Data extraction and matching The EurHisFirm experience

Sébastien Adam, Jan Annaert, Frans Buelens, Bertrand B. Coüasnon, Boris Cule, Amaury de Vicq, Camille Guerry, Pierre-Cyrille Hautcoeur, Thierry Paquet, Andres Rojas Camacho, et al.

▶ **To cite this version:**

Sébastien Adam, Jan Annaert, Frans Buelens, Bertrand B. Coüasnon, Boris Cule, et al.. Data extraction and matching The EurHisFirm experience. Methodological Advances in the Extraction and Analysis of Historical Data, Kellogg School of Management - Northwestern University, Dec 2021, Chicago/Virtual, United States. hal-03828381

HAL Id: hal-03828381
https://hal.science/hal-03828381

Submitted on 25 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Data extraction and matching
The EurHisFirm experience

https://eurhisfirm.eu/

Sebastien Adam (Université de Rouen Normandie)

Jan Annaert (University of Antwerp & Antwerp Management School)

Frans Buelens (University of Antwerp)

Bertrand Couasnon (Univ Rennes, CNRS, IRISA Rennes)

Boris Cule (University of Tilburg)

Amaury De Vicq (Paris School of Economics)

Camille Guerry (Univ Rennes, CNRS, IRISA Rennes)

Pierre-Cyrille Hautcoeur (EHESS and Paris School of Economics)

Pantelis Karapanagiotis (European Business School - Wiesbaden)

Iwan Le Floch (Univ Rennes, CNRS, IRISA Rennes)

Aurélie Lemaitre (Univ Rennes, CNRS, IRISA Rennes)

Thierry Paquet (Université de Rouen Normandie)

Johan Poukens (University of Groningen & University of Antwerp)

Angelo Riva (European Business School – Paris / INSEEC U Research Center & Paris School of Economics & Institut Louis Bachelier) – corresponding author

Andres Rojas Camacho (Université de Rouen Normandie)

Abstract

This paper reports results from the design phase of EurHisFirm. Its goal is to integrate isolated and badly accessible financial data sets on 19th and 20th century European companies so that users can query the data as if they reside in one large database. In addition, it wants to stimulate database construction by providing not only methodology and tools to connect to and collaborate with existing ones, but also a collaborative platform, based on machine learning and artificial intelligence, that allows harvesting data in a semi-automatic way. We present the proof-of-concept results of this platform in addition to the performance of matching algorithms, which are necessary to connect and collate the different constituent databases as well as to connect them to contemporary commercial databases.

1. Introduction

In this paper we present results of the EurHisFirm project, which is developed within the framework of the European Union's H2020 Infrastructure Development Program. It brings together economists, historians, IT and information systems scholars and experts in data management from 12 institutions located in 8 European countries. EurHisFirm aims at designing a research infrastructure to collect, connect, collate, enrich, and share detailed, reliable, and standardized long-run European company-level data. As such, it seeks to transform the existing landscape of isolated and inaccessible historical data sets, into a freely-available, integrated, interconnected and interoperable cloud-based network compliant with the FAIR (findable, accessible, interoperable, reusable) data principles (Wilkinson et al., 2016).[1]

EurHisFirm itself does not collect the data on a large scale, as the H2020 Infrastructure Development Program does not finance data collection. Rather it fosters the development of an infrastructure and tools that facilitate the actual data gathering and dissemination. It consists of four distinct phases: (i) design study, (ii) development, (iii) consolidation, (iv) support, of which EurHisFirm has just ended the design study phase. In this phase, it provided proof of concepts, which we summarize in this paper. More specifically, we focus on two interrelated dimensions of the infrastructure: the data extraction platform and matching algorithms. To understand how these fit in the overall project, we first present an overview of its motivation and goals in Section 2. The different components of the intelligent and collaborative platform for the extraction of structured information from images of historical stock exchange price lists are explained in Section 3. We also report on the effectiveness of the platform, based on the tests that were conducted on price lists from Brussels, Paris and Madrid. In Section 4, matching algorithms are discussed and tested. Indeed, as the project envisages pan-European research, it is important that information coming from different exchanges can be linked and matched. It is therefore crucial that entities such as companies and securities, but also persons, can be uniquely identified such that their information that is potentially present in several databases can be retrieved, linked and matched. Again several experiments were set up in which linking and matching algorithms were tested. The tests we present include linking several historical databases to identify cross-listings as well as linking historical price series to present-day series from commercial databases to build long-term price series. We present some concluding remarks in Section 5.

