http://www.eurhisfirm.eu

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

Long-term data for Europe

EURHISFIRM

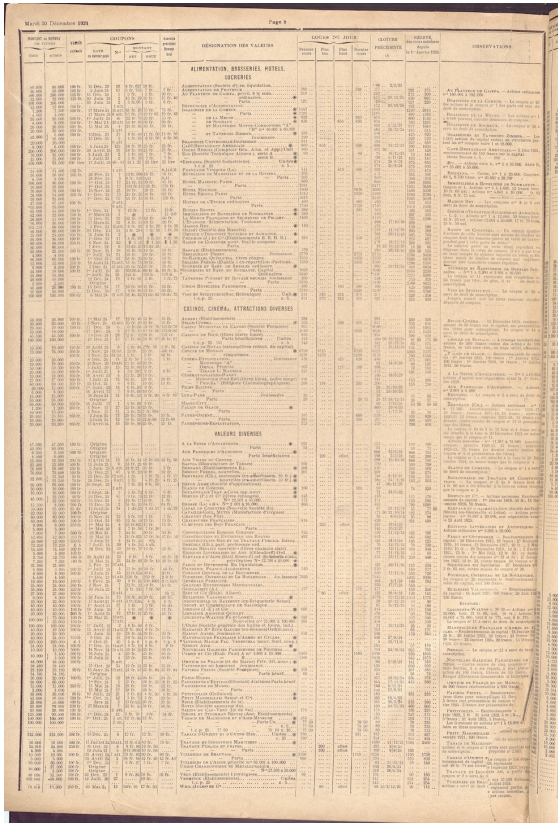
Working Package 7
Data extraction and enrichment system

General Assembly
16 March 2019

Bertrand Coüasnon, IRISA / INSA Rennes
Thierry Paquet, LITIS / University Rouen-Normandie
Gabriel Schneider, Hebis / Goethe University
Paris School of Economics

WP7 - Objectives

- Prototype development
  - Intelligent and collaborative system
  - Extraction of structured information
  - From Images of historical documents
    - Related to companies' financial and economic activities
- Printed sources
  - Securities price lists / Yearbooks
- Extraction
  - Dates, Person names, Company names
  - Numerical values of securities prices…
- Flexible and generic system
  - Cross checking
    - Between pages of documents
    - External sources (databases, web-based resources)



http://www.eurhisfirm.eu

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

WP7 - Selected Document Dataset

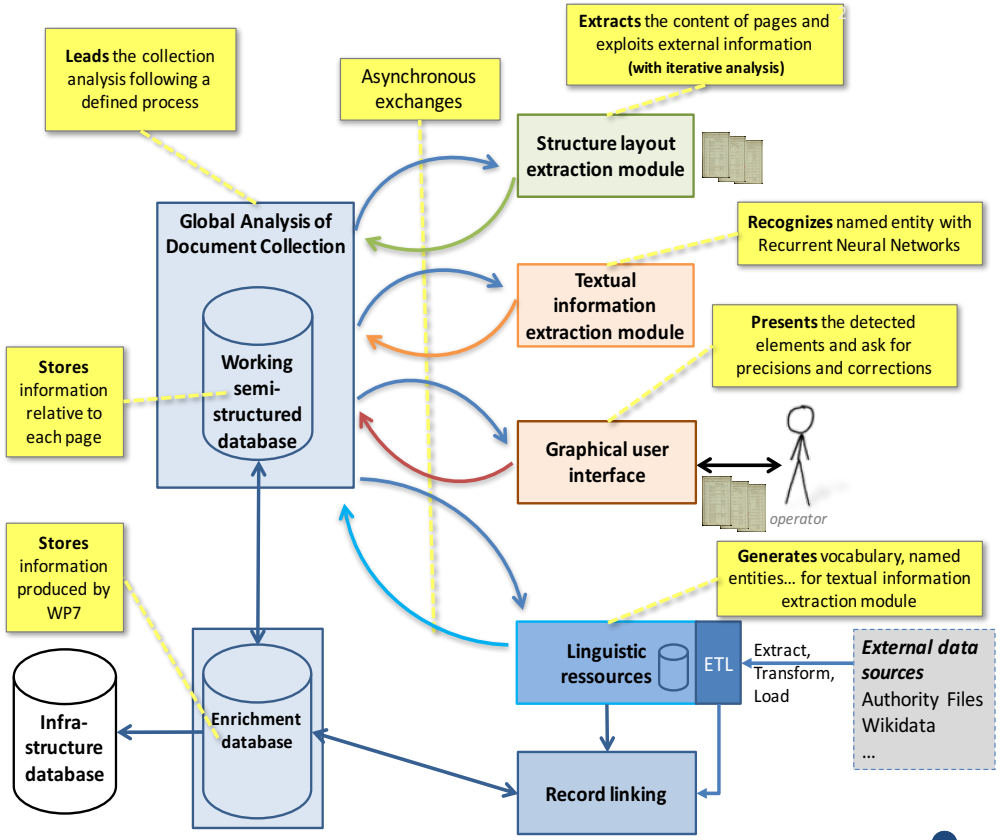
- **Document samples dataset**
  - **Build, validate, evaluate the recognition system**
  - **Selected by the steering committee**
  - **Among more than 30 yearbooks and stock price lists**
  - **Representative of difficulties, document quality**
- **Selection**
  - **3 yearbooks, 3 securities price lists, 3 languages, 4 countries**

