses). See Table 1, we now have a global score of 2.69 (last year was 11.01): 0.20 for the titles and 0.97 for the

This project has received funding from

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

rubrics, which are almost perfectly recognised, 2.08 for the generic tables, 0.65 for the balance sheets and 7.27 for the extreme courses which still have a simplistic description because we don't need a complete analysis of this type of table.

Class	Items count	Old score	Current score
Images	61	11.01	2.69
Titles	61	14.96	1.25
Rubrics	792	3.06	0.20
Tables	52	35.13	2.08
Balance sheets	53	9.08	0.65
Extreme courses	18	/	7.27

Table 1: ZoneMap score on the French yearbook (lower is better)

For the Handbuch yearbook, we added more images to the corpus for a total of 126, with 154 titles and 1704 rubrics. See Table 3, we have a global error rate of 1.57 (3.4 last year, see Table 2): 15.19 for the titles and 1.81 for the rubrics.

Class	Items count	Old score	Current score
Images	35	3.4	2.05
Titles	48	7.55	7.45
Rubrics	524	3.19	1.83

Table 2: ZoneMap score on the German yearbook, old corpus (lower is better)

Class	Items count	Current score
Images	126	1.57
Titles	154	15.19
Rubrics	1704	1.81

Table 3: ZoneMap score on the German yearbook, new corpus (lower is better)

Note: Titles score in Handbuch increases a lot for little mistakes (such as miss-sizing the subtitle). Comparing the scores between the corpuses may be irrelevant because we added 91 images, this is why we have two different scores.

For the new Madrid yearbook, we annotated 72 pages, with 17 titles, 286 rubrics and 66 tables. See Table 4, the global score is 2.84: 0.44 for the titles, 2.10 for the rubrics, and 5.51 for the tables.

This project has received funding from

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

Class	Items count	Current score
Images	72	2.84
Titles	17	0.44
Rubrics	286	2.10
Tables	66	5.51

Table 4: ZoneMap score on the spanish yearbook (lower is better)

2.1.2.2 Evaluation of table analysis

Because ZoneMap score is only based on bounding boxes, it was not entirely relevant for evaluating tables. We could know if tables were well sized, but not how well the elements inside were recognised and linked together.

We needed a second metric, and the more appropriate one was published by Xu Zhong in March 2020. It is based on the Tree-Edit-Distance-based Similarity (TEDS) [Pawlik and Augsten 2016]. Tables are converted to HTML format, which uses a tree-like structure and enables us to compute the TEDS between hypothesis and ground truth. By using trees, we ensure that the score depends on the links between table elements (rows, cells, braces, …) and not only on their coordinates in the image.

This method was presented alongside the PubTabNet dataset for table evaluation (used in ICDAR 2021 Competition on Scientific Literature Parsing), and is easy to use thanks to the Python script provided by the authors.

The TEDS here is the percentage of similarity between two tables: higher is better, with 100 being the perfect score. The global score on a corpus is the average value of the scores of each table, taking into account the number of cells in the tables. If a table contains 100 cells and another only has 50, the former will weigh twice as much in the global score.

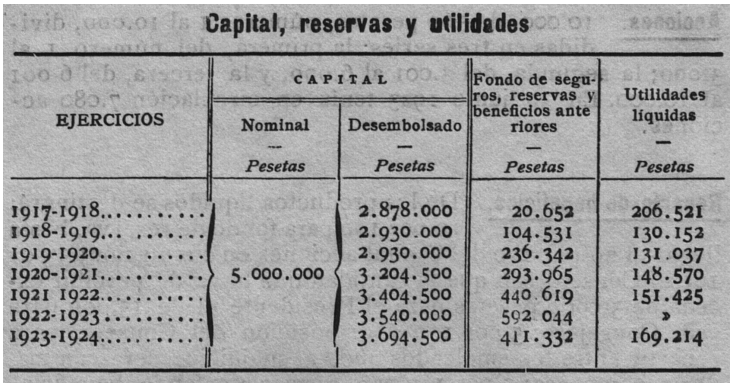
To build the evaluation corpus, we annotated 56 tables in the Desfossés yearbook and 100 tables in the Madrid yearbook. Once again, there is no data to be extracted from Handbuch tables.

Yearbook	Desfossés	Madrid
Current score	96.75%	93.31%

In Figure 11, an example of a Spanish table with its HTML equivalent is presented. The first one is the ground truth, and the bottom one is the hypothesis after analysis. The score for this table alone is 94.2%. This example shows the state of the analysis for a score close to the average score on the whole corpus.

This project has received funding from

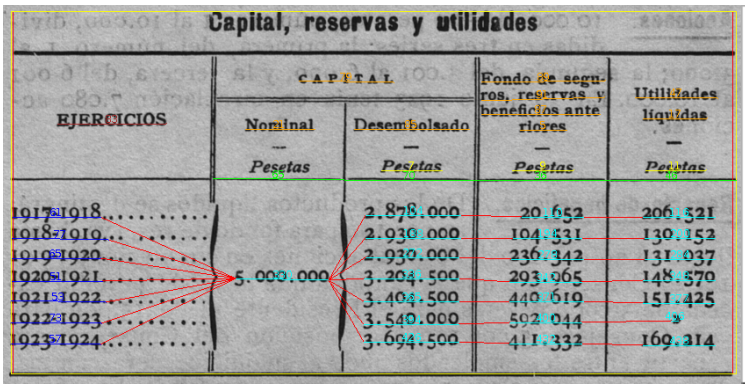
the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489



Capital, reservas y utilidades

EJERCICIOS	CAPITAL Nominal — Pesetas	CAPITAL Desembolsado — Pesetas	Fondo de seguros, reservas y beneficios anteriores — Pesetas	Utilidades líquidas — Pesetas
1917-1918		2.878.000	20.652	206.521
1918-1919		2.930.000	104.531	130.152
1919-1920		2.930.000	236.342	131.037
1920-1921	5.000.000	3.204.500	293.965	148.570
1921 1922		3.404.500	440.619	151.425
1922-1923		3.540.000	592 044	»
1923-1924		3.694.500	411.332	169.214

Title	Column	Column	Column	Column
	Currency			
Row	Cell	Cell	Cell	Cell
Row		Cell	Cell	Cell
Row		Cell	Cell	Cell
Row		Cell	Cell	Cell
Row		Cell	Cell	Cell
Row		Cell	Cell	»
Row		Cell	Cell	Cell



Capital, reservas y utilidades

EJERCICIOS	CAPITAL Nominal — Pesetas	CAPITAL Desembolsado — Pesetas	Fondo de seguros, reservas y beneficios anteriores — Pesetas	Utilidades líquidas — Pesetas
1917-1918		2.878.000	20.652	206.521
1918-1919		2.930.000	104.531	130.152
1919-1920		2.930.000	236.342	131.037
1920-1921	5.000.000	3.204.500	293.965	148.570
1921 1922		3.404.500	440.619	151.425
1922-1923		3.540.000	592 044	»
1923-1924		3.694.500	411.332	169.214

Title	Column	Column	Column	Column
		Currency		
Row	Cell	Cell	Cell	Cell
Row		Cell	Cell	Cell
Row		Cell	Cell	Cell
Row		Cell	Cell	Cell
Row		Cell	Cell	Cell
Row		Cell	Cell	
Row		Cell	Cell	Cell

Figure 11: Example of a Spanish table with its HTML equivalent. The first one is the ground truth, and the bottom one is the hypothesis after analysis.

2.2 Information Extraction System (Task 7.3)

2.2.1 System overview

In this section, we show the new results obtained with the extraction system that was presented in D7.2. Let us recall that financial yearbooks contain a lot of textual 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 because the alphanumeric information of interest is occurring inside specific linguistic patterns with many variations due to phrasing variations. One typical example of such a rubric is the capital rubric which is present in every yearbook, and from which we want to extract the capital evolution through the years. Typically, we want to extract a set of triplets each containing the following information (amount - date - currency).

The system that was developed and presented in D7.2 is based on machine learning algorithms that are trained to optimize their performance using manually annotated data. The more annotated data the better

This project has received funding from

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

performance we can get, but we showed that an average precision higher than 0.90% was obtained with less than 3 000 annotated rubrics. Of course, the performance will vary depending on the type of rubric.

2.2.2 Processing the Spanish 1931 Madrid yearbook

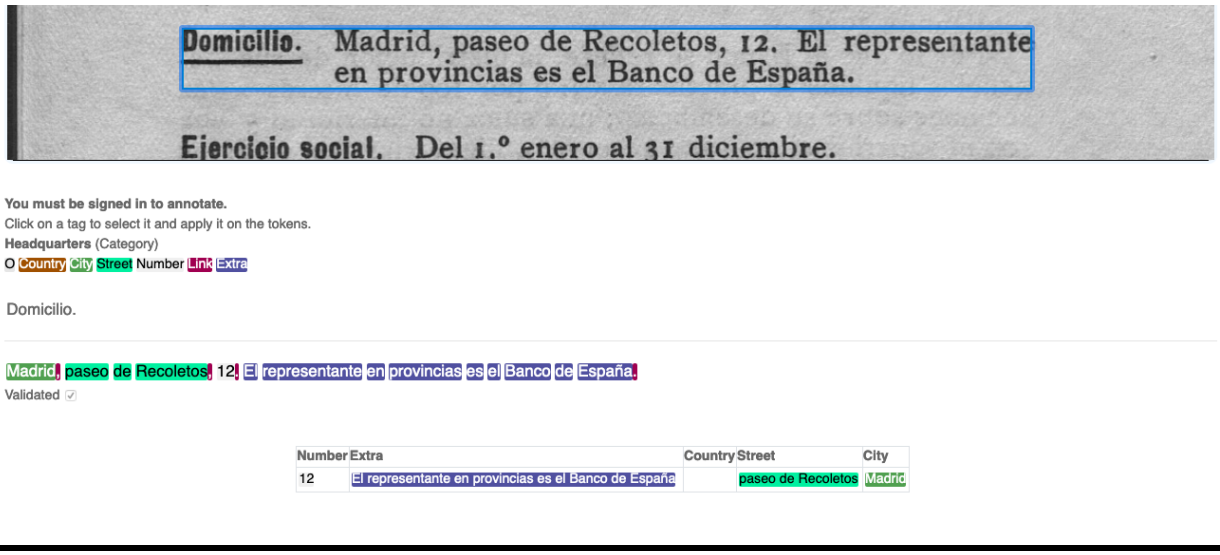
As was planned in our agenda, during the last period we experimented with the extraction of information on the Spanish 1931 Madrid yearbook, after having successfully demonstrated the approach on the French Desfossé 1962, and the German Handbuch 1914-15, see D7.2. In this respect we followed the protocol that was designed during our earlier experiments, which comprises the following steps:

1. Define the relevant rubrics to process, following the requirements of the historians
2. Define a TAG set to allow the extraction of the information needed
3. OCR each rubric
4. manually tag a certain amount of rubric with the collaborative web-based annotation tool that was presented in D7.2.
5. Train and test the extraction system
6. Process the whole yearbook with the system trained in 5
7. Show the results on the internet and export the results in a database.

In collaboration with the historian colleagues, we choose to conduct our experimentations on 5 rubrics (**Headquarters, Founding, Capital, Shares, Bonds**). A total of 24 TAGS have been defined to allow the information extraction of these 5 rubrics. Notice that most of these TAGS are not specific to the Spanish yearbook, only 7 of them were introduced specifically.

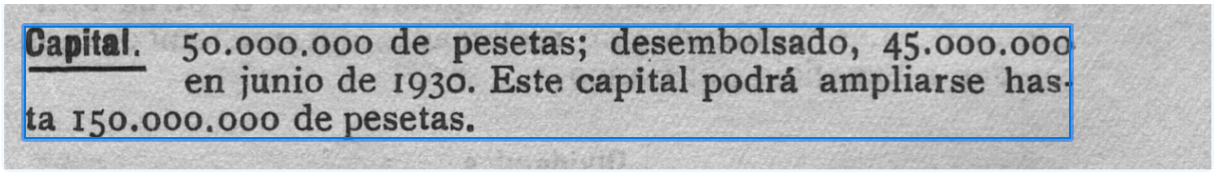
In the following we show some illustrative examples of the 5 rubrics that have been processed, and that are browsed thanks to our web-based annotation and viewing tool.

HEADQUATER rubric



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

CAPITAL rubric



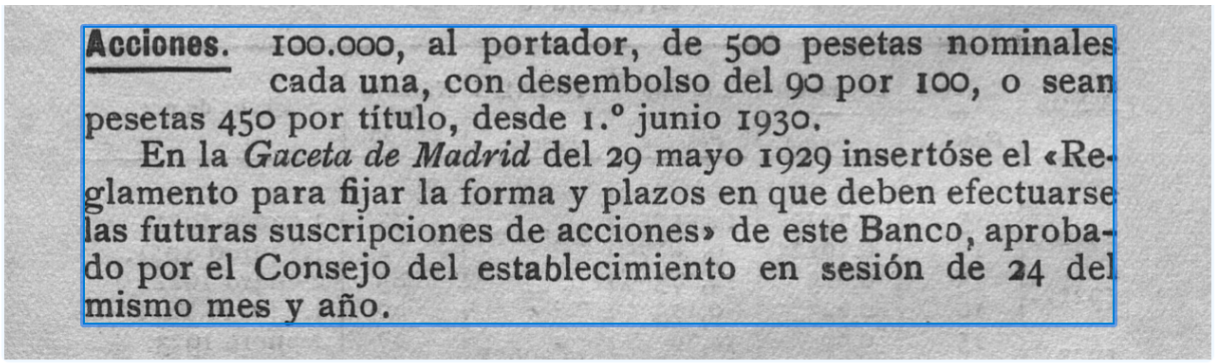
You must be signed in to annotate.
Click on a tag to select it and apply it on the tokens.
Capital (Category)
O Last-Amount Currency Initial-Date Initial-Amount Change-Amount Change-Date Link

Capital.

en junio de 1930. Este capital podrá ampliarse has- ta 150. 000. 000 de pesetas.
Validated

Type	Amount	Currency	Date
Last	150. 000. 000	pesetas	

SHARES rubric



You must be signed in to annotate.
Click on a tag to select it and apply it on the tokens.
Shares (Category)
O Number-of-Shares Type-Share Amount Type-Amount Date Numeration Link

Acciones.