2. EurHisFirm goals

EurHisFirm is rooted on the profound belief that policy measures to address the challenges our society faces would benefit from a deeper understanding of Europe's social and economic long-term trajectories. However, this critical historical understanding of our society remains largely unfulfilled because we lack the requisite empirical basis. One noteworthy shortcoming is the lack of detailed, harmonized, high-quality and long-term data on European companies. Although within academia considerable resources have been devoted to the construction of historical datasets, their aims are typically limited. Moreover, such datasets are scattered and disperse, and they do not satisfy the FAIR data principles. Their access too often depends on the willingness of the person(s) having built them.

[1] For more information on the FAIR data principles, see: https://www.go-fair.org/fair-principles/.

They lack systematic comparative or diachronic analytical purposes. Consequently, due to the absence of permanent infrastructures, harmonization, and universal access, these databases' potential value is largely lost to the public. On the other hand, the few historical series stored in commercial databases are often not entirely suitable for research, despite their widespread use in academia and business. The datasets may have been built using poorly documented sources that were easy-to-find, but not necessarily appropriate or accurate. Unsurprisingly, this makes analysis that relies on such datasets prone to be erroneous.

This scarcity of long-term (micro) data is particularly notable at the European level. In contrast, in the US, acknowledging the enormous research potential of this sort of long-term (micro) data, enormous resources have been devoted to constructing such databases. Notable examples include The Collaborative for Historical Information and Analysis (CHIA), The Wharton Research Data Services (WRDS), and The Center for Research in Security Prices (CRSP). By linking academic and research institutions, CHIA sets out to sustain a Human System Data Resource;[2] whereas WRDS offers its users access to over 250 terabytes of data across multiple disciplines, including accounting, banking, economics, healthcare, insurance and marketing.[3] Last but certainly not least, CRSP is the most widely used financial database, containing prices and dividends for shares listed on the New York Stock Exchange and other trading avenues from 1926.[4] The recent merger between the CRSP and Compustat (an infrastructure that stores accounting and other data on US firms) has further expanded the existing research possibilities. Because of their dominant position in data collection, American companies are frequently - and at least implicitly - deemed representative for their global (including European) counterparts. Lessons are consequently drawn from their experiences that are allegedly - but are often not entirely - applicable elsewhere. This is problematic because it might lead to possibly biased or outright incorrect conclusions, thus undermining research validity and inhibiting policy implications that are applicable outside of the US (EurHisFirm D1.14, 2021).

In sum, there is currently a pressing need for high quality, long-term, empirical European (micro)data. As mentioned before, the lack of this sort of data limits the development of sound models for analyzing structural and cyclical changes. Such models are crucial for understanding the dynamics between financial, economic, and societal changes. Our ongoing capacity to design effective policy measures, in turn, depends on our understanding of both past and current dynamics. While the Strategy Report on Research Infrastructure acknowledges the crucial advances of Big Data and identifies them as promising tools in social sciences and the humanities, the so-called "born-digital" big data still lack the historical depth that "born-on-paper" data can provide.[5] It calls for developing innovative Research Infrastructures which (1) fully exploit Europe's rich historical heritage, (2) rely on new technologies to analyze and make publicly available processed data collections to a wide variety of stakeholders, and (3) can serve as a benchmark for future research projects.