TIME PERIOD	YEARBOOKS	SECURITIES PRICE LISTS
Before WWI	Germany 1914-15 Handbuch	Belgium (in French) 1875, 1878 (or 1912 - under investigation)
Interwar	Spain 1929-1930	Spain 1934
Post WWII	France (web-linking). (Desfossés 1962)	France (web-linking). (Cote : 1 July 1961 - 30 June 1962)

http://www.eurhisfirm.eu
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

Architecture Design of an Adaptable System

- Combination
  - Layout Analysis
  - Textual Information Extraction
  - Linguistic Resources from Data Web Linking
- Local Semi-Structured Database
  - Results organized by page
- Enrichment Database
  - Results of WP7
  - Ready to be Inserted in the the Common Data Model of WP5

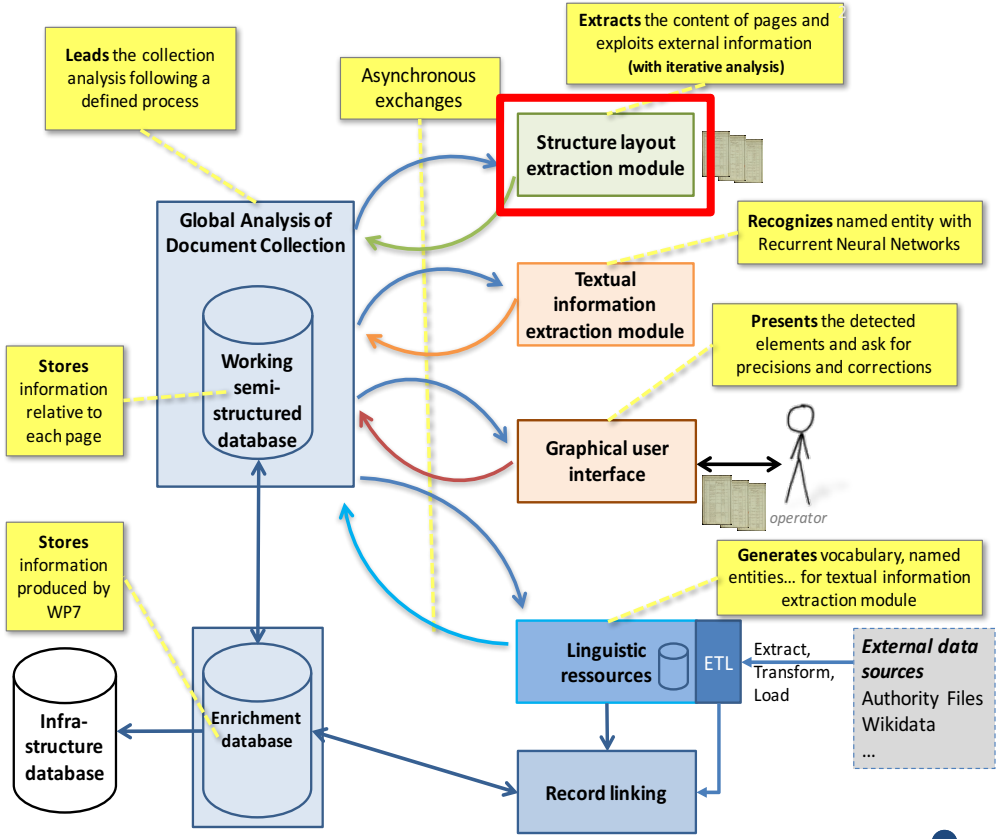


http://www.eurhisfirm.eu
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

Architecture Design of an Adaptable System

Task 7.2: Irisa, Insa Rennes

- Combination
  - Layout Analysis
  - Textual Information Extraction
  - Linguistic Resources from Data Web Linking
- Local Semi-Structured Database
  - Results organized by page
- Enrichment Database
  - Results of WP7
  - Ready to be Inserted in the the Common Data Model of WP5



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

Task 7.2: Library of Document Components Detectors

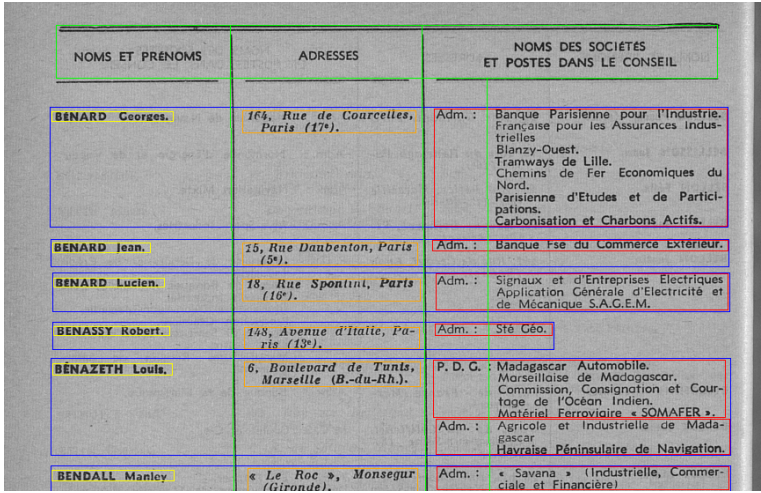
- Detection
  - Table rulings
  - White separators
  - Fragmented text line detection
- Combination of
  - Deep learning
  - Syntactical method
- Library shared with French national project: HBDEX
- Example on French securities price list (Le Parquet, 1962)