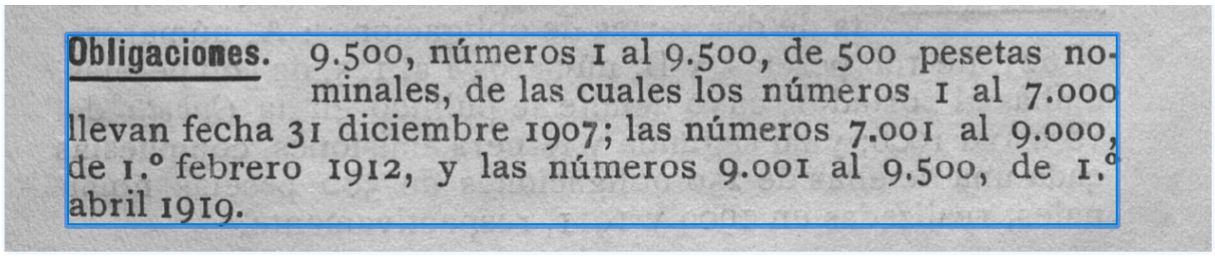
100. 000, al portador, de 500 pesetas nominales cada una, con desembolso del 90 por 100, o sean pesetas 450 por título, desde 1.º junio 1930. En la Gaceta de Madrid del 29 mayo 1929 insertóse el« Re- glamento para fijar la forma y plazos en que deben efectuarse las futuras suscripciones de acciones» de este Banco, aproba- do por el Consejo del establecimiento en sesión de 24 del mismo mes y año.
Validated

Numeration	Amount	Number-of-Shares	Type-Share	Type-Amount	Date
	500	100. 000	al portador	pesetas	1.º junio 1930

BONDS rubric

This project has received funding from

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



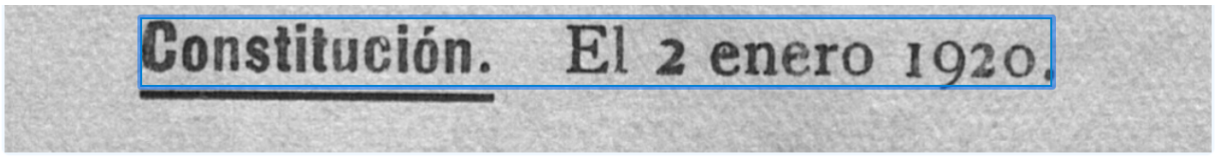
You must be signed in to annotate.
Click on a tag to select it and apply it on the tokens.
Bonds (Category)
O Link Interests Date Currency Type-Amount Amount Number-of-Bond Type-Bond Bond-Link

Obligaciones.

9. 500, números 1 al 9. 500, de 500 pesetas no- minales, de las cuales los números 1 al 7. 000 llevan fecha 31 diciembre 1907; las números 7. 001 al 9. 000, 1. @ de 1. f febrero 1912, y las números 9. 001 al 9. 500, de abril 1919.
Validated

Amount	Currency	Type-Bond	Type-Amount	Number	Date	Interests
	pesetas	1 al 9. 500	500	9. 500		
		1 al 7. 000			31 diciembre 1907	
		7. 001 al 9. 000			f febrero 1912	
		9. 001 al 9. 500			abril 1919	

FOUNDING rubric



You must be signed in to annotate.
Click on a tag to select it and apply it on the tokens.
Founding (Category)
O Date

Constitución.

El 2 enero 1920.
Validated

Date
2 enero 1920

This project has received funding from

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

2.2.3 Evaluation on all corpuses

2.2.3.1 Annotated data for the French Yearbook

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 5: French Yearbook labeled data per rubric

2.2.3.2 French Yearbook Desfossés 1962, per rubric Tag evaluation

Administrators	Precision	Recall	$F_{\beta=1}$
admin_comment	50.00%	42.86%	46.15
admin_comment_link	0.00%	0.00%	0.00
admin_family_name	98.04%	98.44%	98.24
admin_name	97.87%	98.76%	98.31
admin_next_link	97.77%	98.24%	98.00
admin_pos_link	99.01%	99.01%	99.01
admin_position	99.01%	99.01%	99.01
admin_title	0.00%	0.00%	0.00
Overall	97.30%	97.55%	97.42

Table 6: Tag evaluation results French Administrators rubric

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

Headquarters	Precision	Recall	$F_{\beta=1}$
arrond	98.08%	98.08%	98.08
comment	0.00%	0.00%	0.00
company	14.29%	25.00%	18.18
country	66.67%	20.00%	30.77
district	55.26%	84.00%	66.67
link	86.33%	90.24%	88.25
nextlink	54.55%	61.22%	57.69
no-str	92.94%	94.05%	93.49
str	87.63%	87.63%	87.63
town	80.09%	87.44%	83.60
type	61.90%	61.90%	61.90
Overall	80.49%	85.34%	82.85

Table 7: Tag evaluation results French Headquarters rubric

Founding	Precision	Recall	$F_{\beta=1}$
chg-duree	0.00%	0.00%	0.00
chg-enddate	40.00%	50.00%	44.44
chg-startdate	70.00%	46.67%	56.00
chg-status	46.15%	50.00%	48.00
ini-duree	97.16%	100.00%	98.56
ini-enddate	88.24%	93.75%	90.91
ini-startdate	95.32%	96.45%	95.88
ini-status	95.78%	93.53%	94.64
link	96.45%	96.17%	96.31
Overall	94.66%	94.22%	94.44

Table 8: Tag evaluation results French Founding rubric

Capital	Precision	Recall	$F_{\beta=1}$
Ini-amount	93.02%	93.02%	93.02
Ini-date	97.56%	93.02%	95.24
chg-amount	97.55%	95.50%	96.51
chg-date	93.93%	95.64%	94.77
currency	96.62%	92.45%	94.49
last-amount	94.12%	92.31%	93.20
last-date	0.00%	0.00%	0.00
link	50.35%	42.60%	46.15
Overall	90.26%	86.93%	88.57

Table 9: Tag evaluation results French Capital rubric

This project has received funding from

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

Operations	Precision	Recall	$F_{\beta=1}$
arrond	0.00%	0.00%	0.00
comment	16.67%	20.00%	18.18
district	100.00%	100.00%	100.00
link	87.50%	63.64%	73.68
nextlink	79.31%	92.00%	85.19
no-str	0.00%	0.00%	0.00
str	0.00%	0.00%	0.00
town	86.27%	91.67%	88.89
type	94.44%	100.00%	97.14
Overall	82.71%	76.92%	79.71

Table 10: Tag evaluation results French Operations rubric

Sales	Precision	Recall	$F_{\beta=1}$
amount	97.93%	97.93%	97.93
currency	94.92%	95.73%	95.32
link	97.05%	98.29%	97.66
year	97.93%	99.58%	98.75
Overall	97.26%	98.19%	97.72

Table 11: Tag evaluation results French Sales rubric

Financial Year	Precision	Recall	$F_{\beta=1}$
end-date	100.00%	87.50%	93.33
start-date	85.71%	100.00%	92.31
Overall	92.86%	92.86%	92.86

Table 12: Tag evaluation results French Financial Year rubric

Coupons	Precision	Recall	$F_{\beta=1}$
Coup_amount	97.85%	96.81%	97.33
Coup_comment	86.40%	88.52%	87.45
Coup_currency	93.80%	94.90%	94.35
Coup_date	96.93%	98.37%	97.64
Coup_modality	36.36%	21.05%	26.67
Coup_nb	96.96%	97.48%	97.22
Coup_security	83.78%	79.49%	81.58
Overall	95.89%	96.17%	96.03

Table13: Tag evaluation results French Coupons rubric

This project has received funding from

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

2.2.3.3 Annotated data for the German Yearbook

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 14: German Yearbook labeled data per rubric

2.2.3.4 German Yearbook Handuch 1914-1915, per rubric Tag evaluation

Capital	Precision	Recall	$F_{\beta=1}$
Cap-decr	0.00%	0.00%	0.00
Cap-incr	84.91%	90.00%	87.38
Ini-amount	94.12%	69.57%	80.00
Ini-date	100.00%	85.71%	92.31
capital_part	100.00%	100.00%	100.00
chg-amount	90.00%	69.23%	78.26
chg-date	83.08%	75.00%	78.83
currency	95.52%	87.27%	91.21
last-amount	95.00%	100.00%	97.44
link	89.40%	73.77%	80.84
share_number	100.00%	73.85%	84.96
share_price	96.83%	88.41%	92.42
share_type	92.16%	68.12%	78.33
Overall	92.17%	79.79%	85.54

Table 15: Tag evaluation results German Capital rubric

This project has received funding from

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

Balance Sheet	Precision	Recall	$F_{\beta=1}$
amount	96.35%	96.06%	96.21
amount_total	94.12%	94.12%	94.12
currency	100.00%	100.00%	100.00
day	100.00%	100.00%	100.00
item	91.74%	91.20%	91.47
item_total	94.12%	94.12%	94.12
link_date	100.00%	100.00%	100.00
link_item	66.67%	14.29%	23.53
link_part	96.36%	97.85%	97.10
month	100.00%	95.00%	97.44
part_assets	94.74%	94.74%	94.74
part_liabilities	100.00%	100.00%	100.00
year	95.00%	95.00%	95.00
Overall	95.12%	94.31%	94.71

Table 16: Tag evaluation results German Balance Sheet rubric

Founding	Precision	Recall	$F_{\beta=1}$
concession	100.00%	70.00%	82.35
day	97.62%	92.48%	94.98
effect	100.00%	94.12%	96.97
founder	97.06%	97.06%	97.06
general_assembly	100.00%	100.00%	100.00
ini-day	91.33%	95.14%	93.20
ini-month	75.95%	83.92%	79.73
ini-year	91.72%	88.89%	90.28
link	88.46%	84.15%	86.25
link_between	16.67%	50.00%	25.00
link_date	96.67%	94.91%	95.78
link_yearbook	88.57%	91.18%	89.86
month	78.95%	68.18%	73.17
registration	98.78%	95.29%	97.01
year	89.39%	88.72%	89.06
yearbook	88.57%	91.18%	89.86
Overall	91.25%	89.99%	90.62

Table 17: Tag evaluation results German Founding rubric

This project has received funding from

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

Financial Year	Precision	Recall	$F_{\beta=1}$
calendar	100.00%	96.67%	98.31
day_end	97.14%	91.89%	94.44
day_start	97.14%	89.47%	93.15
link_between	91.67%	89.19%	90.41
link_within	95.71%	90.54%	93.06
month_end	97.14%	94.44%	95.77
month_start	94.44%	89.47%	91.89
Overall	96.01%	91.38%	93.64

Table 18: Tag evaluation results German Financial Year rubric

Voting Right	Precision	Recall	$F_{\beta=1}$
currency	95.38%	93.94%	94.66
currency_amount	95.38%	93.94%	94.66
link	92.31%	87.50%	89.84
link_share_amount	0.00%	0.00%	0.00
max	0.00%	0.00%	0.00
share	89.43%	89.43%	89.43
share_amount	86.03%	90.70%	88.30
share_type	0.00%	0.00%	0.00
vote	98.33%	92.19%	95.16
vote_amount	99.17%	90.91%	94.86
Overall	91.98%	89.80%	90.88

Table 19: Tag evaluation results German Voting Right rubric

2.2.3.5 Annotated data for the Spanish Yearbook

Spanish 1931 Madrid Yearbook	
TREATED RUBRIC	**LABELED RUBRICS FOR THE SYSTEM**
Capital	196
Founding	202
Headquarters	206
Shares	194
Bonds	137

Table 20: Spanish Yearbook labeled data per rubric

This project has received funding from

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

2.2.3.6 Spanish 1931 Madrid Yearbook, per rubric Tag evaluation

Bonds	Precision	Recall	$F_{B=1}$
Date	85.32%	73.51%	78.97
Link	89.29%	59.62%	71.50
Type-Bond	92.73%	15.41%	26.42
Amount	64.66%	82.69%	72.57
Currency	93.18%	22.78%	36.61
Number-of-Bond	40.54%	53.01%	48.67
Type-Amount	53.36%	44.64%	53.19
Interests	98.87%	73.60%	84.38
Overall	80.52%	72.06%	72.08

Table 21: Tag evaluation results Spanish Bonds rubric

Capital	Precision	Recall	$F_{B=1}$
Last-Amount	97.52%	98.59%	98.05
Link	86.74%	74.97%	80.43
Currency	100.00%	49.40%	66.02
Initial-Amount	93.96%	93.96%	93.96
Change-Date	95.39%	94.24%	94.81
Change-Amount	92.76%	90.27%	91.50
Initial-Date	100.00%	62.15%	58.61
Overall	92.32%	88.19%	88.05

Table 22: Tag evaluation results Spanish Capital rubric

Founding	Precision	Recall	$F_{B=1}$
Date	99.78%	99.59%	99.68
Overall	99.78%	99.59%	99.68

Table 23: Tag evaluation results Spanish Founding rubric

Headquarters	Precision	Recall	$F_{B=1}$
City	96.86%	89.17%	92.86
Street	95.05%	97.62%	96.31
Number	94.65%	97.41%	96.01
Link	89.36%	93.87%	91.56
Extra	97.20%	97.81%	97.50
Country	100.00%	56.44%	40.89
Overall	94.52%	94.53%	94.44

Table 24: Tag evaluation results Spanish Headquarters rubric

This project has received funding from

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

Shares	Precision	Recall	$F_{B=1}$
Number-of-Shares	98.24%	84.67%	90.95
Link	88.39%	71.62%	79.12
Amount	97.29%	88.11%	92.47
Type-Amount	96.97%	91.00%	93.89
Type-Share	88.90%	75.34%	81.56
Date	77.90%	84.43%	81.03
Numeration	92.57%	90.00%	91.26
Overall	88.09%	87.07%	87.15

Table 25: Tag evaluation results Spanish Shares rubric

2.2.3.7 Global results in yearbooks

In the following tables, we show the performance of the extraction process in the Spanish 1931 Madrid Yearbook. The two tables show the average performance (first table) and the detailed per rubric performance. For sake of comparison with the results presented in D7.2 last year, both tables recall the results obtained on the French Desfossé, and the German Handbuch. We can see that a similar level of performance was obtained on the three Yearbooks.