EurHisFirm responds to this call by building a research infrastructure that not only integrates scattered databases, but also offers the necessary methodology and tools to help building new large-scale

[2] For more information on CHIA, see: http://chia.pitt.edu/.
[3] For more information on WRDS, see: https://wrds-www.wharton.upenn.edu/.
[4] For more information on CRSP, see: http://www.crsp.org/.
[5] For more information on European Strategy Forum on Research Infrastructures (ESFRI) and the Strategy Report on Research Infrastructures, see: http://roadmap2018.esfri.eu/.

historical datasets. By doing so, it hopes to stimulate research on the long-term development of European companies over the nineteenth and twentieth centuries from a diversity of perspectives: (i) innovation, (ii) business history, (iii) political economy, and (iv) banking and finance. The data collection to allow for this type of research will primarily consist of organizational (e.g., juridical status, voting and governance rules), financial (e.g., financing, accounting, market data), geographical (e.g., location of headquarters and units of production), socio-technical (e.g., patents) and sociological (e.g., directors, staff careers, shareholders, and partners) information on firms.

Evidently, this sort of historical firm-level data exists in a wide range of formats, digital or otherwise. Countless researchers, both within and outside of the EurHisFirm project, have been collecting their own research material and have been constructing their own databases for a long time. Unfortunately, these databases have their own idiosyncrasies, are sometimes hard to find and even more difficult to access by outsiders. Their maintenance may be hazardous when the original research team moves to other projects, questioning their sustainability. Also constructing new databases involves huge start-up costs. The EurHisFirm research infrastructure attempts to overcome these problems in several ways.

First, building on experiences from earlier research projects conducted on the national level for France (*Données financières historiques,* DFIH) and Belgium (*Studiecentrum voor Onderneming en Beurs,* SCOB),[6] it developed a common data model that provides a set of standards that accommodates data specificities varying both over time and location, while at the same time allowing to compare data across different countries and periods as much as possible (Annaert et al., 2012). Each individual database developer can decide the degree by which her data conform to the standard and use the support and tools provided by the research infrastructure to migrate towards the common standard. This common standard is necessary to interconnect the participating databases. Of course, one of the main challenges to the operational integration of the different existing and future databases EurHisFirm faces is the unique identification at European level of firms, persons and securities. In Section 4, we discuss some of the difficulties that are encountered when information originating from different databases is matched and illustrate the performance of matching algorithms.

Second, existing databases are also bound to constantly change and expand over time as new users amend them or add new information. Therefore, there is a need for an autonomous, self-contained environment, distinctive from these individual databases, which allows the research teams to collaboratively import, edit, and use the combined datapool. Within the EurHisFirm project, Wikibase is currently being used to develop this collaborative environment. Its intended purpose is to facilitate the process of matching all types of entities between various sources, and ultimately visualize and export the results of this collaborative work for the user's own needs. Wikibase itself does not provide the tool to find the matches between entities in separate databases but allows to register and share those matches in such a way that others can build on them. In this way, many matching techniques and processes could be considered, implemented, and used independently by any interested actor. It is important to note that the matching processes can be run outside of Wikibase while their findings

[6] E.g., see the French research project DFIH (https://dfih.fr/) or SCOB (https://www.uantwerpen.be/en/research-groups/scob/) for the Belgian experience.

would be automatically registered and then verified manually, all in a centralized and open database on the web: the EurHisFirm Wikibase platform. In other words, "Wikibase is not the database to rule them all, but the database to link them all" (EurHisFirm M6.1, 2020). As such, Wikibase is a concrete example to demonstrate that EurHisFirm does not merely strive to conduct and encourage empirical research, but also seeks out to promote methodological and epistemological advances in digital humanities and social sciences to develop innovative tools and deepen our understanding of the practical applications of these techniques.

Third, the source material the EurHisFirm project envisages is immense and digitalisation may be considered punitive. However, given that sources such as price lists and yearbooks are very structured, harvesting information is amenable to some automatization. EurHisFirm provides an intelligent and collaborative extraction platform. Importantly, software is developed and trained to recognize the content of digitalized images of the original sources like price lists and yearbooks as information about securities, companies, or other entities and to store it as such in a structured database. As such we hope to industrialize archival work thus unlocking the multitude of paper sources that are available to document the rich European financial history of the last two centuries for systematic research. We provide a more detailed overview of the platform and its underlying processes in Section 3.