http://www.eurhisfirm.eu
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

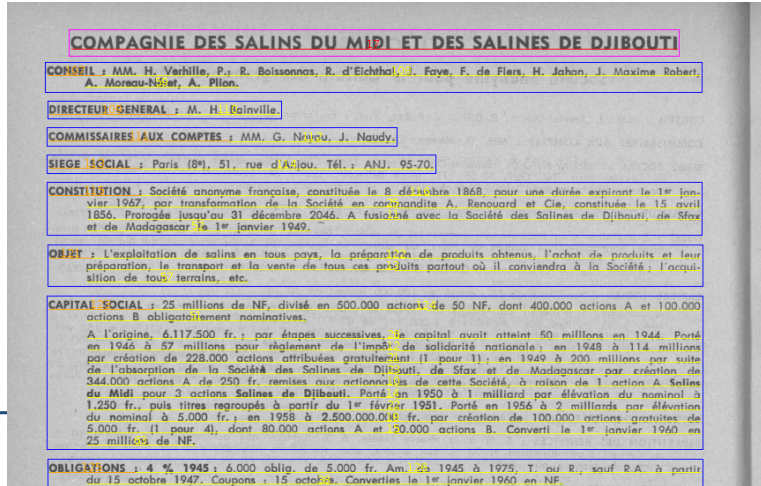
HeBIS GOETHE UNIVERSITÄT FRANKFURT AM MAIN UMR IRISA

Task 7.2 - Results on Structure Layout Extraction

- Structure recognition on French yearbooks (Desfossés, 1962)
  - Description of the structure
  - Detection of administrators
    - Name
    - Address
    - Function, company
  - Detection of emitters
    - Name
    - Rubric
      - Title
      - Set of paragraphs



NOMS ET PRÉNOMS	ADRESSES	NOMS DES SOCIÉTÉS ET POSTES DANS LE CONSEIL	
BENARD Georges.	165, Rue de Courcelles, Paris (17e).	Adm. :	Banque Parisienne pour l'Industrie. Française pour les Assurances Industrielles Blanzy-Ouest. Tramways de Lille. Chemins de Fer Economiques du Nord. Parisienne d'Etudes et de Participations. Carbonisation et Charbons Actifs.
BENARD Jean.	15, Rue Daubenton, Paris (5e).	Adm. :	Banque Fse du Commerce Extérieur.
BENARD Lucien.	18, Rue Spontini, Paris (16e).	Adm. :	Signaux et d'Entreprises Electriques Application Générale d'Electricité et de Mécanique S.A.G.E.M.
BENASSY Robert.	148, Avenue d'Italie, Paris (13e).	Adm. :	Sté Géo.
BÉNAZETH Louis.	6, Boulevard de Tunis, Marseille (B.-du-Rh.).	P. D. G. :	Madagascar Automobile. Marseillaise de Madagascar. Commission, Consignation et Courtage de l'Océan Indien. Matériel Ferroviaire « SOMAFER ».
		Adm. :	Agricole et Industrielle de Madagascar Havraise Péninsulaire de Navigation.
BENDALL Manley	« Le Roc », Monsegur (Gironde).	Adm. :	« Savana » (Industrielle, Commerciale et Financière)



COMPAGNIE DES SALINS DU MIDI ET DES SALINES DE DJIBOUTI

CONSEIL : MM. H. Verhille, P. R. Boissonnas, R. d'Eichthal, J. Faye, F. de Flers, H. Jahan, J. Maxime Robert, A. Moreau-Néret, A. Plion.

DIRECTEUR GÉNÉRAL : M. H. Bainville.

COMMISSAIRES AUX COMPTES : MM. G. Noyou, J. Naudy.

SIÈGE SOCIAL : Paris (8e), 51, rue d'Anjou. Tél. : ANJ. 95-70.

CONSTITUTION : Société anonyme française, constituée le 8 décembre 1868, pour une durée expirant le 1er janvier 1967, par transformation de la Société en commandite A. Renouard et Cie, constituée le 15 avril 1856. Prorogée jusqu'au 31 décembre 2046. A fusionné avec la Société des Salines de Djibouti, de Sfax et de Madagascar le 1er janvier 1949.

OBJET : L'exploitation de salins en tous pays, la préparation de produits obtenus, l'achat de produits et leur préparation, le transport et la vente de tous ces produits partout où il conviendra à la Société ; l'acquisition de tous terrains, etc.

CAPITAL SOCIAL : 25 millions de NF, divisé en 500.000 actions de 50 NF, dont 400.000 actions A et 100.000 actions B obligatoirement nominatives.

