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  - From Images of historical documents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)
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

<table>
<thead>
<tr>
<th>TIME PERIOD</th>
<th>YEARBOOKS</th>
<th>SECURITIES PRICE LISTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before WWI</td>
<td>Germany 1914-15 Handbuch</td>
<td>Belgium (in French)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1875, 1878 (or 1912 - under investigation)</td>
</tr>
<tr>
<td>Interwar</td>
<td>Spain 1929-1930</td>
<td>Spain 1934</td>
</tr>
</tbody>
</table>
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 Common Data Model of WP5

---

**Global Analysis of Document Collection**

**Working semi-structured database**

- **Stores information relative to each page**
- **Stores information produced by WP7**

**Enrichment database**

**Infra-structure database**

- **Extracts** the content of pages and exploits external information (with iterative analysis)
- **Recognizes** named entity with Recurrent Neural Networks
- **Presents** the detected elements and ask for precisions and corrections
- **Generates** vocabulary, named entities... for textual information extraction module

**Graphical user interface**

**Asynchronous exchanges**

**Structure layout extraction module**

**Textual information extraction module**

**Linguistic resources**

**ETL**

**Record linking**

**External data source**

- Authority Files
- Wikidata...
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 Common Data Model of WP5

[Diagram showing the architecture design with various components and their interactions.]

- **Global Analysis of Document Collection**
- **Working semi-structured database**
- **Infra-structure database**
- **Enrichment database**
- **Linguistic resources**
- **ETL**
- **Record linking**

- **Structure layout extraction module**
- **Asynchronous exchanges**
- **Textual information extraction module**
- **Graphical user interface**
- **Operator**
- **Generates vocabulary, named entities... for textual information extraction module**
- **Extracts the content of pages and exploits external information (with iterative analysis)**
- **Recognizes named entity with Recurrent Neural Networks**
- **Presents the detected elements and ask for precisions and corrections**

- **Stores information relative to each page**
- **Stores information produced by WP7**
- **Extract, Transform, Load**
- **Authority Files : Wikidata**

- **Long-term data for Europe**

---

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)
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

---

<table>
<thead>
<tr>
<th>NOMS ET PRÉNOMS</th>
<th>ADDRESSES</th>
<th>NOMS DES SOCIÉTÉS ET POSTES DANS LE CONSEIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BENARD Gérard</td>
<td>16, Rue de Courcelles, Paris (18)</td>
<td>Adm. Banque Particulière pour l'Industrie, financée par les Assurances Indemnité-Banque de l'Ile-de-France, 16, Rue St-Pontin, Lyon (69)</td>
</tr>
<tr>
<td>BENARD Jean</td>
<td>16, Rue Dumbarton, Paris (17)</td>
<td>Adm. Banque de Commerce Extérieur</td>
</tr>
<tr>
<td>BENARD Louis</td>
<td>83, Rue St-Pontin, Paris (100)</td>
<td>Adm. Simeaux et Jardineurs Electriques Application Générale d'Electricité et de Mécanisme S.A.M.E.</td>
</tr>
<tr>
<td>BENASSI Robert</td>
<td>29, Avenue d'Italie, Paris (19)</td>
<td>Adm. Mél F.C.</td>
</tr>
<tr>
<td>BÉNAZETH Louis</td>
<td>6, Rue Raynouard de Pons, Marseille (B.-d'Aubon)</td>
<td>P. D. G.</td>
</tr>
<tr>
<td>BENAUD Manon</td>
<td>« Le Roc », Monseigneur Beaufils</td>
<td>Adm. « Sèvres »</td>
</tr>
</tbody>
</table>

---

Long-term data for Europe

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

http://www.eurhisfirm.eu
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 Common Data Model of WP5
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.