Fourth, conceived to meet the need for a benchmark Research Infrastructure as outlined by the H2020's Report on Research Infrastructure, to bring together the isolated and often inaccessible historical datasets into an interconnected and interoperable cloud-based network, EurHisFirm will develop two types of data access: 'on-site' (researchers visit and cooperate with the dataset producers on a project while learning about the data and tools) and remote access (researchers have remote access to the data and the platforms). The plan is to grant access according to accreditations delivered based on standard rules elaborated in compliance with the European Charter for Access to Research Infrastructures. In the first stage, access to data is on-site, but over time EurHisFirm processes will allow for full remote and unrestricted access to data, in line with the progress of the common platforms, and in accordance with the FAIR data principles.

Finally, we also want the research infrastructure to be sustainable, in the sense that the data will remain available to users everywhere in the most user-friendly way as possible. As such EurHisFirm will benefit from the unique experiences of its various members, including the Consortium of European Social Science Data Archives (CESSDA ERIC) and GESIS.[7] The former's in-depth knowledge in European infrastructures will ensure EurHisFirm to exploit synergies, to reflect on sustainability, and to enable complementarity and coherence within the Social Sciences and Humanities Open Cloud (SSHOC) and the European Open Science Cloud (EOSC).[8] In addition, EurHisFirm member GESIS's vast knowledge in data quality, documentation, and harmonization will further guide EurHisFirm to realize these complementarities with existing European Research Infrastructures. The latter is crucial because

[7] For more information on CESSDA and GESIS, see respectively: https://www.cessda.eu/ and https://www.gesis.org/home.

[8] For more information on SSHOC and EOSC, see respectively: https://cordis.europa.eu/project/id/823782 and https://eosc-portal.eu/.

the commitment to integrate into the broader European Research Infrastructure ecosystem is central to the EurHisFirm design.[9]

3. Platform for the extraction of structured information

An intelligent and collaborative platform for the extraction of structured information from images of historical stock exchange price lists is developed. It embeds a system to connect the data extracted to the corresponding security. The platform performs two main tasks: the recognition of the document structure using a cross validation module, and the definition of a general-purpose text recognizer (OCR) to read characters. In the context of the ANR project HBDEX (Exploitation of Big Historical Data for the Digital Humanities: application to financial data), it has been tested at full scale on price lists from the Paris unofficial market: "la Coulisse", harvesting data from 1871 to 1961, amounting to some 235,000 pages and more than 30 million lines. However, it is built in a generic way to be able to analyze and extract data from price lists of different origins. In the EurHisFirm project, we demonstrate its generic capacities by applying it to a test set of official price lists from Brussels, Madrid and the Paris Parquet.

We first present in Section 3.1. the strategy and the transversal analysis of a collection of documents, which drives the complete process of data extraction. It combines the global meta table structure to understand the tables of tables organization on price list pages (Section 3.2), the generic table content description for table understanding (Section 3.3) and the general-purpose text recognizer (Section 3.4). Each section also presents results and the evaluation of each process. In Section 3.5, a global evaluation is given for the complete Information Extraction System on price lists. Links to examples of the XML results of the application of the Information Extraction System on 1899 price lists for "La Coulisse" are given in the Internet appendix.

3.1. Strategy and transversal analysis

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 1) which allows a cross validation of various information in the document collection through a transversal analysis of documents. 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, by using redundancies found in the sequence of daily quotations. At each iteration, when needed the system generates questions for expert users in an asynchronous way.

Insert Fig 1 about here

Each iteration consists of five steps:

1. a first structural analysis with a grammatical description to produce a hypothesis,
2. the transcription of the text-lines localized in step 1,
3. a sequential analysis for the validation of the elements extracted from the image,

[9] The SSHOC project is coordinated by CESSDA ERIC; three other institutional members of the EurHisFirm consortium (EEP-PSE, UA, SAFE) are already members of SSHOC as representatives for EurHisFirm.

4. an eventual call to a user interface;
5. a new structural analysis that integrates the knowledge obtained in steps 2, 3 and 4.