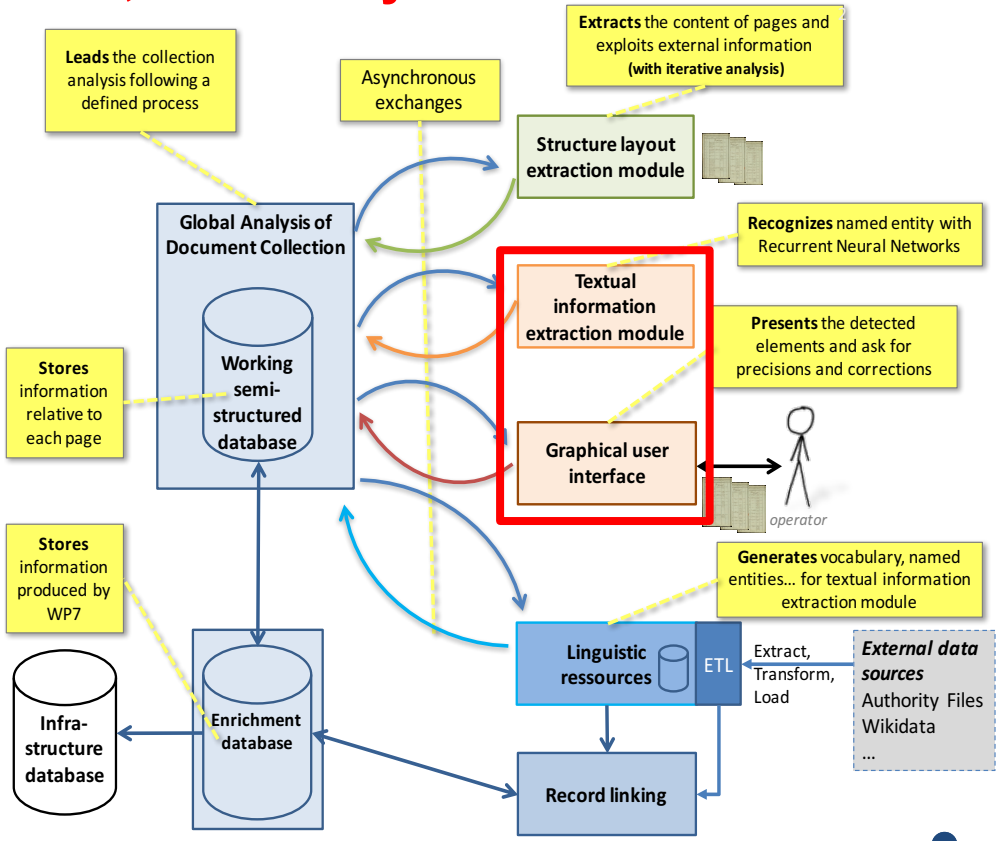
A l'origine, 6.117.500 fr. ; par étapes successives, le capital avait atteint 50 millions en 1944. Porté en 1946 à 57 millions pour règlement de l'impôt de solidarité nationale ; en 1948 à 114 millions par création de 228.000 actions attribuées gratuitement (1 pour 1) ; en 1949 à 200 millions par suite de l'absorption de la Société des Salines de Djibouti, de Sfax et de Madagascar par création de 344.000 actions A de 250 fr. remises aux actionnaires de cette Société, à raison de 1 action A Salines du Midi pour 3 actions Salines de Djibouti. Porté en 1950 à 1 milliard par élévation du nominal à 1.250 fr., puis titres regroupés à partir du 1er février 1951. Porté en 1956 à 2 milliards par élévation du nominal à 5.000 fr. ; en 1958 à 2.500.000.000 fr. par création de 100.000 actions gratuites de 5.000 fr. (1 pour 4), dont 80.000 actions A et 20.000 actions B. Converti le 1er janvier 1960 en 25 millions de NF.

OBLIGATIONS : 4 % 1945 : 6.000 oblig. de 5.000 fr. Am. de 1945 à 1975. T. au R., sauf R.A. à partir du 15 octobre 1947. Coupons : 15 octobre. Converties le 1er janvier 1960 en NF.

Architecture Design of an Adaptable System

Task 7.3: Litis, University of Rouen

- Combination
  - Layout Analysis
  - Textual Information Extraction
  - Linguistic Resources from Data Web Linking
- Local Semi-Structured Database
  - Results organized by page
- Enrichment Database
  - Results of WP7
  - Ready to be Inserted in the the Common Data Model of WP5



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

Task 7.3: Optical Character Recognition (OCR)

- Neural Network encodes the input image and proposes appearance probabilities for all possible characters in the image
- Language Model helps to constraint the recognition hypotheses to the most linguistically correct ones

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

Task 7.3: Information extraction levels

- Level 1 : 1st paragraph contains company name

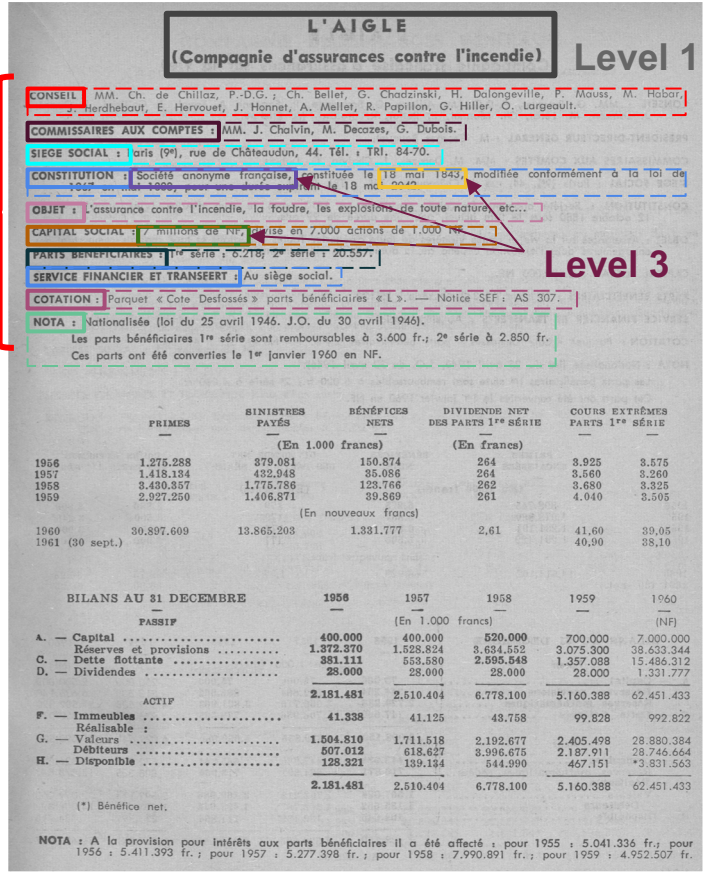
  Ex. L'AIGLE (Compagnie d'assurances contre l'incendie)

  - Level 2 : set of paragraphs identified by keywords that specify the type of entity to be extracted.