**INPUT IMAGE**

**ENCODING**

**RECOGNITION HYPOTHESES**

- list of characters

**DECODING**

**LANGUAGE MODEL**

- ROUGE
- ROUGE 0.87
- CROIX 0.06
- ROBOT 0.02
- RATIO 0.02
- TOUCH 0.01
- ROUGE 0.01

**TRANSCRIPT:**

ROUGE
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
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

<table>
<thead>
<tr>
<th>Entity name</th>
<th>Parag. text</th>
<th>Type of Named entities</th>
<th>Named entities detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company name</td>
<td>1st parag. text</td>
<td>Company name</td>
<td>L’AIGLE (Compagnie d’assurances contre l’incendie)</td>
</tr>
<tr>
<td>CONSEIL</td>
<td>Text</td>
<td>Name</td>
<td>P. de Séroux</td>
</tr>
<tr>
<td>CAPITAL</td>
<td>Text</td>
<td>Function</td>
<td>P.-D.G.H.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Task 7.3: Extraction primary results

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

<table>
<thead>
<tr>
<th>Level 2 keywords</th>
<th>Level 3 Named Entities</th>
<th>Precision TP/(TP+FP)</th>
<th>Recall TP/(TP+FN)</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conceil (alternatives)</strong></td>
<td>Admin_name</td>
<td>99.36%</td>
<td>98.52%</td>
<td>98.94%</td>
</tr>
<tr>
<td></td>
<td>Admin_function</td>
<td>92.59%</td>
<td>98.68%</td>
<td>95.54%</td>
</tr>
<tr>
<td><strong>Siege social</strong></td>
<td>departement</td>
<td>96.84%</td>
<td>92.93%</td>
<td>94.85%</td>
</tr>
<tr>
<td></td>
<td>town</td>
<td>83.13%</td>
<td>90.79%</td>
<td>86.79%</td>
</tr>
<tr>
<td></td>
<td>address</td>
<td>98.75%</td>
<td>81.44%</td>
<td>89.27%</td>
</tr>
<tr>
<td><strong>Constitution</strong></td>
<td>initial company state</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td>start date</td>
<td>100.00%</td>
<td>99.00%</td>
<td>99.50%</td>
</tr>
<tr>
<td></td>
<td>end date</td>
<td>99.02%</td>
<td>96.19%</td>
<td>97.58%</td>
</tr>
<tr>
<td></td>
<td>company transformation type</td>
<td>100.00%</td>
<td>92.00%</td>
<td>95.83%</td>
</tr>
<tr>
<td></td>
<td>transformation date</td>
<td>52.63%</td>
<td>55.56%</td>
<td>54.05%</td>
</tr>
<tr>
<td><strong>capital</strong></td>
<td>Initial capital</td>
<td>92.31%</td>
<td>92.31%</td>
<td>92.31%</td>
</tr>
<tr>
<td><strong>service financier et transfert</strong></td>
<td>Location</td>
<td>30.00%</td>
<td>60.00%</td>
<td>40.00%</td>
</tr>
<tr>
<td></td>
<td>Institution name</td>
<td>72.47%</td>
<td>97.73%</td>
<td>83.23%</td>
</tr>
<tr>
<td><strong>OVER ALL</strong></td>
<td></td>
<td>85.93%</td>
<td>88.86%</td>
<td>86.76%</td>
</tr>
</tbody>
</table>
Task 7.3: Extraction primary results

- Precise specification → good extraction

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

*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.*
Task 7.3: Extraction primary results

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 lexical features.