Insert Fig 2 about here

An example of the global strategy for price lists extraction is presented in Figure 2: it is made of a sequence of four iterations to cross validate columns, then sections, then stock names, to finish with the remaining stock fields. With the user interface we validate certain elements that could not be validated automatically because the information is missing from the images. For example, when a new stock appears on the market, we need to know whether it is really a new stock or an existing stock whose name has changed. If it is a new stock, we need an expert to define the unique ID with which it should be associated. Figure 3 shows another example: the string "m." has been correctly recognized as Mark, 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 together with all the original information needed to provide answers: the current stock quotation (in the middle), the previous day's quotation for the same stock (above) and the next day's quotation (below). This global strategy with cross validation, allows to improve the quality of the data extraction, while reducing drastically the number of questions for experts, thanks to the collection redundancies. The global strategy is built on the formalization of user interactions proposed in (Chazalon et al., 2011). An example of this Expert User Interface is presented in the video Demo 1 in the internet appendix.

Insert Fig 3 about here

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, is able to extract data in price lists when tables are vertically aligned. But the tables in the EurHisFirm test dataset 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 4). We therefore developed the recursive table structure analysis based on previous work done in (Coüasnon, 2006) with the DMOS-PI method and its b-dimensional grammatical formalism EPF (Enhanced Position Formalism). With EPF we describe a global meta table structure, whatever the number of rows and the number of columns it is made of, and try to detect recursively in each detected cell, another recursive table structure. From this description the recognition system of the DMOS-PI method can detect the global meta table structure in any price list table.

Insert Fig 4 about here

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. Even when documents are degraded or if a line

segment representing a line border is damaged is the system able to correctly detect the global table structure. Four possible types of structures 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 to extract thin lines, thick lines, multiple lines (Lemaitre et al., 2009). 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. Subsequently, the price list data extraction system is applied separately on each cell.

3.2.2. Evaluation

To evaluate our system, we used the ZoneMap metric (Galibert et al., 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.
We built three different corpora, one for each type of price list we work on: Brussels, Paris Parquet and Madrid. Each corpus contained 30 pages and we obtained a global score of 0.52 for Brussels, 0.72 for Parquet and 1.37 for Madrid (see Table 1).

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

Table 1: ZoneMap score on the Brussels, Parquet and Madrid price lists (lower is better)

Madrid's score is slightly higher because its price list's pages are more degraded than for the other two, and its structure is somewhat harder to detect. Its score is still very low, however, demonstrating that the recognition is very good. The score shown in the table is the mean of the score per page. To get a better idea of the meaning of the score, see the example in Figure 5.

Insert Fig 5 about here

Here we obtain a score of 0.88, because we correctly extracted each of the five tables composing the structure. The score is not zero 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 were not detected. A score containing no issues is generally near one or even zero.

3.3. Generic description of table content

3.3.1. System

The first step of the generic table content extraction system is a table structural analysis of pages. This structural analysis is done with a combination of deep-learning and syntactic approaches. In order to localize text lines within the page, we use an existing system based on deep learning, called Aru-Net (Grüning et al., 2019). Aru-Net is a fully convolutional network which follows a U-net architecture with residual blocks. Aru-Net produces images in which each pixel has a probability of belonging to a text-line (Figure 6 (b)). Text-lines are then extracted from the probability maps produced by the network thanks to simple filtering operations (i.e., gaussian filter and hysteresis thresholding).

Insert Fig 6 about here

The localized text-lines and vertical rulings (extracted with a Kalman filter) are used as terminals, stored in perceptive layers of the bi-dimensional description of price list tables made in EPF from the DMOS-PI method (Coüasnon, 2006; Lemaitre et al., 2009). As the physical organization of price lists varies across exchanges and over time, it is essential the price list table description be sufficiently generic to be able to describe all these variations in a simple way. We therefore developed a generic architecture of the data extraction system.