    Ex1. CONSTITUTION
    Ex2. CAPITAL SOCIAL

    - Level 3 : Text passage containing Named Entities specifying the (multiple) attributes of each entity

      Ex1. Company status : Société anonyme Française
      Start date : 18 mai 1843
      Capital amount : 7 millions de NF



L'AIGLE
(Compagnie d'assurances contre l'incendie)

CONSEIL : MM. Ch. de Chillaz, P.-D.G. ; Ch. Bellet, G. Chadzinski, H. Dalongeville, P. Mauss, M. Habar, J. Herdhebaut, E. Hervouet, J. Honnet, A. Mellet, R. Papillon, G. Hiller, O. Largeault.

COMMISSAIRES AUX COMPTES : MM. J. Chalvin, M. Decazes, G. Dubois.

SIEGE SOCIAL : Paris (9e), rue de Châteaudun, 44. Tél. : TRI. 84-70.

CONSTITUTION : Société anonyme française, constituée le 18 mai 1843, modifiée conformément à la loi de 1867 et du 1 août 1889, pour une durée expirant le 18 mai 2043.

OBJET : L'assurance contre l'incendie, la foudre, les explosions de toute nature, etc...

CAPITAL SOCIAL : 7 millions de NF, divisé en 7.000 actions de 1.000 NF.

PARTS BENEFICIAIRES : 1re série : 6.218; 2e série : 20.557.

SERVICE FINANCIER ET TRANSFERT : Au siège social.

COTATION : Parquet « Cote Desfossés » parts bénéficiaires « L ». — Notice SEF : AS 307.

NOTA : Nationalisée (loi du 25 avril 1946. J.O. du 30 avril 1946).
Les parts bénéficiaires 1re série sont remboursables à 3.600 fr.; 2e série à 2.850 fr.
Ces parts ont été converties le 1er janvier 1960 en NF.

	PRIMES	SINISTRES PAYÉS	BÉNÉFICES NETS	DIVIDENDE NET DES PARTS 1re SÉRIE	COURS EXTRÊMES PARTS 1re SÉRIE	
	—	—	—	—	—	—
		(En 1.000 francs)		(En francs)		
1956	1.275.288	379.081	150.874	264	3.925	3.575
1957	1.418.134	432.948	85.086	264	3.560	3.260
1958	3.430.357	1.775.786	129.766	262	3.680	3.325
1959	2.927.250	1.406.871	39.869	261	4.040	3.505
		(En nouveaux francs)				
1960	30.897.609	13.865.203	1.331.777	2,61	41,60	39,05
1961 (30 sept.)					40,90	38,10

BILANS AU 31 DECEMBRE	1956	1957	1958	1959	1960
	—	—	—	—	—
PASSIF		(En 1.000 francs)			(NF)
A. — Capital	400.000	400.000	520.000	700.000	7.000.000
Réserves et provisions	1.372.370	1.528.824	3.634.552	3.075.300	38.633.344
C. — Dette flottante	381.111	553.580	2.595.548	1.357.088	15.486.312
D. — Dividendes	28.000	28.000	28.000	28.000	1.331.777
	2.181.481	2.510.404	6.778.100	5.160.388	62.451.433
ACTIF					
F. — Immeubles	41.338	41.125	43.758	99.828	992.822
Réalisable :					
G. — Valeurs	1.504.810	1.711.518	2.192.677	2.405.498	28.880.384
Débiteurs	507.012	618.627	3.996.675	2.187.911	28.746.664
H. — Disponible	128.321	139.134	544.990	467.151	*3.831.563
	2.181.481	2.510.404	6.778.100	5.160.388	62.451.433

(*) Bénéfice net.

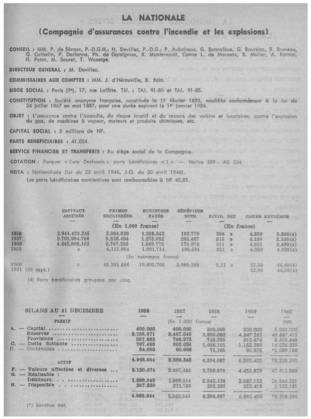
NOTA : A la provision pour intérêts aux parts bénéficiaires il a été affecté : pour 1955 : 5.041.336 fr.; pour 1956 : 5.411.393 fr.; pour 1957 : 5.277.398 fr.; pour 1958 : 7.990.891 fr.; pour 1959 : 4.952.507 fr.