Yearbook	# Rubrics	# Tags	# Manually Tagged Rubrics	# Machine Tagged Rubrics	Precision
French	8	43	1415	19339	90.80%
German	6	58	2494	18711	91.75%
Spanish	5	24	935	2007	91.03%

Table 26: Global analysis of the Yearbooks and Precision

This project has received funding from

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

Rubric	French Precision (%)	German Precision (%)	Spanish Precision (%)
Administrators	97.3		
Headquarters	80.4		94.5
Founding	94.7	91.2	99.7
Capital	90.2	92.2	92.3
Operation	82.7		
Sales	97.3		
Financial year	92.9	96.0	
Coupon	95.9		
Balance Sheet		95.1	
Voting Right		92.0	
Loss and Profit Account		84.0	
Shares			88.0
Bonds			80.5

Table 27: Global precision per tag

This project has received funding from

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

2.3 Application of the Yearbook Information Extraction System to three yearbooks

We have applied the data extraction system on the three yearbooks (Handbuch 1913-1914, Madrid 1931 and 1935, Desfossés 1962) ***for a total of 7,453 pages and it generated 6,009 issuers, 73,853 rubrics, 2,086 tables and 1,665 balance sheets***. Using these detected rubrics, ***we analyzed a total of 19 types of rubrics in three different languages and distinguished 126 different types of Named Entities***. ***The system was able to tag and extract automatically more than 40 000 named entities exploiting 4844 hand labelled data with a precision higher than 90%.***

The quality of the data extraction follows the different performance and recognition rates we have presented for each yearbook in this document. All the data extracted have been inserted in the WP7 Data Extraction Viewer. This viewer allows users to browse all images and the data extracted. An example on the French Desfossés 1962 is presented for the "LA FONCIERE" company with the rubric Capital selected. Once this rubric is open, the user can access the different modification of the capital automatically extracted in a structured way. On the same page the Balance Sheet is selected, we can see the hierarchical structure of the table, where the "C- Dette flottante" is selected and belongs to the liabilities section, and the extracted data is presented in Figure with for each value the associated information given by the table (liabilities, year, item, amount, currency). This demonstrator is available at this address for the three yearbooks: http://litis-eurhisfirm.univ-rouen.fr/pivan/

This project has received funding from

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

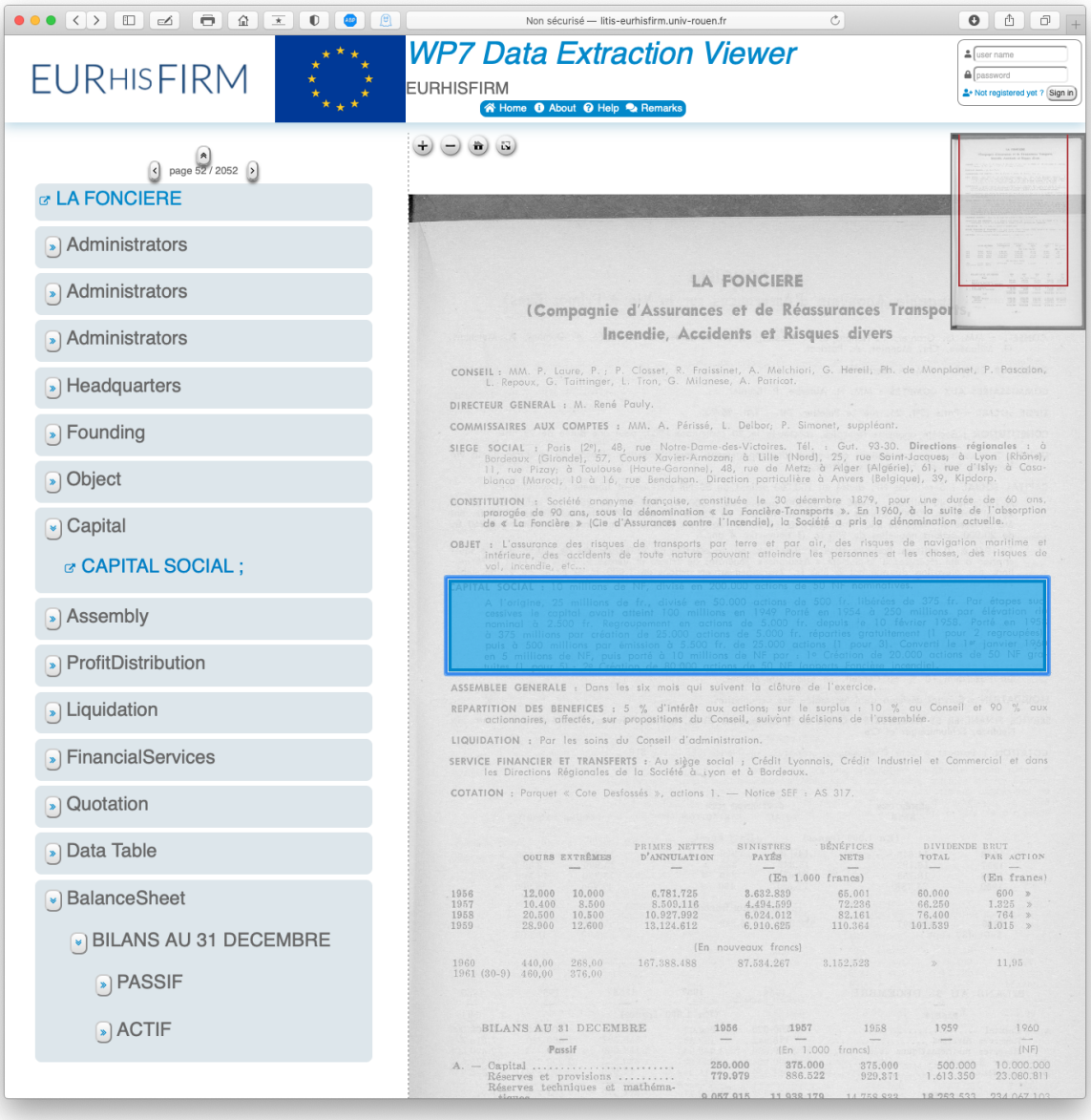


Figure 12: WP7 Data Extraction Viewer - "LA FONCIERE" company with the rubric Capital selected

This project has received funding from

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

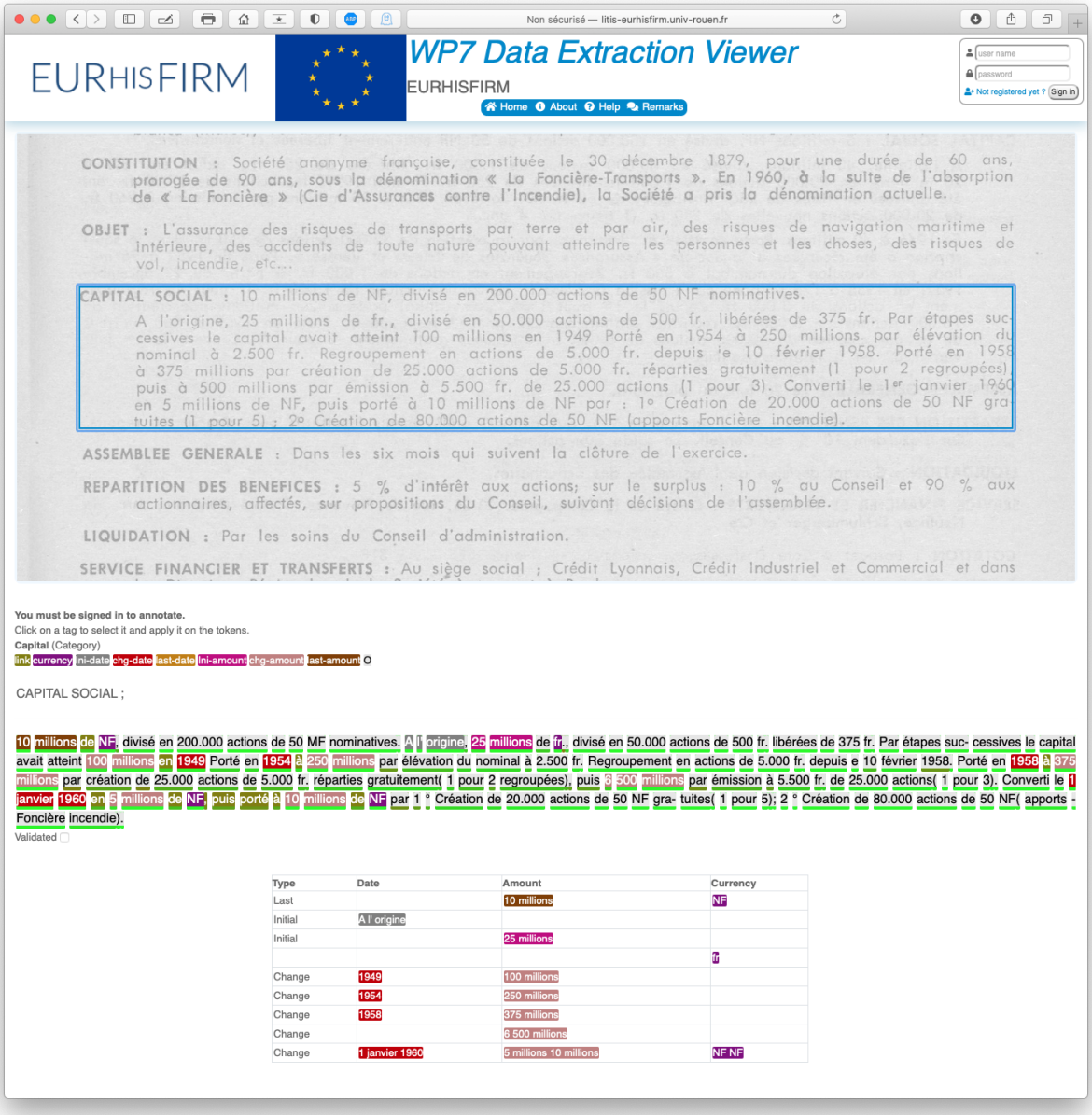


Figure 13: WP7 Data Extraction Viewer - Capital rubric with all the capital changes automatically extracted

This project has received funding from

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

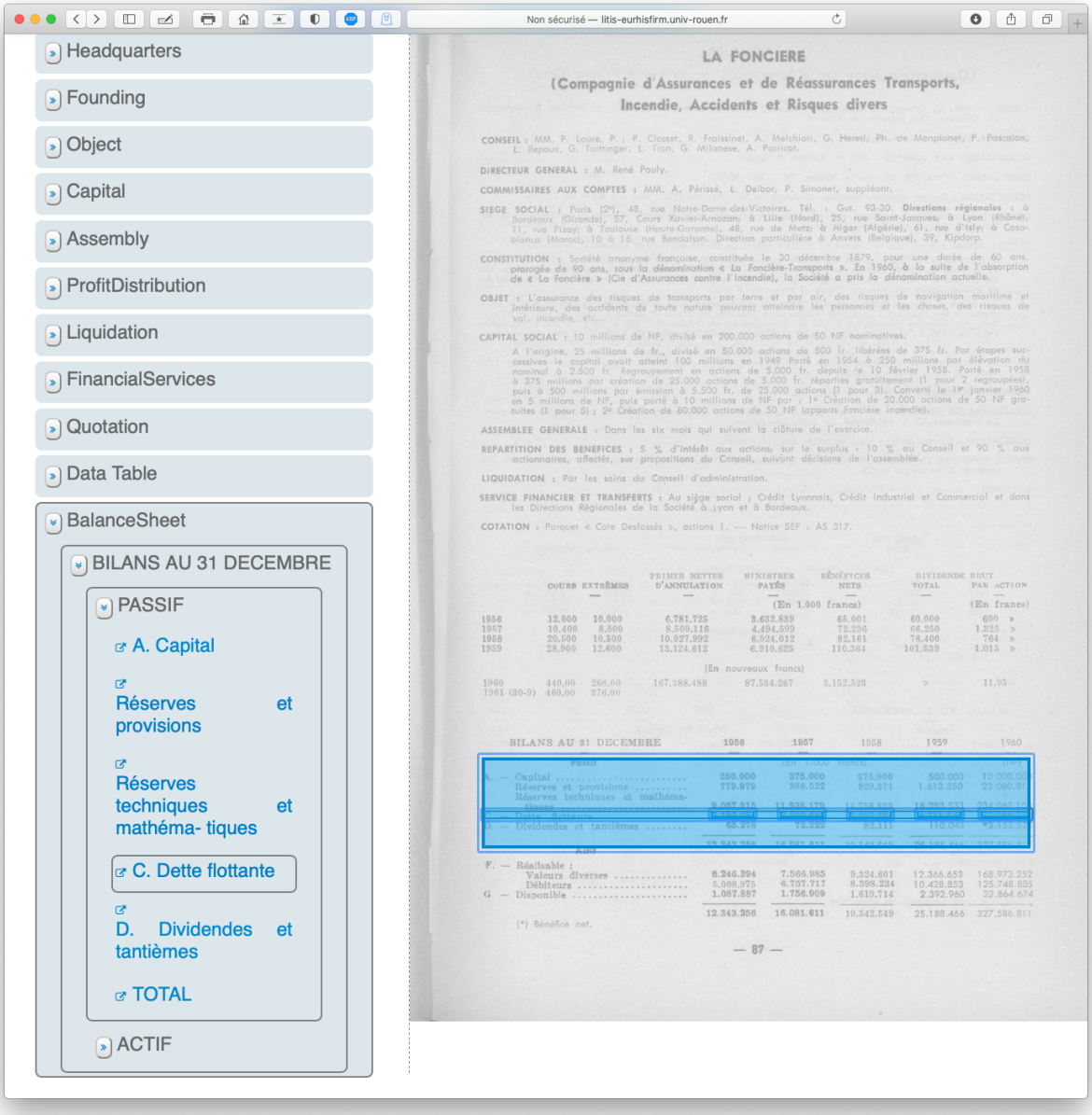


Figure 14: WP7 Data Extraction Viewer - Balance Sheet rubric selected with the detected hierarchy of the table. The "C. Dettes flottante" liabilities line is selected