To define a generic data extraction system, the core recognition of the structure is the same from one price list table to another and we just have to specify the characteristics of each document before starting the generic analysis of the document structure. Therefore, the system always analyzes a document in the same way, with a specification of the characteristics that slightly modify 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(commentTitleMultipleLinesStaggered)(
 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(commentTitleMultipleLinesStaggered)(
page P)))))))))))))))))))).
```

In this way, each type of document is characterized by a list of attributes and rather than by their origin. These attributes could be used again for future price list documents, to produce in an easy way an adapted version combining in a different way the different characteristics (e.g., a stock quotation running over multiple lines).

For the Brussels, Paris Parquet and Madrid price lists major structure adaptations were needed. All these adaptations describe new characteristics, which are added to a library, ready to be reused for other price lists. To illustrate the idea, we discuss three characteristics that required adaptations to the structure.

The first is about the disposition of the section titles. In our corpora two different structures regarding sections were detected (Figure 7 and 8).

Insert Figs 7 and 8 about here

When the section title is outside of a column, we have to look for new sections across the entire width of the table. In this case we also know that no column separators should be next to a section title. We developed this variation and, with the generic aspect of our system, we only have to specify for each corpus whether their sections titles are outside or inside columns.

A second example is about stock quotations running across multiple lines. Initially, we defined a price list description where a table is analyzed line by line and where it was impossible for a stock quotation to run over multiple lines. But, in all three EurHisFirm corpora this possibility does occur. We therefore described two other line types: *staggered lines* (Figure 9) and *stock on multiple lines* (Figure 10).

Insert Figs 9 and 10 about here

The system needs to detect when a title contains data on multiple lines, in order to avoid splitting one stock into two. In the case of staggered lines, their particular structure allows them to be detected: 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. Stocks 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.
Multiple lines structure description also needs to take into account the 'idem' symbol (Figure 11). In some price lists like the Paris unofficial list "La Coulisse", they can contain dash symbols under the part that had to be repeated. Here in the example of Figure 11, 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 the data linked to it.

Insert Fig 11 about here

The third example deals with 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 12).

Insert Fig 12 about here

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 if they should contain useful information.

3.3.2. Evaluation Columns & Headers recognition and validation

We test the importance 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” and select the first page of each day of quotation from 1899 to 1915. This subset contains 4,055 images.
The results produced by step (1) alone (results before validation) results in 320 errors (7.89%), which can be reduced by the full strategy (results after validation) to only 18 errors (0.44%).
Figure 13 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. In a few cases, when the degradation of the document is too strong, the transversal analysis cannot correct errors, and has an impact on column width (see Figure 14). This configuration explains the remaining 18 errors on a total of 4,055 pages.

Insert Figs 13 and 14 about here

3.4. General-purpose text recognizer (OCR)

3.4.1. System overview

High performance character recognition can be achieved when large amounts of training data are available. The strategy developed 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. Figure 15 shows the recognition pipeline that was implemented. It is composed of a deep neural network made of Convolutional Layers (CNN) that process the input image of text lines provided by the structural analysis module of tables, that was presented above. It is 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 first mode is called a raw OCR which outputs the sequence of characters with the highest probabilities.

Insert Fig 15 about here

The second recognition mode is column specific. The system discards any sequence of characters that is not allowed by the grammar that expresses the content of the column. This recognition mode is triggered by the structural analysis of the table that detects each specific column. The Grammar OCR output (see figure 16) is the sequence of characters with the highest probability allowed by the grammar. This second recognition mode requires 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 kind of information that occurs similarly every day over a very long period of time, whereas in some other cases the values may be different every day. They represent prices,