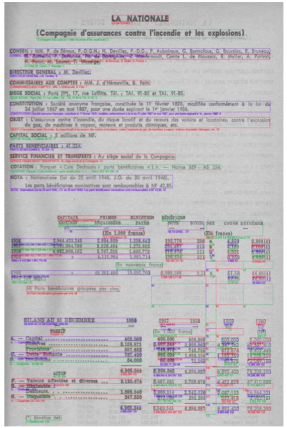
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

Task 7.3: Rule based information extraction pipeline

Extraction rules + keywords approximation = Handcrafted Extraction Patterns

OCR for document image → Paragraph extraction → Level 1&2 information extraction → Level 3 information extraction → extracted named entities


LA NATIONALE
(Compagnie d'assurances contre l'incendie et les explosions)


LA NATIONALE
(Compagnie d'assurances contre l'incendie et les explosions)

Entity name	Parag. text
Company name	1st parag. text
CONSEIL	Text
CAPITAL	Text
...	...

Type of Named entities	Named entities detected
Company name	L'AIGLE (Compagnie d'assurances contre l'incendie)
Name	P. de Séroux
Function	P.-D.G.H.
...	...

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

Task 7.3: Extraction primary results

http://www.eurhisfirm.eu

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

Level 3 evaluation : These results have been obtained by manually annotating 100 paragraphs for each level 2 keyword

Level 2 keywords	Level 3 Named Entities	Precision TP/(TP+FP)	Recall TP/(TP+FN)	F1-score
Conceil (alternatives)	Admin_name	99.36%	98.52%	98.94%
	Admin_function	92.59%	98.68%	95.54%
Siege social	departement	96.84%	92.93%	94.85%
	town	83.13%	90.79%	86.79%
	address	98.75%	81.44%	89.27%
Constitution capital	initial company state	100.00%	100.00%	100.00%
	start date	100.00%	99.00%	99.50%
	end date	99.02%	96.19%	97.58%
	company transformation type	100.00%	92.00%	95.83%
	transformation date	52.63%	55.56%	54.05%
	Initial capital	92.31%	92.31%	92.31%
service financier et transfert	Location	30.00%	60.00%	40.00%
	Institution name	72.47%	97.73%	83.23%
OVER ALL		**85.93%**	**88.86%**	**86.76%**

Task 7.3: Extraction primary results

- Precise specification → good extraction

Level 2 : Constitution
Level 3 : successful extraction

Société anonyme française, constituée en 1873, sous le nom d'Atlas (Vie), transformée le 12 octobre 1880 sous le titre actuel, pour une durée de 99 ans.

initial company status (Green)
start date(Red)
end date(Blue)
transform. status (Brown)
transform. date (Gray)

Level 2 keywords	Level 3 Named Entities	Precision TP/(TP+FP)	Recall TP/(TP+FN)	F1-score
Conceil (alternatives)	Admin_name	99.36%	98.52%	98.94%
	Admin_function	92.59%	98.68%	95.54%
Siege social	departement	96.84%	92.93%	94.85%
	town	83.13%	90.79%	86.79%
	address	98.75%	81.44%	89.27%
Constitution	initial company state	100.00%	100.00%	100.00%
	start date	100.00%	99.00%	99.50%
	end date	99.02%	96.19%	97.58%
	company transformation type	100.00%	92.00%	95.83%
	transformation date	52.63%	55.56%	54.05%
capital	Initial capital	92.31%	92.31%	92.31%
service financier et transfert	Location	30.00%	60.00%	40.00%
	Institution name	72.47%	97.73%	83.23%
OVER ALL		**85.93%**	**88.86%**	**86.76%**

http://www.eurhisfirm.eu
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

Task 7.3: Extraction primary results

http://www.eurhisfirm.eu

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

Level 2 : Service financier et transfert

Level 3 : successful extraction

Paris : **Crédit Lyonnais**

Location.(**Green**)

Financial institution name (**Blue**)

Level 3 : False positive extraction

Actions : **Siège social** ; **obligations** : **B.N.C.I.**

Location.(**Green**),

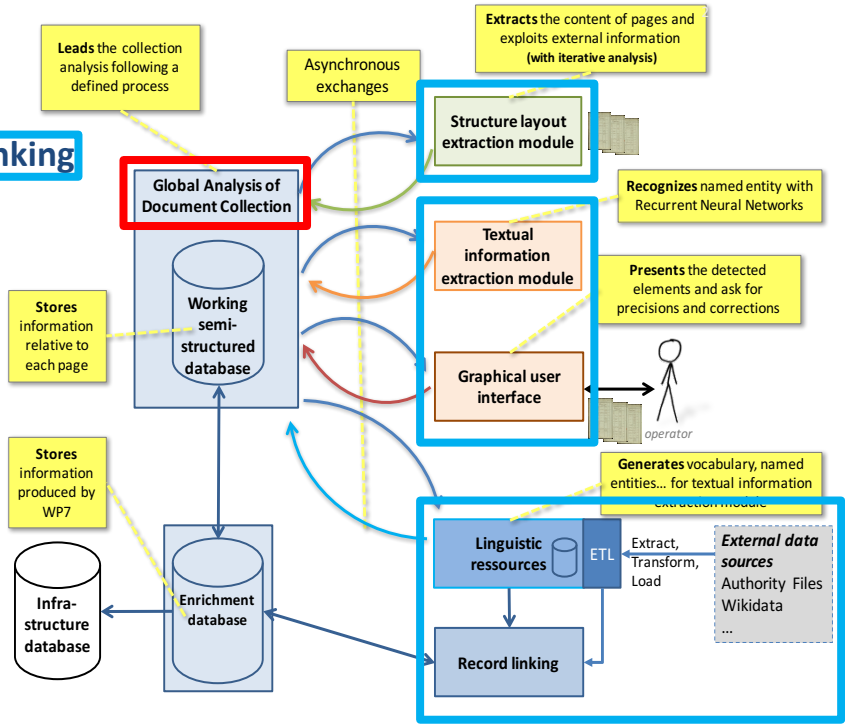
financial institution name (**Blue**)

Add more specification (specialization) to the extraction pattern by considering an enriched mixture of typographical and and lexical features.