This project has received funding from

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

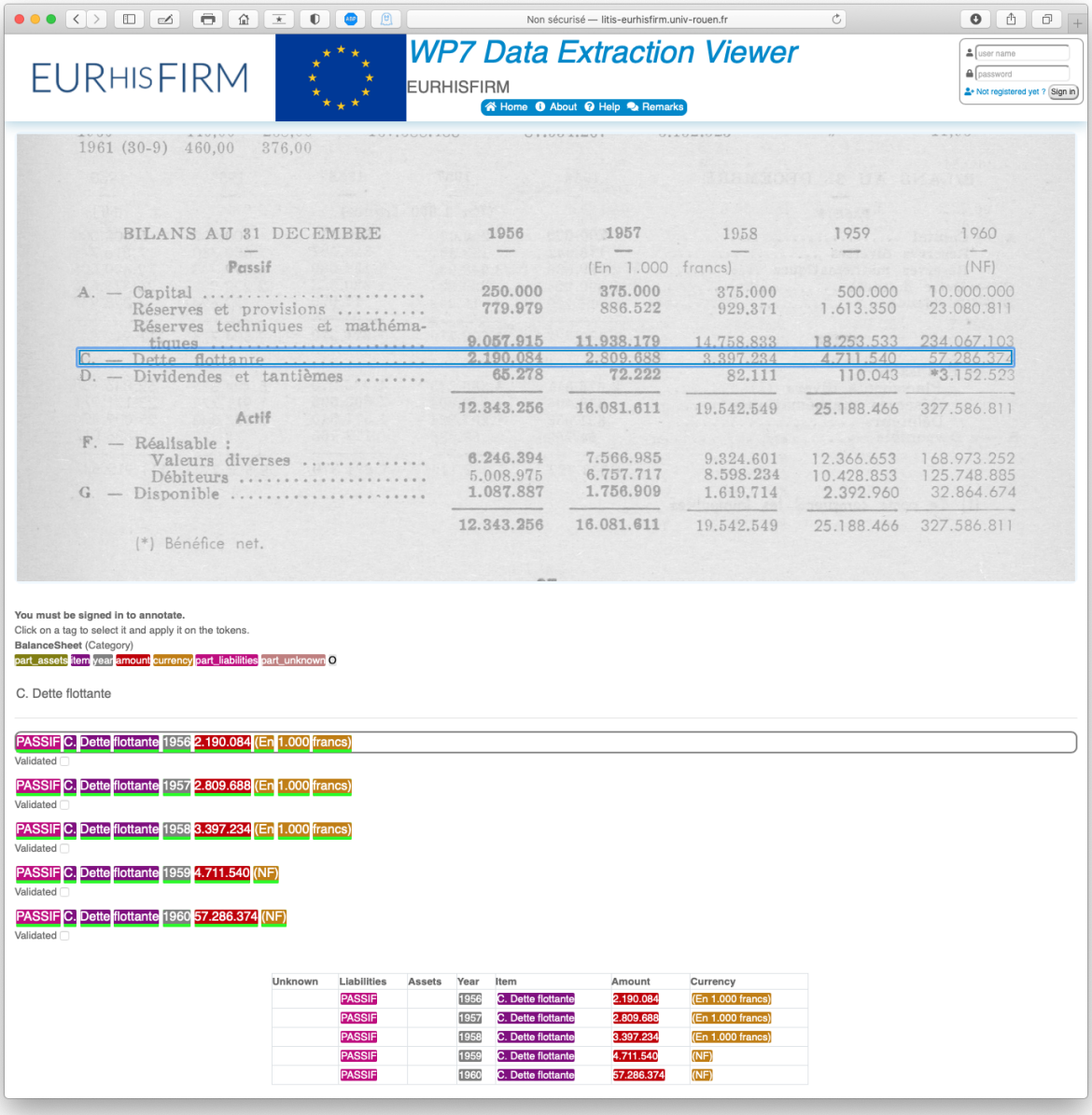


Figure 15: WP7 Data Extraction Viewer - Liabilities line "C. Dette flottante" of the Balance Sheet rubric with all the data automatically extracted

This project has received funding from

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

3 Price List Extraction system

This section will be about the updates we made on the price list extraction system since deliverable D7.2. The system consists of two main tasks: the document structure recognition using a cross validation module, and the definition of a general-purpose text recognizer (OCR). Most of our work focused on generalizing our processes in order to be able to analyse and extract data from price lists of different origins.

3.1 Strategy and transversal analysis

As presented in D7.2, we design a global strategy to take advantage of the sequentiality of the collection and correct errors in noisy documents. Our global strategy is based on an iterative process (see Figure 16). The aim of each iteration is to recognize and validate a structural or textual element of the documents: columns, sections, stock names (table entry), and other fields.

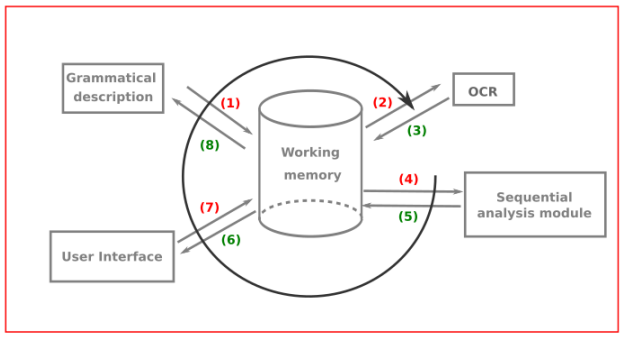


Figure 16: Overview of the global strategy

Each iteration consists of 5 steps:

1. a first structural analysis with a grammatical description to produce hypothesis,
2. the transcription of the text-lines localized in step 1,
3. a sequential analysis for the validation of the extracted elements from the image,
4. an eventual call to a user interface;
5. a new structural analysis that integrates the knowledge obtained in steps 2, 3 and 4.

This project has received funding from

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

With the user interface, we validate certain elements that could not be validated automatically because the information is not on the images. For example, when a new stock appears on the market, we need to know if it is really a new stock or an existing stock that has changed its name. Figure 17 shows another example: the text "m." has been correctly recognized, but we cannot automatically determine whether this currency refers to the German, Polish or Finnish mark. This interface presents the different questions to an expert with all the information needed to answer: the current stock quotation (in the middle), the previous day for the same stock (above) and the next day (below). The objective is to reduce at a minimum the number of questions asked to experts, thanks to the collection redundancies.

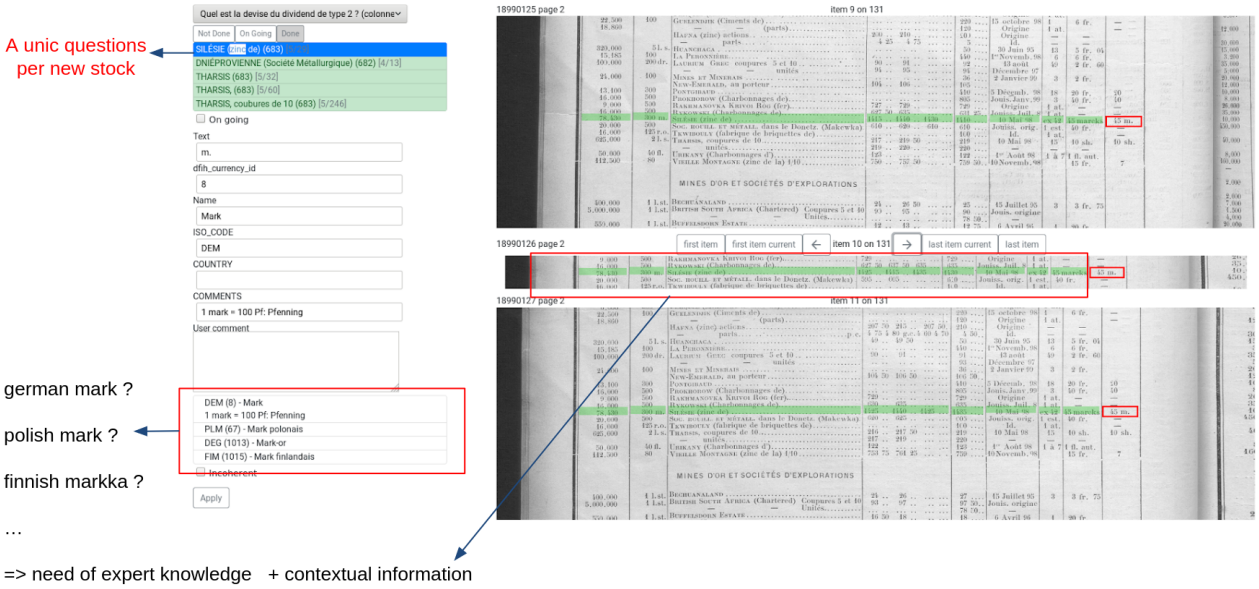


Figure 17: User interface - case of ambiguous interpretation

3.2 Global Meta Table Structure: tables of tables

3.2.1 System

The structure of price lists can vary depending on their origin and their date. The system developed in the context of the ANR project HBDEX, detailed in deliverable D7.2 is able to extract data in price lists when tables are vertically aligned. The tables that we need to detect are not ordered or presented in the same way from one price list to another. For example, a global meta table can group several price lists in each of its cells, with in some cases, a price list which starts in one cell and continues in the next cell (Figure 15). We therefore developed the recursive table structure analysis based on previous work done in [Coüasnon 2006]. It describes a global meta table structure whatever the number of rows, whatever the number of columns it is made of, and tries to detect recursively in each detected cell, another recursive table structure. This description is done using a bi-dimensional grammatical description of table structures and is able to detect the global meta table structure in any price lists tables.

This project has received funding from

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

The system can use double or thick vertical or horizontal line borders, and/or understand the recursive organization of the table to detect the global table. In case documents are degraded or if a line segment representing a line border is damaged, the system is also able to correctly detect the global table structure. There are actually four possible types of structures that can be recursively detected:

- A bi-dimensional table made at minimum of four cells organized in two rows and two columns
- A horizontal mono-dimensional table made at minimum of two cells
- A vertical mono-dimensional table made of two cells
- A cell

This table structure system is built on a multiresolution line segment detection [Lemaitre 2009] to extract thin lines, thick lines, multiple lines. The analysis and recognition of the global table organization is also able to understand the reading order of each cell containing a price list table. It uses the table organization and the coherence of the headers found in each price list table. Then, the price list data extraction system is applied separately on each cell.

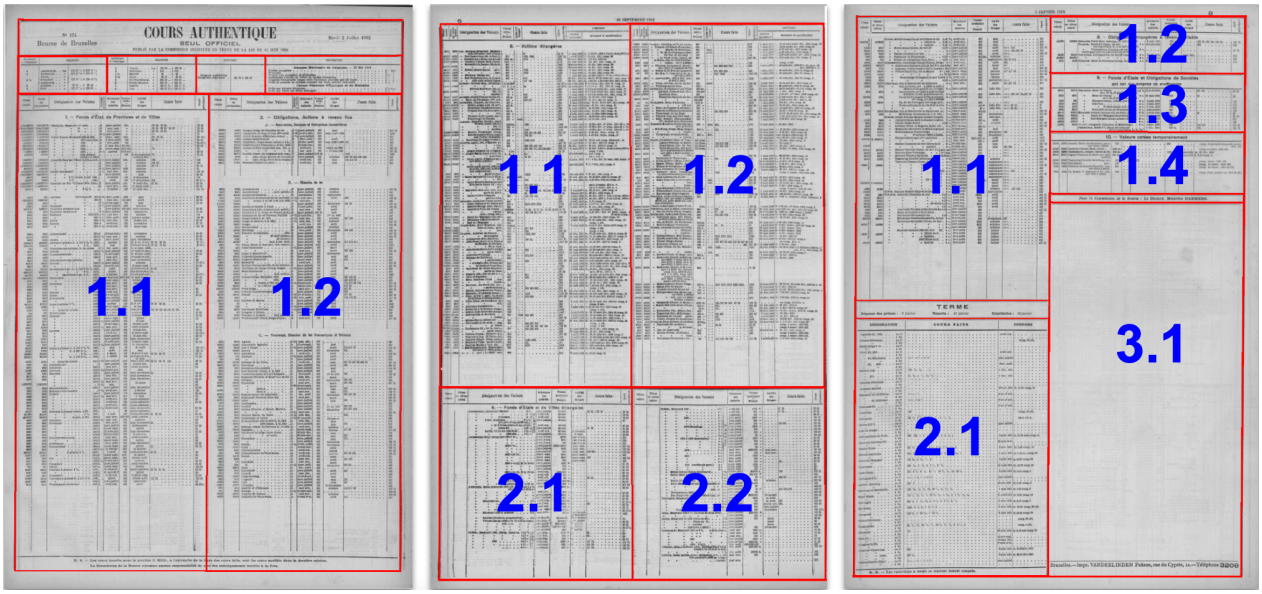


Figure 18: Brussels 1912 Price Lists: Global meta table recognition with reading order recognition

This project has received funding from

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

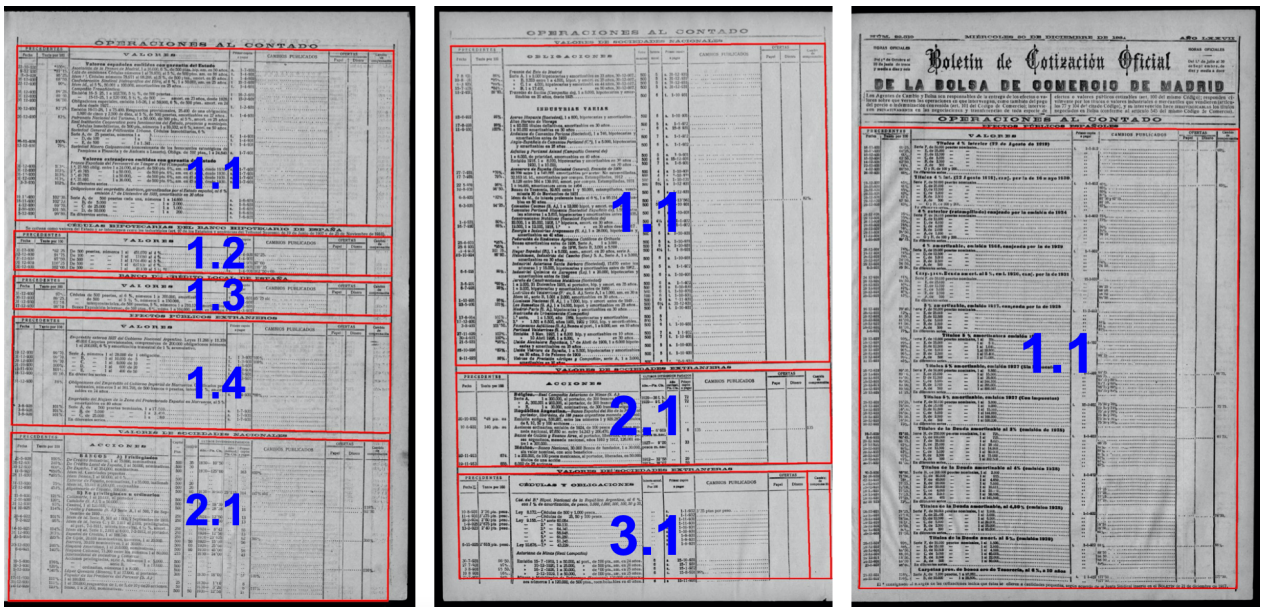


Figure 19: Madrid 1931 Price Lists: Global meta table recognition with reading order recognition