dates, or any numerical values that are encoded in a specific format. A grammar describes the encoding used in the price lists for any such column.

Insert Fig 16 about here

3.4.2. Datasets and Training methodology

The cornerstone in the design of a deep neural network is to get sufficient labelled training data to put the performance to its limit. The training data should be representative of the targeted corpus. But we perfectly know that each corpus may present some specific fonts and styles that a generic recognizer may not have encountered before during its design phase. This is why we adopted a two-stage training strategy of our OCR. A first stage consists in training a general-purpose OCR with as much training data as possible with as little labelling effort as possible. The OCR was first trained on pages from the *French Pricelist La Coulisse*. The advantage of this first data set is to make our OCR perform very accurately on the different types of paper, fonts, and characters of this corpus. This strategy will reward us with a generic OCR which will not be specialized and thus ready for adaptation to other corpora. If we started from scratch for every corpus, we would have to undergo the same procedure of annotating a big volume of data for the OCR to start behaving correctly, and then only in that corpus. Having a generic OCR from the beginning allows us to easily shift to any corpus with the least possible amount of data. This is the second stage of the training process, which consists of training a specific OCR using a few annotated pages from the new corpus.
A first labelled dataset was built using a commercial OCR to automatically label a dataset composed of 75,000 images extracted from pages of the *French Pricelist La Coulisse*. Of course, at this stage the produced labels were prone to errors, as a commercial OCR is optimized to perform better on modern printed documents than on old ones. A second dataset made of 5,000 line images from *La Coulisse* 1899 and 70,000 line images from *La Coulisse* 1924 was manually annotated so as to provide a sufficient amount of verified training and test data. All in all, we were able to build a specialized training dataset of 46,000 images from *La Coulisse 1924* with verified annotations, and a generic training dataset composed 75,000 + 12,000 images by mixing the data that were automatically labelled by the commercial OCR with data from *La Coulisse 1899 and 1924 that were manually annotated.*
An example of the interface for OCR annotation for building training data is presented in the video Demo 2 in the internet appendix.

3.4.3. Data augmentation

Data augmentation is now a well-known technique to train a Deep Neural Network to its limits without the need to increase the size of the training dataset. This technique is based on modifying the training images randomly during training. 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 table 2 we present the different treatments that are included in the augmentation process with an example and an explanation on how it helps.

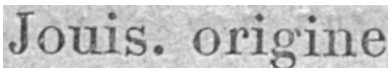
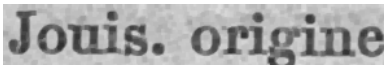
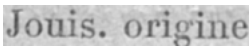
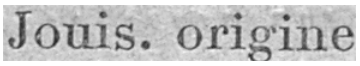
Original image	
Erosion/Dilatation	
Lightening / Contrast modification	
Change of Resolution (DPI change)	
Elastic Distortion	
Gaussian noise	
Bounding Box modification	
Image sharpness modification	

Table 2: Examples of every data augmentation technique

3.4.4. Performance evaluation

The performance of the system is presented by computing the Character Error Rate (CER). For every experiment, the test data set contains hand labelled data that were excluded from the training processes to keep the data as new as possible for the OCR and to have a fair evaluation protocol.

In the following tables we present the performance of the system for *different columns* so that we can characterize the impact of the grammar in percent compared to the performance of the raw OCR. As we can see in table 3, the specialized OCR performs very well on the *La Coulisse 1924* corpus, with a CER that is below 1% except for the *DESIGNATION DES VALEURS* column which is the more complex and gets a CER of 2.61%. Here we can see the grammar reduces the CER to 0.24%, which is more than acceptable. One should not forget, however, that this performance was achievable only because we were able to provide the system with the specific grammar of each column, which is the list or the regular expression that expresses the expected entries in table 3.

#	Column Name	CER raw OCR	CER OCR with grammar	Improve-ment
1	MONTANT ou NOMBRE DE TITRES ÉMIS	0.07 %	0.04 %	35.40 %
2	MONTANT ou NOMBRE DE TITRES ADMIS	0.17 %	0.14 %	21.82 %