Level 2 keywords	Level 3 Named Entities	Precision TP/(TP+FP)	Recall TP/(TP+FN)	F1-score
Conceil (alternatives)	Admin_name	99.36%	98.52%	98.94%
	Admin_function	92.59%	98.68%	95.54%
Siege social	departement	96.84%	92.93%	94.85%
	town	83.13%	90.79%	86.79%
	address	98.75%	81.44%	89.27%
Constitution capital	initial company state	100.00%	100.00%	100.00%
	start date	100.00%	99.00%	99.50%
	end date	99.02%	96.19%	97.58%
	company transformation type	100.00%	92.00%	95.83%
	transformation date	52.63%	55.56%	54.05%
	Initial capital	92.31%	92.31%	92.31%
service financier et transfert	Location	30.00%	60.00%	40.00%
	Institution name	72.47%	97.73%	83.23%
OVER ALL		**85.93%**	**88.86%**	**86.76%**

Future Work : Task 7.1 Global Analysis of Collections

- Combination
  - Task 7.2: Structure layout extraction
  - Task 7.3: Textual information extraction
  - Task 7.4: Linguistic Resources from Data Web Linking
- Objective
  - Maximize recognition quality
  - Minimize user interaction
- Iterative analysis
  - Multiple analysis of the same page
  - Asynchronous interaction
  - Progressive integration of contextual information
- Cross-checking among pages
  - Transverse analysis
  - Detect stability/break on textual information
  - Check coherence with financial knowledge formalization on securities values



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

Future Work: Task 7.1, 7.2 - Global and Page Analysis

- Iterative analysis: at each iteration
  - Process a collection of pages
  - Layout Analysis
    - Grammatical description of pages (Task 7.2)
    - Generate all similar names in the collection
  - Sequence analysis
    - Done by textual extraction (Task 7.3)
    - Detect stable names, new names
    - Improve recognition quality
    - Link with unique ID from infrastructure database

- Example on stock price lists
  - Iteration 1: Recognition of columns names + Sequence analysis
  - Iteration 2: Recognition of section names + Sequence analysis
  - Iteration 3: Recognition of securities names + Sequence analysis
  - Iteration 4: Recognition of securities prices + Sequence browse left to right
    - Financial rules => Check coherence of securities prices
  - Iteration 5: Recognition of securities prices + Sequence browse right to left
    - Financial rules => Check coherence of securities prices

http://www.eurhisfirm.eu
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

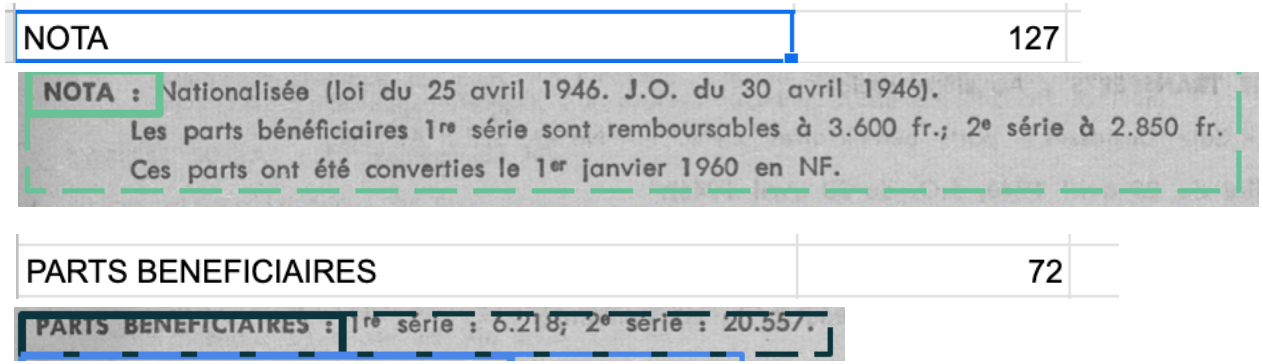
Future Work : Task 7.3 Textual Information Extraction

http://www.eurhisfirm.eu

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

The developed modules will be bricked together to create the process in which

- **the user will take a central place to validate the discovered new items that were not specified**



- **while the machine will specialize its extraction patterns through active learning exploiting the false positive patterns rejected by the user**

Future Work: Task 7.3 - System Workflow

http://www.eurhisfirm.eu
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 777489