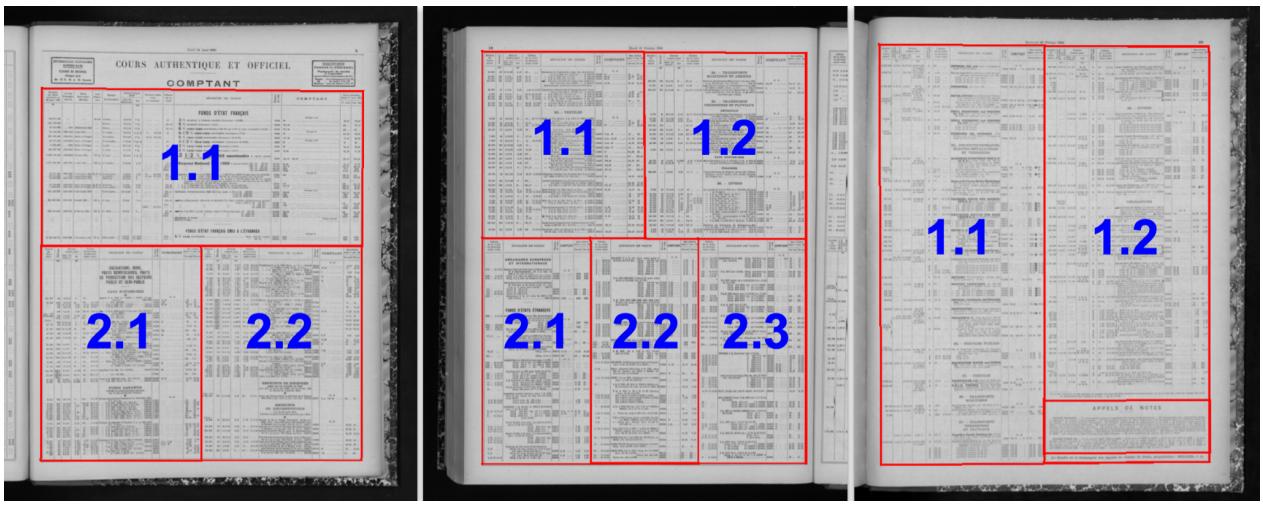


Figure 20: Paris 1961 Price Lists: Global meta table recognition with reading order recognition

3.2.2 Evaluation results

In order to evaluate our system, we used the ZoneMap metric [Galibert 2014] which is a metric we already used for yearbooks. ZoneMap results are based on bounding box similarity, and tend towards zero when hypothesis and ground truth are close.

This project has received funding from

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

We built three different corpus, one for each type of price list we have to work on : Brussel, Parquet and Madrid. Each corpus contained 30 pages and we obtained a global score of 0.52 for Brussel, 0.72 for Parquet and 1.37 for Madrid (see Table 28).

Origin	Images count	Tables count	Current score (ZoneMap)
Brussel	30	120	0.52
Parquet	30	71	0.72
Madrid	30	52	1.37

Table 28: ZoneMap score on the Brussel, Parquet and Madrid price lists (lower is better)

Madrid's score is slightly higher because pages are more degraded in this price list than the other two, and the structure is a little harder to detect, but the score is still very low, showing that the recognition is very good. The score shown in the table is the mean of the score per page. To get an idea of the meaning of the score we can see an example in Figure 21.

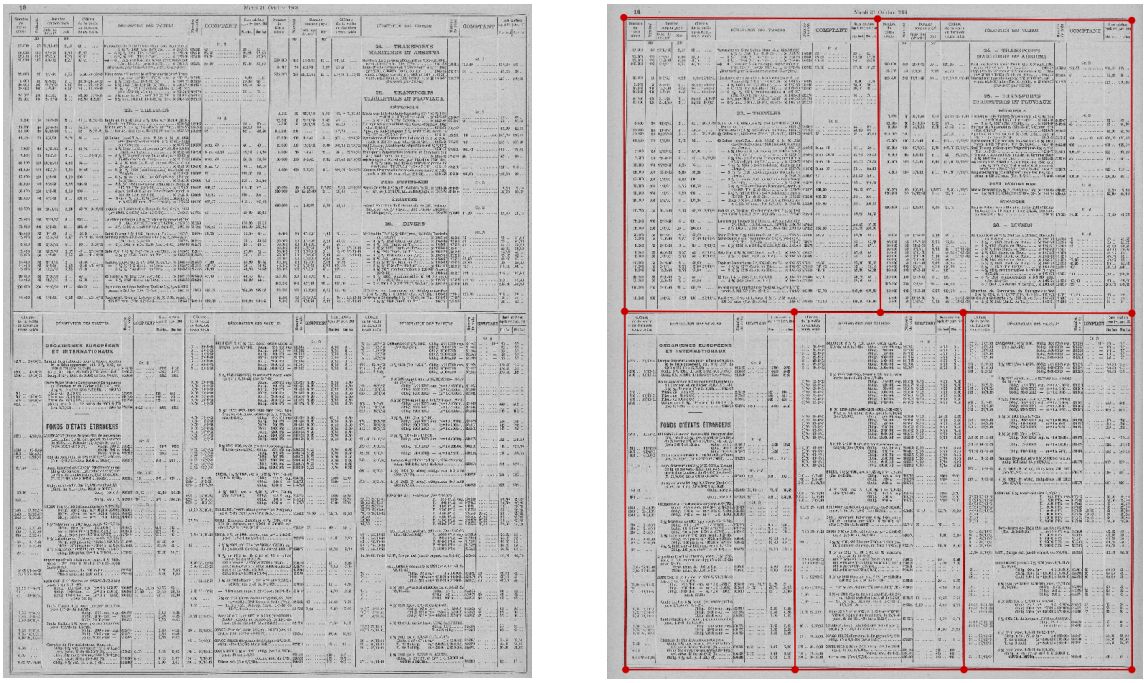


Figure 21: Meta-structure extracted on a page from Parquet price lists

Here we obtain a score of 0.88, because we correctly extracted each of the five tables composing the structure. We do not obtain a score of 0 because corners coordinates vary a bit from the ground truth. Otherwise, ZoneMap is mostly influenced by mistakes made on the detected area. For instance, a really high score (more than 1000) would appear if an entire table isn't detected. A score containing no issues is generally near 0 or 1.

This project has received funding from

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

3.3 Generic description of table content

3.3.1 System

In this section, we will see the updates we made since D7.2 regarding price lists data extraction. We focused on making our system generic, and then we upgraded the structures taken into account by our data extractor.

3.3.1.1 Genericity

We updated our data extraction system to make it generic and easily adaptable to new structures and price lists. In order to do this, we used the same approach as in yearbooks. Our system always analyses a document the same way, but we specify characteristics that slightly modifies the process when needed.

For instance, here is how we declare Brussels and Paris Parquet price lists specificities:

```
priceList P  ::=

``(typePage      "French_Parquet")      &&

(DECLARE(idemFirstLine)                  (
(DECLARE(doubleSeparator)                (
(DECLARE(titleMultipleLinesStaggered)    (
(DECLARE(columnSectionTitle)             (
(DECLARE(grFilter)                       (
(DECLARE(useNearestLine)(
(DECLARE(commentTitleMultipleLinesStagger
ed)(
 page P)))))))))))))).
```

```
priceList P  ::=

``(typePage Belgium_Bruxelles") &&

(DECLARE(doubleSeparator) (
(DECLARE(thickSeparator) (
(DECLARE(titleMultipleLines) (
(DECLARE(titleMultipleLinesStaggered) (
(DECLARE(smallSectionTitle) (
(DECLARE(outOfColumnSectionTitle) (
(DECLARE(quotationMark) (
(DECLARE(titles)(
(DECLARE(noBanner)(
(DECLARE(useNearestLine)(
(DECLARE(commentTitleMultipleLinesStaggere
d)(
page P)))))))))))))))))))).
```

That way, each type of document is characterized by a list of attributes and not only by their origin. These attributes could be used again in future price lists documents, to produce in an easy way an adapted version combining in a different way the different characteristics (stock on multiple lines for example).

We will now detail the major structure adaptations we had to work on for Brussel, Paris Parquet and Madrid price lists. All these adaptations describe new characteristics which are added to the library, the same way it is done on yearbooks.

This project has received funding from

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

3.3.1.2 Evolution in structural analysis

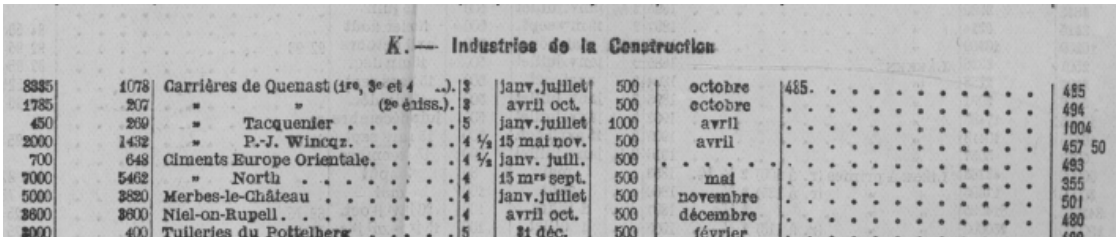
Section titles

The first characteristic we had to work on was the disposition of the section titles. We found two different structures regarding sections (Figure 22 and 23).

13. — HOTELS, CASINOS, THERMALISME

	NF		NF							Gr. 3		
16.000	50	....	.. ..	53,70	5/6/61	Eaux Minérales et Bains de Mer, act. (ex-droits, ex-c. 2 et 3 du 10/1/61) ... ×	517889	.....	.....	56 ..	52,25
2.000	50	....	.. ..	52,70	16/6/61	— Promesses (Jce 1/1/61) ... I	85234	.....	.....	54 ..	52 ..
....	...	....	.. ..	0,15	10/2/61	— Droits de souscription (c. 2) *Ivrables* ...	85234	....	.....	0,25	0,10
....	...	....	.. ..	6,10	13/7/61	— Droits d'attribution (nos 1 à 16000) (c. 3) ...	89235	....	.....	7 ..	6,10
173.400	50	1/6/61	6,84	760 ..	....	Evian-les-Bains (S.A. Eaux Min. d'), act. (ex-c. 13).	3402	755	.....	900 ..	650 ..
15.000	300	3/7/61	57,50	2738 ..	....	Grand Hôtel, act. (ex-c. 10) ... ×	563579	2710	.....	3298	2195
59.115	100	5/6/61	B 7,35	336 ..	....	Lutétia (Hôtel), act. (ex-c. 14) ... ×	3777	350	.....	385 ..	284 ..

Figure 22: Example of a section title outside of columns in Brussels price list



K. — Industries de la Construction

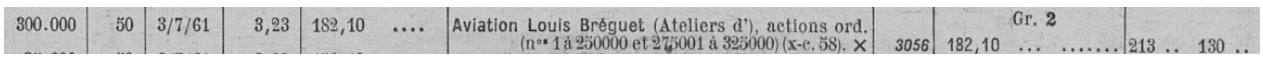
8335	1078	Carrières de Quenast (1re, 3e et 4 ..)	3	janv.juillet	500	octobre	485.		485
1785	207	» » (2e émiss.)	3	avril oct.	500	octobre			494
450	269	» Tacquenier	5	janv.juillet	1000	avril			1004
2000	1432	» P.-J. Wincqz	4 ½	15 mai nov.	500	avril			457 50
700	648	Ciments Europe Orientale	4 ½	janv. juill.	500				493
7000	5462	» North	4	15 mrs sept.	500	mai			355
5000	3820	Merbes-le-Château	4	janv.juillet	500	novembre			501
3600	3800	Niel-on-Rupell	4	avril oct.	500	décembre			480
2000	400	Tuileries du Poltelberg	5	31 déc.	500	février			

Figure 23: Example of a section title inside of a column in Parquet price list

When the section title is outside of a column, we have to look for new sections in all the width of the table. In this case we know as well that there should not be columns separators next to a section title. We developed this variation and, with the generic aspect of our system, we only have to specify depending on the corpus, whether their sections titles are outside or inside columns

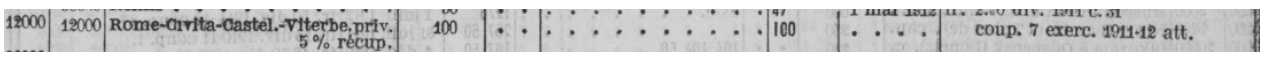
Multiple lines

Our data extractor was previously analysing a table line by line and it was not possible for one stock to be on multiple lines. In our three corpora, we saw that it was possible to see titles on multiple lines. We currently described two other types of line: *staggered lines* (Figure 24) and *stock on multiple lines* (Figure 25).



300.000	50	3/7/61	3,23	182,10	....	Aviation Louis Bréguet (Ateliers d'), actions ord. (nos 1 à 250000 et 275001 à 325000) (x-c. 58). ×	3056	Gr. 2 182,10		213 ..	130 ..

Figure 24: Example of a staggered line from Parquet price list



12000	12000	Rome-Civita-Castel.-Viterbe.priv. 5 % récup.		100				100		coup. 7 exerc. 1911-12 att.	

Figure 25: Example of a stock on multiple line from Brussels price list

This project has received funding from

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

We need to be able to detect when a title contains data on multiple lines in order not to split one stock in two.

In the case of staggered lines, we identify them thanks to their particular structure: the first half is written on the first line, the second half on the last line and the stock name is written in several lines in between.

Stock on multiple lines consist of a stock written on only one line except for the name which can be on several lines. In order to detect this case, we have to check under each stock if there are lines only containing data in the stock name column, and decide whether we should concatenate these or not.

When adding multiple lines structures, we had to think as well about the 'idem' symbol (Figure 26).


Aviation Louis Bréguet (Ateliers d'), actions ord. (nos 1 à 250000 et 275001 à 325000) (x-c. 58). ×
— Actions B de priorité (nos 250001 à 275000) (ex-c. 58)..... ×

Figure 26: Example of an 'idem' symbol (dash) after a stock on multiple lines from Parquet price list

In our original data extractor, we worked on documents where we had dash symbols under the part that had to be repeated. Here in the figure above, we have only one dash symbol standing for *Aviation Louis Bréguet*. For the first version of our system, we decided to repeat the entire first line of the multiple line because we did not have the required data to identify and separate the base stock name from data linked to it.

Titles

Another structure that we had to deal with were titles. Once we have the cells composing the meta-structure, we apply our data extraction system in each of these as seen previously. Sometimes these cells can contain a framed title or a blank space rather than a table (Figure 27).