<table>
<thead>
<tr>
<th>Level 2 keywords</th>
<th>Level 3 Named Entities</th>
<th>Precision TP/(TP+FP)</th>
<th>Recall TP/(TP+FN)</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concel (alternatives)</td>
<td>Admin_name</td>
<td>99.36%</td>
<td>98.52%</td>
<td>98.94%</td>
</tr>
<tr>
<td></td>
<td>Admin_function</td>
<td>92.59%</td>
<td>98.68%</td>
<td>95.54%</td>
</tr>
<tr>
<td>Siege social</td>
<td>departement</td>
<td>96.84%</td>
<td>92.93%</td>
<td>94.85%</td>
</tr>
<tr>
<td></td>
<td>town</td>
<td>83.13%</td>
<td>90.79%</td>
<td>86.79%</td>
</tr>
<tr>
<td></td>
<td>address</td>
<td>98.75%</td>
<td>81.44%</td>
<td>89.27%</td>
</tr>
<tr>
<td>Constitution</td>
<td>initial company state</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td>start date</td>
<td>100.00%</td>
<td>99.00%</td>
<td>99.50%</td>
</tr>
<tr>
<td></td>
<td>end date</td>
<td>99.02%</td>
<td>96.19%</td>
<td>97.58%</td>
</tr>
<tr>
<td></td>
<td>company transformation type</td>
<td>100.00%</td>
<td>92.00%</td>
<td>95.83%</td>
</tr>
<tr>
<td></td>
<td>transformation date</td>
<td>52.63%</td>
<td>55.56%</td>
<td>54.05%</td>
</tr>
<tr>
<td>capital</td>
<td>Initial capital</td>
<td>92.31%</td>
<td>92.31%</td>
<td>92.31%</td>
</tr>
<tr>
<td>service financier et transfert</td>
<td>Location</td>
<td>30.00%</td>
<td>60.00%</td>
<td>40.00%</td>
</tr>
<tr>
<td></td>
<td>Institution name</td>
<td>72.47%</td>
<td>97.73%</td>
<td>83.23%</td>
</tr>
<tr>
<td>OVER ALL</td>
<td></td>
<td>85.93%</td>
<td>88.86%</td>
<td>86.76%</td>
</tr>
</tbody>
</table>
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
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
Future Work: Task 7.3 Textual Information Extraction

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

  - NOTA

  Les parts bénéficiaires 1ère série sont remboursables à 3.600 fr.; 2e série à 2.850 fr.
  Ces parts ont été converties le 1er janvier 1960 en NF.

  - PARTS BENEFICIAIRES


- 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

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 Common Data Model of WP5
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
  o → 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

<table>
<thead>
<tr>
<th></th>
<th>U. Antwerpen</th>
<th>PSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of records</td>
<td>2370</td>
<td>7682</td>
</tr>
<tr>
<td>Number of supposed matches</td>
<td></td>
<td>144</td>
</tr>
<tr>
<td>Number of linking framework matches</td>
<td></td>
<td>58 supposed matches (~ 40%) + 95 other matches + 183 „reviews“</td>
</tr>
</tbody>
</table>
Results

- Linking to GND, DBpedia and Wikidata started

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>p</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://hebis.de/eurhisfirm/CFS-000098">http://hebis.de/eurhisfirm/CFS-000098</a></td>
<td><a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#type">http://www.w3.org/1999/02/22-rdf-syntax-ns#type</a></td>
<td>schema:Organization</td>
</tr>
<tr>
<td>2</td>
<td><a href="http://hebis.de/eurhisfirm/CFS-000098">http://hebis.de/eurhisfirm/CFS-000098</a></td>
<td><a href="http://hebis.de/eurhisfirm/externalId">http://hebis.de/eurhisfirm/externalId</a></td>
<td>&quot;BAYHANDE&quot;</td>
</tr>
<tr>
<td>3</td>
<td><a href="http://hebis.de/eurhisfirm/CFS-000098">http://hebis.de/eurhisfirm/CFS-000098</a></td>
<td>dc:source</td>
<td>&quot;CFS&quot;</td>
</tr>
<tr>
<td>4</td>
<td><a href="http://hebis.de/eurhisfirm/CFS-000098">http://hebis.de/eurhisfirm/CFS-000098</a></td>
<td>schema:identifier</td>
<td>&quot;CFS-000098&quot;</td>
</tr>
<tr>
<td>5</td>
<td><a href="http://hebis.de/eurhisfirm/CFS-000098">http://hebis.de/eurhisfirm/CFS-000098</a></td>
<td>schema:legalName</td>
<td>&quot;Bayerische Handelsbank&quot;</td>
</tr>
<tr>
<td>6</td>
<td><a href="http://hebis.de/eurhisfirm/CFS-000098">http://hebis.de/eurhisfirm/CFS-000098</a></td>
<td><a href="http://hebis.de/eurhisfirm/sector">http://hebis.de/eurhisfirm/sector</a></td>
<td>&quot;Bank&quot;</td>
</tr>
</tbody>
</table>
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