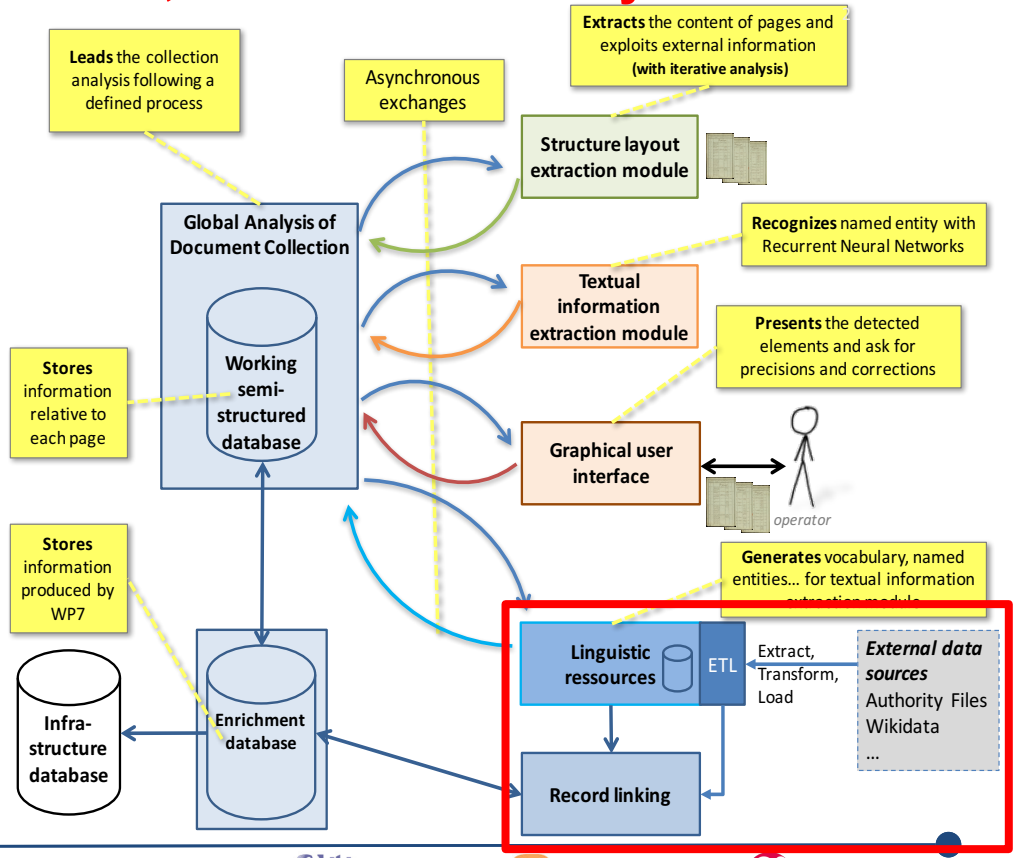
1) The user will introduce new knowledge to the system, by annotating new discovered items.
2) The new validated data will be introduced in the lexical database in order for the system to enrich the current extraction patterns
3) The system will benefit from the rejected items by learning new extraction patterns
4) Use some internal validation rules (Lexicons , cross-references within the corpus) whenever possible. Or have external cross-referencing (Hebis)

- From Handcrafted Extraction Patterns to Learnable Extraction Patterns
- Requires less specification efforts at the price of labelling (annotating) some examples
- Generalize to any language
- Assume OCR performs sufficiently well to run OCR and extraction sequentially

Architecture Design of an Adaptable System

Task 7.4: Hebis, Goethe University

- Combination
  - Layout Analysis
  - Textual Information Extraction
  - Linguistic Resources from Data Web Linking
- Local Semi-Structured Database
  - Results organized by page
- Enrichment Database
  - Results of WP7
  - Ready to be Inserted in the the Common Data Model of WP5



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

WP 7.4

Automated Linking of Named Entities

Gabriel Schneider, HeBIS

Agenda

- Introduction
- Using a linking framework
- Results
- Challenges
- Outlook

Introduction

- Different trusted sources in the web contain relevant data
  - → Enrichment of EURHISFIRM data
- → Identification of same entities in different data sources

Introduction

- Manual linking is time-consuming & prone to errors
  - → automated approach
- → Usage of a linking framework

Using a linking framework

- Linked Data Linking Framework
- Linking of two sources at a time
- Comparison of entities
  - → computation of a sameness score
- Linking result als linked data triple

Using a linking framework

- Different measures available
  - Measures for: strings, vector spaces, point-set, topological, temporal
- Combination of multiple attributes & measures possible

Results

- Data from University of Antwerpen vs. Paris School of Economics

	U. Antwerpen	**PSE**
Number of records	2370	7682
Number of supposed matches	144	
Number of linking framework matches	58 supposed matches (~ 40%) + 95 other matches + 183 „reviews“	

Results

- Linking to GND, DBpedia and Wikidata started

	s	p	o
1	<http://hebis.de/eurhisfirm/CFS-000098>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	schema:Organization
2	<http://hebis.de/eurhisfirm/CFS-000098>	<http://hebis.de/eurhisfirm/externalId>	"BAYHANDE"
3	<http://hebis.de/eurhisfirm/CFS-000098>	dc:source	"CFS"
4	<http://hebis.de/eurhisfirm/CFS-000098>	schema:identifier	"CFS-000098"
5	<http://hebis.de/eurhisfirm/CFS-000098>	schema:legalName	"Bayerische Handelsbank"
6	<http://hebis.de/eurhisfirm/CFS-000098>	<http://hebis.de/eurhisfirm/sector>	"Bank"
7	<http://hebis.de/eurhisfirm/CFS-000098>	owl:sameAs	<http://d-nb.info/gnd/17306-X>

Challenges

- Sources in different languages
- Varying data quality
- Resolution of ambiguities
  - → usage of more attributes (address data etc.)
- → Need for standardised data

Outlook

- Infrastructure for testing automated linking available
- Optimisation of configuration for evaluation of linking framework
  - → reference data needed

Outlook

- Continuation of linking to authority files (GND, GLEI, VIAF, ISNI etc.)
- Integration of infrastructure into project workflows
- Other possible use cases