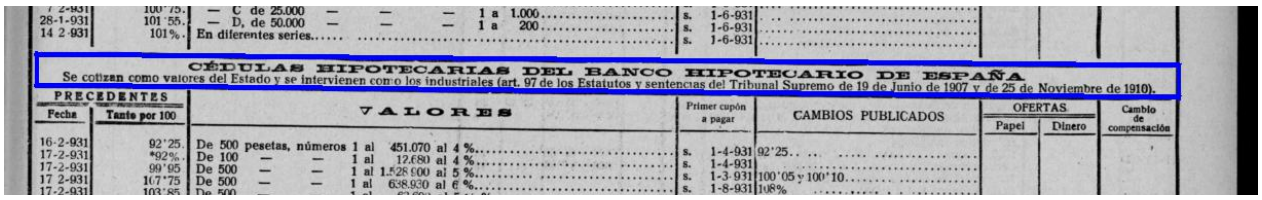

CÉDULAS HIPOTECARIAS DEL BANCO HIPOTECARIO DE ESPAÑA
Se cotizan como valores del Estado y se intervienen como los industriales (art. 97 de los Estatutos y sentencias del Tribunal Supremo de 19 de Junio de 1907 y de 25 de Noviembre de 1910).

Figure 27: Example of a title identified in blue from Madrid price list

Instead of ignoring these, we added in our system the possibility to identify titles in addition to tables. We can thereafter apply a specific method on titles.

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

3.3.2 Evaluation result

3.3.2.1 Columns & Headers recognition and validation

We test the interest of the transversal analysis strategy (see section 3.1) for the correction of recognition errors in noisy documents. We chose a subset of the collection of "La Coulisse" (French unofficial market). We select the first page of each day of quotation from 1899 to 1915. This subset is composed of 4,055 images.

We consider the results produced by step (1) alone (results before validation) and the results produced by the whole strategy (results after validation). We count the number of errors i.e., the number of documents for which the recognized structure is wrong.

		French pricelists (1899 to 1915)
	Num. of pages	4, 055
Grammatical description alone (step (1))	Num. of errors Error rate	320 7.89 %
Whole validation strategy	Num. of errors Error rate	18 0.44%

Table 29: Evaluation on the Grammatical and Strategy analysis

The experiment shows that thanks to our strategy we reduced the error rate from 7.89% to 0.44%. Figure 28 shows a qualitative example of improvement obtained with our strategy where columns not detected when processed as an isolated page, are detected using the collection context. A limitation of our system is that it does not manage to correctly recognize the structure when the degradation of the documents has an impact on column width (see Figure 29). This configuration explains the remaining 18 errors on a total of 4,055 pages.

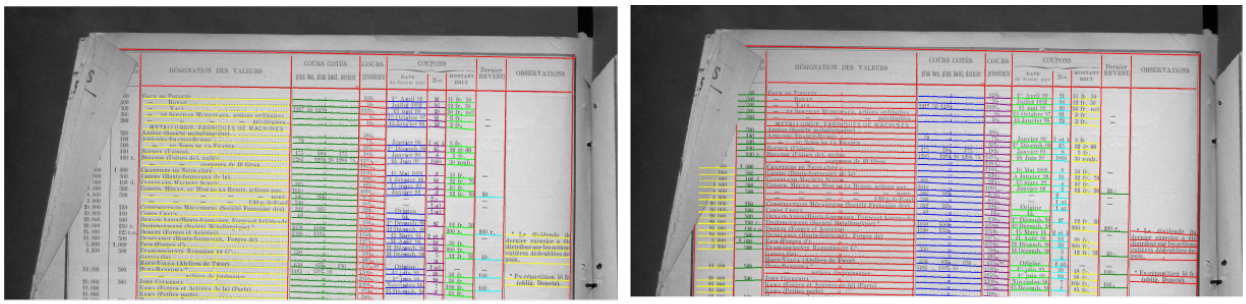
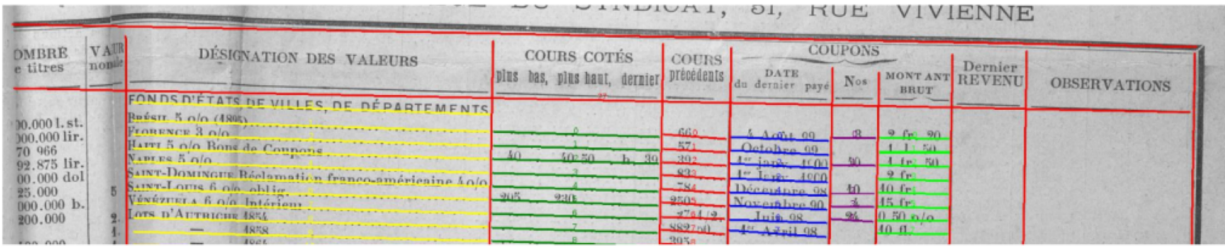


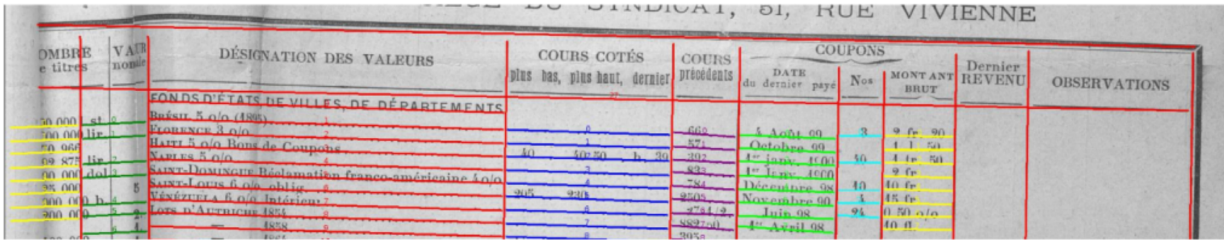
Figure 28: Example of improvement obtained with our strategy (left: before validation; right: after validation)

This project has received funding from

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



(a) Grammatical description alone (before validation)



(b) Global strategy (after validation)

Figure 29: Example of a limit case of our strategy (only 18 errors for 4, 055 pages)

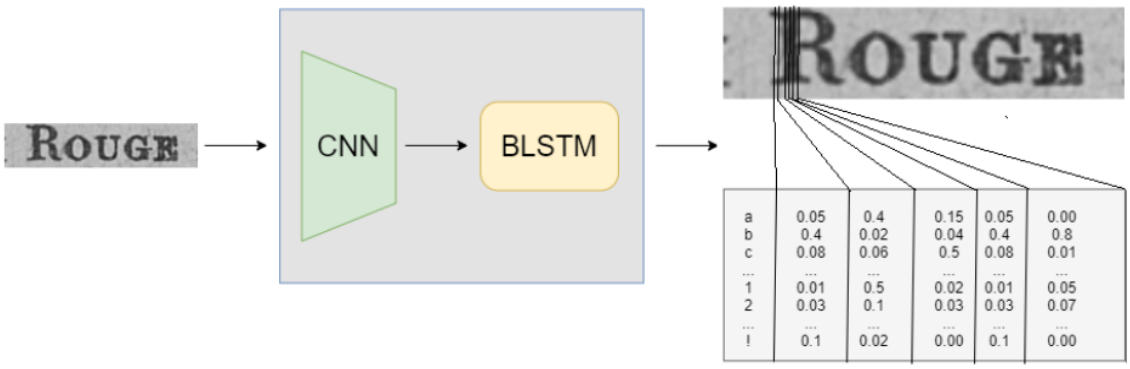
This project has received funding from

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

3.5 General-purpose text recognizer (OCR)

3.5.1 System overview

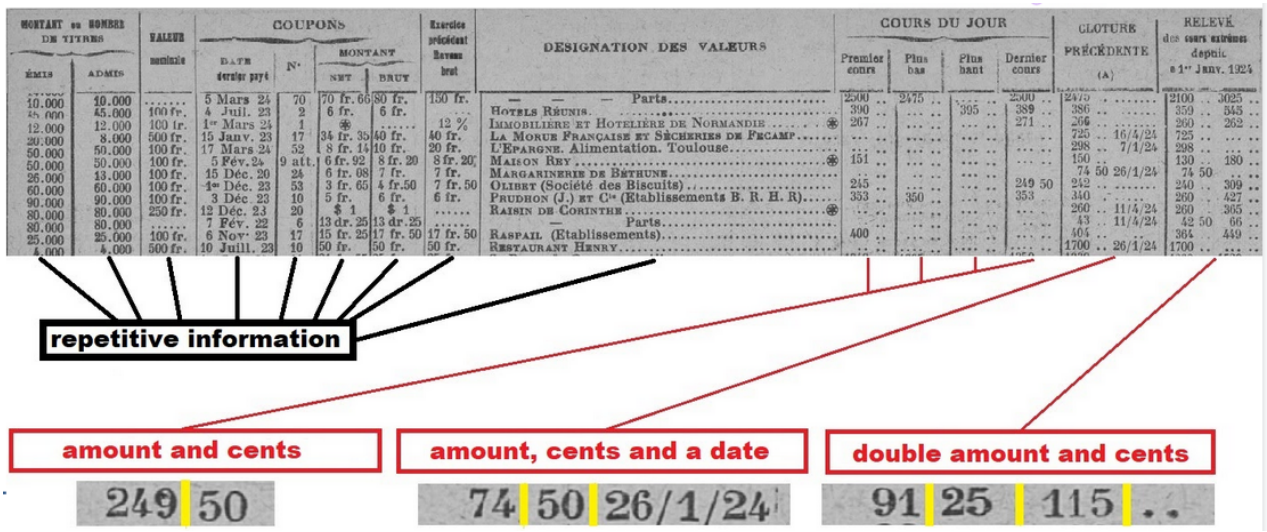
In D7.2 we have demonstrated that high performance character recognition can be achieved when large amounts of training data are available. The strategy is based on specializing a generic OCR to a dedicated corpus for which transcriptions are available so that supervised training of a deep neural network architecture can be pushed to its highest limits in terms of performance. The figure below shows the recognition pipeline which is composed of a deep neural network made of Convolutional Layers (CNN) followed by Recurrent Layers made of Bilateral Long Short Term Memory Networks (BLSTM). This hybrid neural network provides a lattice of character hypotheses with their associated probability. Two recognition modes are available on the system. The raw OCR output is made of the sequence of characters with the highest probabilities.



The Grammar OCR output (see figure below) is the sequence of characters allowed by the grammar, with the highest probability. This second recognition mode discards any sequence of characters that is not allowed by the grammar. There is one specific grammar for each specific column of the price lists. In some cases, the grammar simply encodes a list of possible names, or values. This is typically the case for every information that occurs similarly every day over a very long period of time whereas in some other cases, some other values are different every day, but they represent prices, dates, or any kind of numerical values that is encoded in a specific format. A grammar describes the encoding used in the price lists for any such column.

This project has received funding from

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



The table below recalls the performance for the different modes of the recognizer for some experimentations made on the French corpus *La Coulisse* that were reported in the previous report D.7.2.

OCR	Character Error Rate
OMNIPAGE Nuance	26.64%
LITIS OCR V0 (hand labeled + OMNIPAGE)	12.83%
LITIS OCR V1 (hand labeled, 40 pages, 53 642 fields)	1.52%
LITIS OCR V1 + Grammar	0.55%

Table 30: Evaluation of different OCRs

3.5.2 Data augmentation

During this last period of the project, we have upgraded the capability of the OCR by adding what is called "data augmentation". This method applies random treatments to the data before giving it to the system to learn. The principle lies in creating more data from the available ones. This means that instead of manually labelling 100,000 images, we can label a much smaller number of samples and then increase their number by augmenting the data, applying some random transformations on them during training. In the following we present the different treatments that are included in the augmentation process with an example and an explanation on how it helps.

This project has received funding from

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

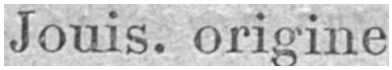
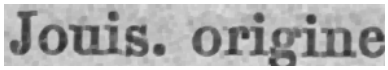
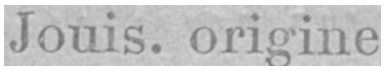
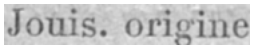
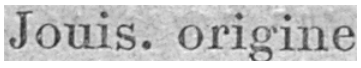
Original image	
3.5.2.1 Erosion/Dilatation	
3.5.2.2 Lightening / Contrast modification	 
3.5.2.3 Change of Resolution (DPI change)	
3.5.2.4 Elastic Distortion	
3.5.2.5 Gaussian noise	
3.5.2.6 Bounding Box modification	
3.5.2.7 Image sharpness modification	

Table 31: Examples of every data augmentation technique

3.5.2.8 Evaluation

In order to prove how data augmentation helps to train a better OCR, we present the following result table in a test made with the *Coulisse dataset* as a benchmark. We observe a 23% relative improvement of the CER which can be considered to be very significant at this high level of performance.

OCR	Character Error Rate
LITIS OCR V1 (hand labelled, 40 pages, 53 642 fields)	1.52%
LITIS OCR V2 (hand labelled, 40 pages, 53 642 fields + Data augmentation)	1.17%

Table 32: Evaluation on the comparison between OCRs with and without data augmentation

This project has received funding from

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

3.5.3 Final experiment results

3.5.3.1 Datasets and Evaluation

As was planned in our agenda, we conducted a full evaluation of LITIS OCR on the different corpora of our work program. The table below reports the amount of training data that were annotated manually and then used for training the OCR either specifically on one dataset when sufficient annotated data are available or by mixing the training datasets when not enough data is available. The best character error rate (CER) is reported for each corpus of our EURHISFIRM benchmark in addition to the performance obtained on the *La Coulisse* corpus, which was made available to us as an extra dataset for comparison purposes. In the results reported we use the raw OCR outputs without any Language Model, and we compute an average performance whatever the type of field considered (stock names, prices etc…).

Corpus	Number of annotated field images	Character Error Rate
French *La Coulisse*	53 642	1.17%
French *Le Parquet*	7 146	1.61%
Belgium *Bruxelles*	35 810	1.88%
Spain *Madrid*	7 705	2.84%

Table 33: Number of labeled data per corpus and Error rate on the corpus

In the following, we report the detailed performance of the OCR for each and every condition of training (without / with data augmentation (DA)) and decoding (without / with language model (LM)), and for each corpus of the benchmark.

This project has received funding from

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

Coulisse Price List	CER Raw	CER LM	CER Raw + DA	CER DA + LM
Number of Titles	1.48%	1.05%	1.10%	0.53%
Nominal Value	1.55%	0.94%	1.13%	0.60%
Name of Stock	2.73%	1.09%	1.39%	0.72%
Prices of the day	11.90%	4.56%	2.12%	1.10%
Yesterday's Price	8.89%	7.50%	1.95%	0.98%
Last day of payment of the Coupon	1.94%	0.65%	0.81%	0.22%
Numbers of the Coupon	0.76%	0.76%	0.70%	0.52%
Gross Amount of the Coupon	0.93%	0.07%	0.84%	0.07%
Last Income	1.03%	0.00%	0.69%	0.09%

Table 34: Evaluation on the Coulisse Corpus per column using the raw and Language Module outputs

Belgium Price List	CER Raw + DA	CER DA + LM
Number of stocks admitted	1.54%	1.10%
Number of active stocks	2.03%	1.02%
Name of Stock	2.83%	0.91%
Nominal Value	1.24%	0.85%
Dates	1.47%	0.84%
Prices of the day	2.25%	1.52%
Previous price	1.56%	1.04%
Last day of payment of the Coupon	1.93%	1.02%
Amount and specification of the coupon	0.96%	0.60%

Table 35: Evaluation on the Belgium Corpus per column using the raw and Language Module outputs

This project has received funding from

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

Parquet Price List	CER Raw + DA	CER DA + LM
Number of stocks admitted	1.77%	1.44%
Nominal Value	0.67%	0.20%
Ex-coupon date	1.25%	1.02%
Coupon net value	1.03%	0.71%
Previous prices	2.52%	1.82%
Name of Stock	2.44%	0.95%
Stock's code number	1.42%	0.54%
Spot prices	2.81%	1.42%
Extreme prices since January 1st	2.19%	1.24%

Table 36: Evaluation on the Parquet Corpus per column using the raw and Language Module outputs

Madrid Price List	CER Raw + DA	CER DA + LM
Date of previous	2.48%	1.88%
Percentage	2.44%	1.75%
Name of Stock	2.98%	1.10%
Nominal value	3.33%	1.89%
Payout	0.98%	0.88%
Date of last paid dividends	2.10%	1.50%
Current year of last paid dividends	1.02%	0.7%
First coupon to be pay from the last paid dividends	2.62%	1.53%
Public changes	2.02%	1.62%
Change of compensation	1.23%	0.82%

Table 37: Evaluation on the Spanish Corpus per column using the raw and Language Module outputs

This project has received funding from

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

3.6 Evaluation and Application of the Price List Information Extraction System to four price lists

We evaluate the complete price lists information extraction system, built on the combination of the global meta table structure recognition (section 3.2), the reading order recognition (section 3.2), the price lists tables recognition (section 3.3), the General-purpose text recognizer (OCR) (section 3.4), driven by the Strategy and transversal analysis (section 3.1).

The evaluation is done on Paris "La Coulisse" price lists 1899 in the context of the French ANR HBDEX project, where the complete extraction of the data combined with expert users interaction and the insertion of all the extracted data in the DFIH database have been done.

3.6.1 Evaluation of Stock identification on Paris "La Coulisse" 1899, 1696 stock lines

We evaluate the complete price lists information extraction system on French price lists from "La Coulisse". We have processed the first 6 months of 1899 and we choose 4 days of quotation each composed of 4 pages (16 images) for the evaluation. This represents a total of 1696 stock lines.

We evaluate the quality of stock identification:

1. without any consideration of the context of the collection
2. with consideration of the context of the collection but without user interactions
3. with consideration of the context of the collection and with user interactions (the whole proposed strategy)

	without collection context	with collection context without user interactions	collection context + user interaction
Nb of true positive	1550	1580	1675
Nb of false positive	147	117	13
Nb of false negative	146	116	21
Precision	0.913	0.931	0.989
Recall	0.914	0.932	0.988
F-measure	0.914	0.931	0.988

Table 38: Evaluation on the Paris La Coulisse 1899 Corpus using the collection context and user interaction

This project has received funding from

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

The experimentation (results presented in table above) shows that our strategy improves **the F-measure from 0.914 without collection context to 0.988 with the collection context and expert user interaction**.

3.6.2 Data Extraction on 6 months of Paris "La Coulisse" 1899 price lists, 536 pages

This quality of data extraction (F-measure of 0.988) is done while, **on 536 pages and 54,603 stock lines from 6 months of La Coulisse 1899, the number of questions to expert users is drastically reduced from 4,061 to 309 with the collection context.** Indeed, to obtain results of the same quality without the exploitation of the context of the collections, 4,061 questions would have been necessary.

On these **536 pages of 6 months of daily quotation, a total of 491, 427 cells, from 54,603 stock lines has been extracted and produced in XML (see Figure 6) after 309 expert users interaction.** All this data has then been inserted in the DFIH database.

This experiment on Paris "La Coulisse" 1899 validates the ability of the Price Lists Information Extraction System to extract all the data found in price lists with a high quality of recognition, while minimizing the expert users interaction.

	without collection context	with collection context
Nb of days	134	134
Nb of images (6 month of quotations)	536	536
Nb of stock lines	54,603	54,603
Nb of questions required to obtained a F-measure ≥ 0.988	4,061	309

Table 39: Paris La Coulisse 1899: Interest of the collection context for drastically reduce expert users interactions

This project has received funding from

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

3.6.3 Application of the Price Lists Information Extraction System on EurHisFirm dataset

We have applied the Price List Information Extraction on the price lists EurHisFirm dataset. We present examples of results on the combination of the global meta table structure, the reading order recognition, the column and header detection, the section detection and the stock lines detection, on Brussels 1912 (Figure 30 and 31), on Madrid 1931 (Figure 32 and 33) and on Paris "Le Parquet" 1961-1962 (Figure 34 and 35).

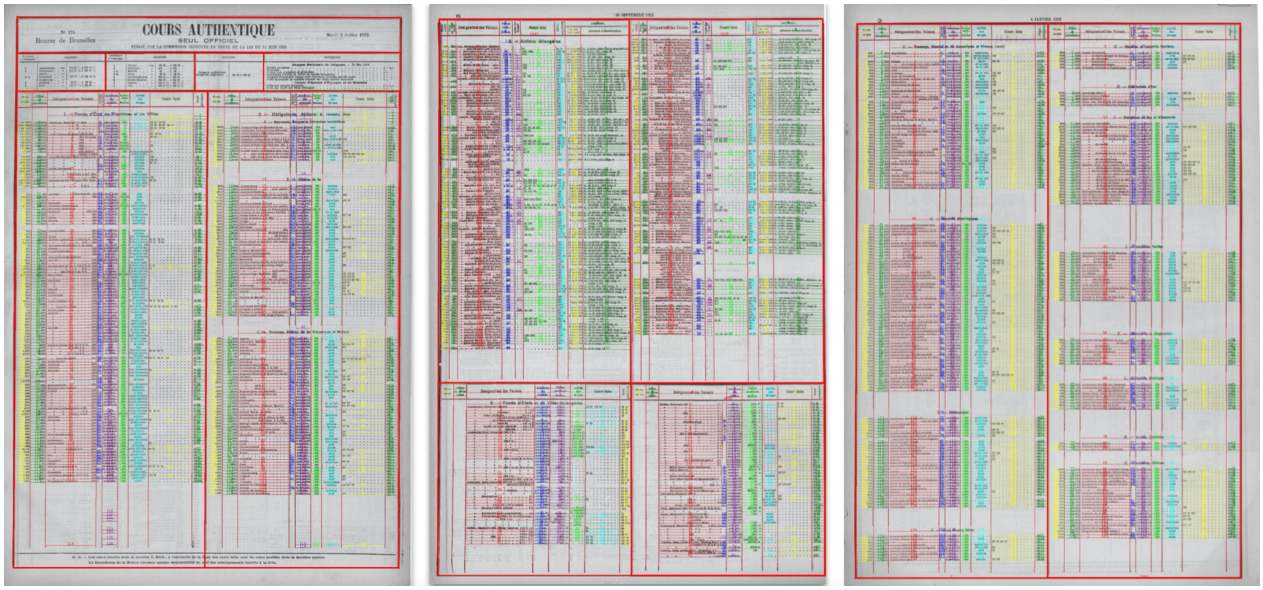


Figure 30: Price Lists Brussels 1912: Results on three pages of global meta table recognition, reading order recognition, price list column and header recognition in each cell

This project has received funding from

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

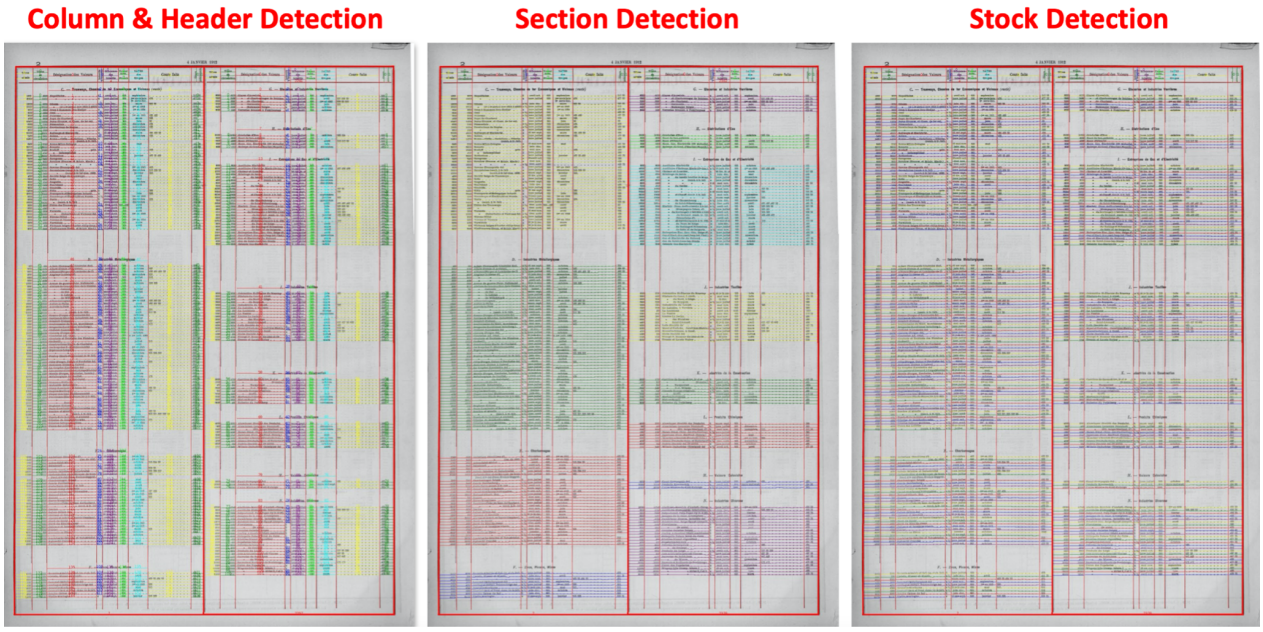


Figure 31: Price Lists Brussels 1912: Results on one page of global meta table recognition, reading order recognition with: price list column and header recognition in each cell, each color represents one column (left); section detection, each color represents one section (middle); stock detection, each color represents one stock (right)

This project has received funding from

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

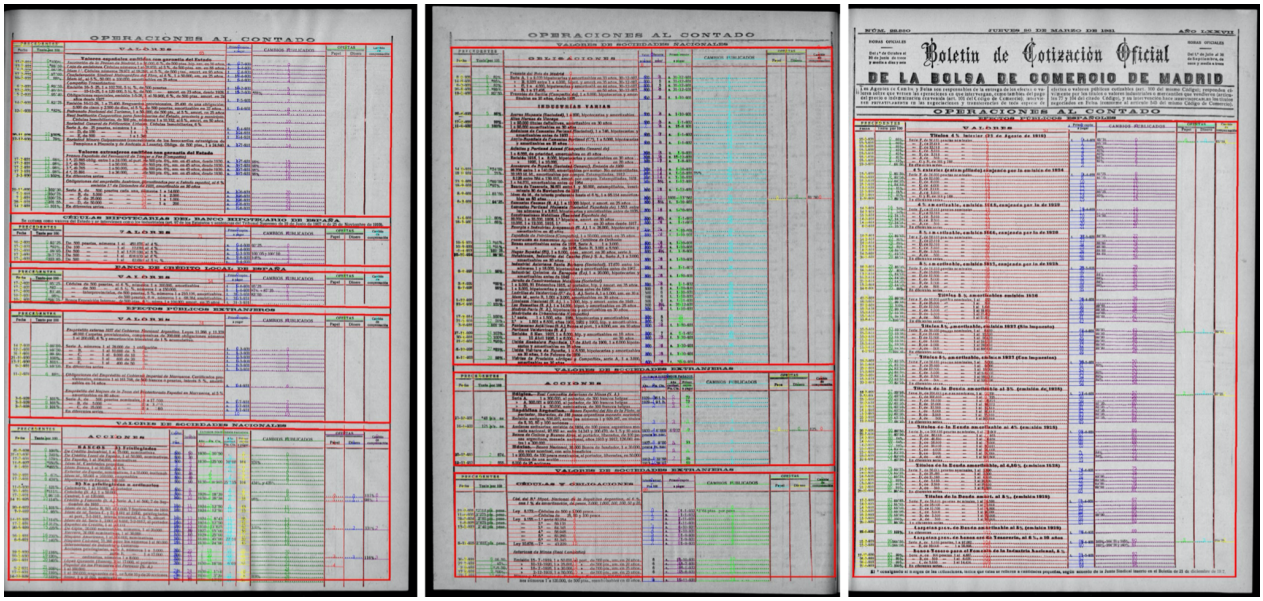


Figure 32: Price Lists Madrid 1931: Results on three pages of global meta table recognition, reading order recognition, price list column and header recognition in each cell

Column & Header Detection **Section Detection** **Stock Detection**

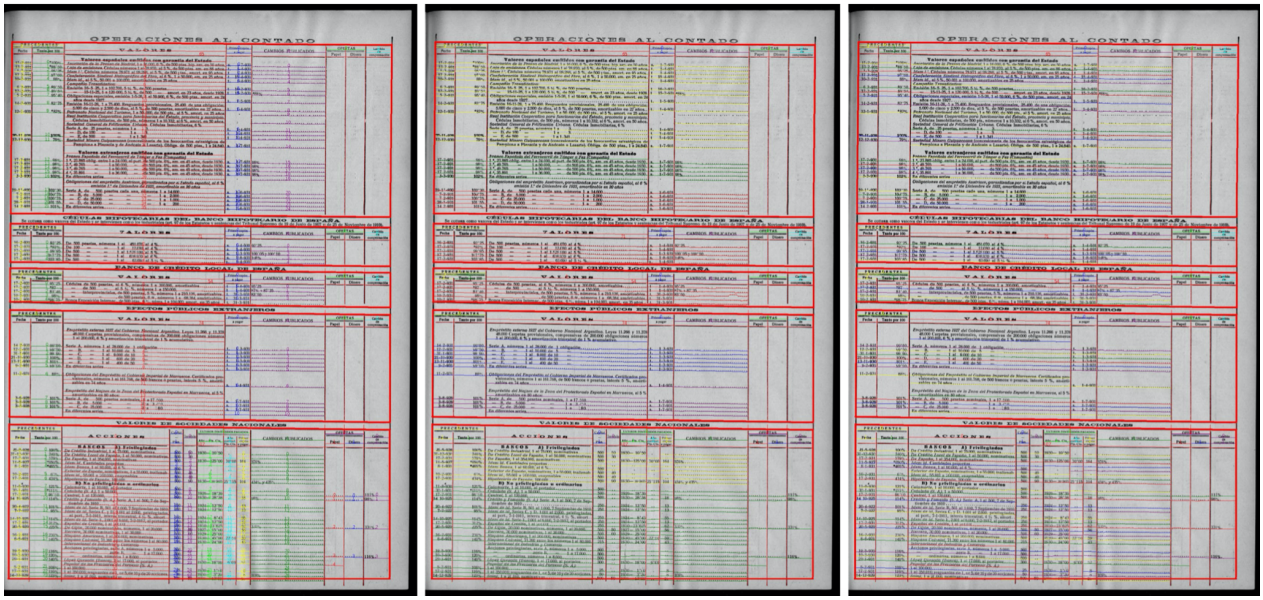


Figure 33: Price Lists Brussels 1912: Results on one page of global meta table recognition, reading order recognition with: price list column and header recognition in each cell, each color represents one column (left); section detection, each color represents one section (middle); stock detection, each color represents one stock (right)

This project has received funding from

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

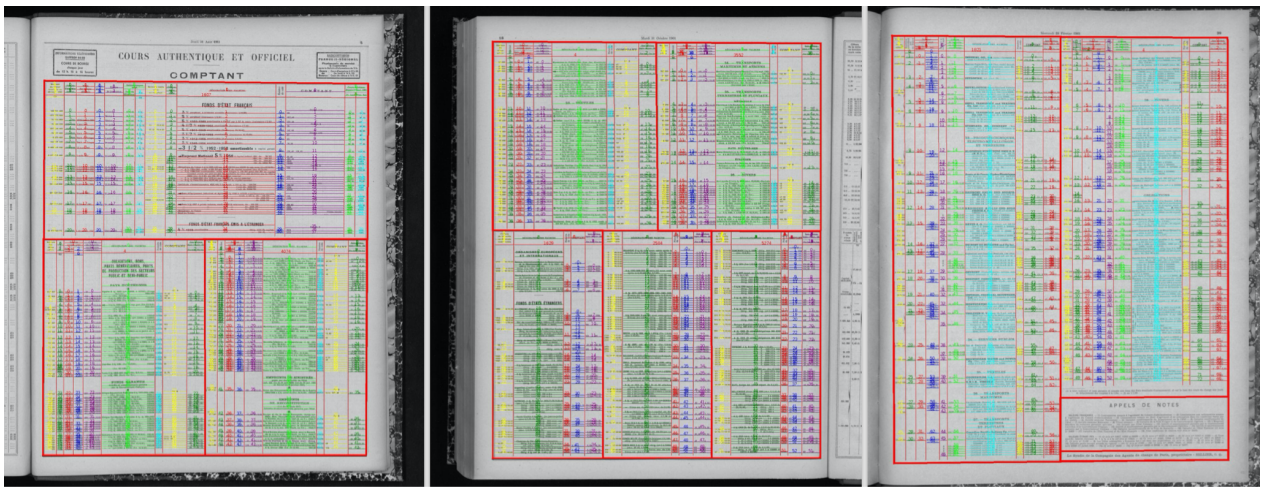


Figure 34: Price Lists Paris "Le Parquet" 1961-1962: Results on three pages of global meta table recognition, reading order recognition, price list column and header recognition in each cell

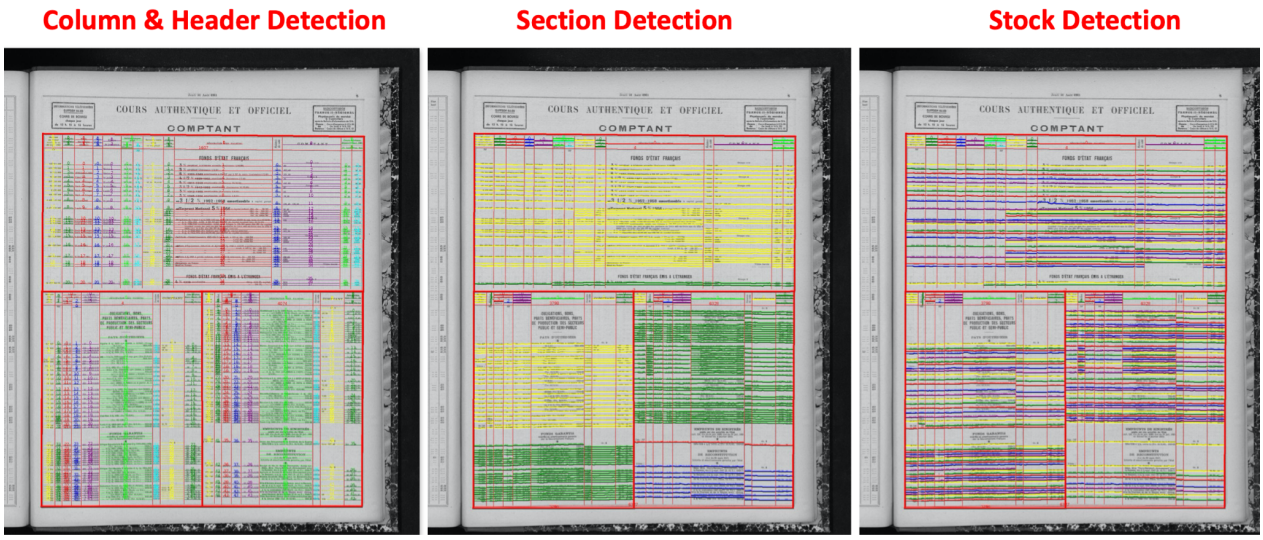


Figure 35: Price Lists Paris "Le Parquet" 1961-1962: Results on one page of global meta table recognition, reading order recognition with: price list column and header recognition in each cell, each color represents one column (left); section detection, each color represents one section (the green section starts in cell 2.1 and ends in cell 2.2)(middle); stock detection, each color represents one stock (right)

This project has received funding from

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

The WP7 Price Lists Information Extraction System has been applied on 1,755 pages of the four price lists datasets produced 176,382 extracted stocks (mainly domestic shares and bounds). More detailed results by price list are given in table 40.

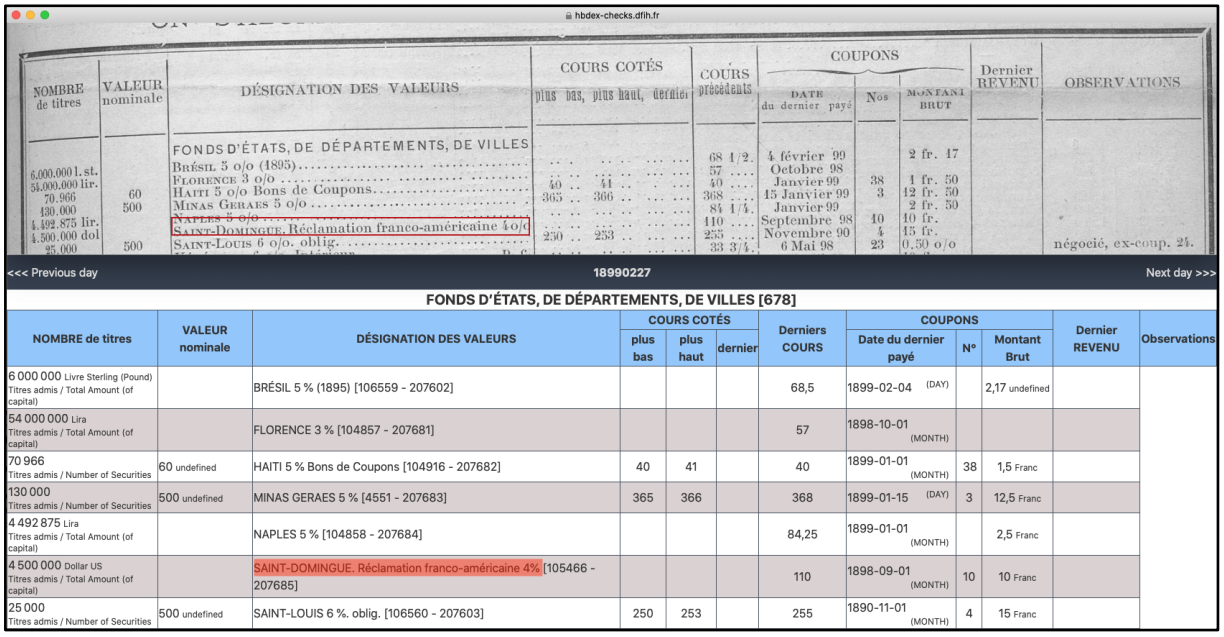
Results of the WP7 Price Lists Information Extraction System		**Paris "La Coulisse" 1899**	**Brussels 1912**	**Madrid 1931**	**Paris "Le Parquet" 1961-1962**
Num. of pages		536	268	534	417
Domestic shares & bounds	**Num. of extracted stocks**	**56,603**	**56,350**	**14,562**	**48,083**
	Mean num. of stocks by day		1,466	600	1, 936
	Num. of sections and subsections		1,197		2 ,128
	Mean num. of section and subsections by day		31		85
Public bounds	**Num. of extracted stocks**				**784**
	Mean num. of stocks by day				31
	Num. of sections and subsections				49
	Mean num. of section and subsections by day				2

Table 40: WP7 global results on the Price Lists

On the price list Paris "La Coulisse" 1899, the WP7 Information Extraction System produced data with an estimated F-measure of 0.988. The processing of **536 pages (6 months of 1899) generated 54,603 stock lines while the number of questions to expert users was reduced from 4,061 to 309 with the collection context.** All the data extracted has been produced in XML (see Figure 36) for an import in the DFIH database.

This project has received funding from

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



FONDS D'ÉTATS, DE DÉPARTEMENTS, DE VILLES [678]

NOMBRE de titres	VALEUR nominale	DÉSIGNATION DES VALEURS	COURS COTÉS plus bas	COURS COTÉS plus haut	COURS COTÉS dernier	Derniers COURS	COUPONS Date du dernier payé	COUPONS N°	COUPONS Montant Brut	Dernier REVENU	Observations
6 000 000 Livre Sterling (Pound) Titres admis / Total Amount (of capital)		BRÉSIL 5 % (1895) [106559 - 207602]				68,5	1899-02-04 (DAY)		2,17 undefined		
54 000 000 Lira Titres admis / Total Amount (of capital)		FLORENCE 3 % [104857 - 207681]				57	1898-10-01 (MONTH)				
70 966 Titres admis / Number of Securities	60 undefined	HAITI 5 % Bons de Coupons [104916 - 207682]	40	41		40	1899-01-01 (MONTH)	38	1,5 Franc		
130 000 Titres admis / Number of Securities	500 undefined	MINAS GERAES 5 % [4551 - 207683]	365	366		368	1899-01-15 (DAY)	3	12,5 Franc		
4 492 875 Lira Titres admis / Total Amount (of capital)		NAPLES 5 % [104858 - 207684]				84,25	1899-01-01 (MONTH)		2,5 Franc		
4 500 000 Dollar US Titres admis / Total Amount (of capital)		SAINT-DOMINGUE. Réclamation franco-américaine 4% [105466 - 207685]				110	1898-09-01 (MONTH)	10	10 Franc		
25 000 Titres admis / Number of Securities	500 undefined	SAINT-LOUIS 6 %. oblig. [106560 - 207603]	250	253		255	1890-11-01 (MONTH)	4	15 Franc		

Figure 36: Example of the WP7 Price lists data extraction system: 27 February 1899 - La Coulisse, Data extracted in XML (bottom) after the process of the page (top) in the context of a collection of 6 months of quotations. Each extracted data is linked to its location in the image (see in red "SAINT-DOMINGUE")

4 WP7 Information Extraction System demonstration

The results of the application of the WP7 Information Extraction System on the German Yearbook 1913-1914 (Handbuch der deutschen Aktiengesellschaften), the Spanish Yearbook Madrid 1931 and the French Desfossés Yearbook 1962 are presented with the WP7 Data Extraction Viewer on this web site:

- http://litis-eurhisfirm.univ-rouen.fr/pivan/

The results of the application of the WP7 Information Extraction System on official price lists Brussels 1912, Madrid 1931 and Paris 1961-1962, are presented with the WP7 Data Extraction Viewer on this web site:

- http://litis-eurhisfirm.univ-rouen.fr/pivan/

The XML results of the application of the WP7 Information Extraction System on price lists Paris "La Coulisse" 1899, done in the French National project ANR HBDEX, are presented on this web site (see Figure 36):

- https://hbdex-checks.dfih.fr/batches/1899S1_current/18990116

This project has received funding from

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

5 References

[Coüasnon 2006] Coüasnon, B. DMOS, a generic document recognition method: application to table structure analysis in a general and in a specific way. *IJDAR* **8,** 111–122 (2006). https://doi.org/10.1007/s10032-005-0148-5

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

[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*.

[Lemaitre 2009] Aurélie Lemaitre, Bertrand Coüasnon, Jean Camillerapp. Use of Perceptive Vision for Rulling Recognition in Ancient Documents. Eighth IAPR International Workshop on Graphics RECognition (GREC 2009), Jul 2009, La Rochelle, France. pp.3-12. ⟨hal-00545052⟩

[Pawlik and Augsten 2016] Mateusz Pawlik, Nikolaus Augsten, Tree edit distance: Robust and memory-efficient, Information Systems, Volume 56, 2016, Pages 157-173, ISSN 0306-4379, https://doi.org/10.1016/j.is.2015.08.004.

This project has received funding from

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