3	VALEUR nominale	0.02 %	0.00 %	77.08 %
4	COUPONS DATE du dernier payé	0.08 %	0.02 %	72.77 %
5	COUPONS No	0.21 %	0.05 %	74.36 %
6	COUPONS MONTANT NET	0.09 %	0.03 %	68.19 %
7	COUPONS MONTANT BRUT	0.09 %	0.05 %	51.84 %
8	Exercice précédent Revenu brut	0.76 %	0.14 %	81.80 %
9	DÉSIGNATION DES VALEURS	2.61 %	0.24 %	90.90 %
10	COURS DU JOUR Premier cours	0.04 %	0.01 %	86.43 %
11	COURS DU JOUR Plus bas	0.02 %	0.00 %	89.11 %
12	COURS DU JOUR Plus haut	0.01 %	0.00 %	91.67 %
13	COURS DU JOUR Dernier cours	0.21 %	0.00 %	97.97 %
14	CLÔTURE PRÉCÉDENTE (A)	0.52 %	0.09 %	83.39 %
15	RELEVÉ des cours extrêmes depuis le 1er Janv. 1923	0.50 %	0.09 %	81.14 %

Table 3: OCR evaluation on La Coulisse 1924 on 49,858 images

Next, we report the average performance of the OCR *at the corpus level* for the different modes of the recognizer and for different corpora.

Table 4 reports the evolution of the average performance of the OCR regarding the training conditions. We can see the impact of having sufficient high quality annotated data (hand labeled). But we can still see the positive impact of using data augmentation, which reduces the CER by 23%. Finally, the positive impact of using grammars is demonstrated by reaching an average CER of 0.55%, a decrease of 53% of the raw CER.

OCR	**Character Error Rate**
Commercial OCR	26.64%
LITIS OCR V0 (hand labeled + Commercial OCR)	12.83%
LITIS OCR V1 (hand labeled, 40 pages)	1.52%
LITIS OCR V2 (hand labelled, 40 pages, Data augmentation)	1.17%
LITIS OCR V1 + Grammar	0.55%

Table 4: Comparing the average performance on La Coulisse for different training conditions, and different recognition modes.

We conducted further experiments on some other corpora of the EurHisFirm benchmark and we report in table 5 the average performance of the OCR. It also 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 was available, or by mixing the training datasets when insufficient data was available. The best character error rate (CER) is reported for each corpus of our benchmark in addition to the performance obtained on the La Coulisse corpus. In the results reported we use the raw OCR outputs without any grammar, and we compute an average performance whatever the type of field considered (stock names, prices etc…).
We can see that a similar level of performance is obtained on each corpus, with the highest CER on the Madrid price list which is the least annotated corpus. It is also a more difficult corpus due to the low printing quality of the original source.

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

Table 5: Number of labeled data per corpus and average CER

We find similar results for the other EurHisFirm corpora (see Tables A1 to A4 in the internet appendix).

3.5. Evaluation and Application of the Price List Information Extraction System

We evaluate the complete price list 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 list 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 the 1899 Paris “La Coulisse” price lists in the context of the French ANR HBDEX project, where both the complete extraction of the data combined with expert user interaction and the insertion of all the extracted data in the DFIH database has been done.

3.5.1. Evaluation of Stock identification on Paris “La Coulisse” 1899

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 select 4 trading days, 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 6: Evaluation on the Paris La Coulisse 1899 Corpus using the collection context and user interaction

The experimentation (results presented in table 6) 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. The F-measure is the harmonic mean of the precision and recall, where precision is the proportion of identified stocks that are relevant, and recall is the proportion of the stocks that are correctly identified. The F-measure gives an indication of the quality of the stock identification, where a score closer to 1 is better.

3.5.2. Data Extraction on 6 months of the 1899 Paris “La Coulisse” price lists

This quality of data extraction (F-measure of 0.988) is achieved while drastically reducing the number of questions to expert users from 4,061 to 309 thanks to the collection context modeling. Without this collection context, it would have been necessary for the experts to answer 4,061 questions to have the same level of quality. These results were obtained on 536 pages and 54,603 stock lines from 6 months of La Coulisse 1899.

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 table 6) after 309 expert user interactions. 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 user interaction. An example of data extracted in XML on Paris “La Coulisse” 1899, is presented in the video Demo 3 (internet appendix).

	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 obtain an F-measure ≥ 0.988	4,061	309

Table 6: Paris La Coulisse 1899: Interest of the collection context for drastically reduce expert user interactions

We have applied the Price List Information Extraction process on the price lists for Brussels (1912), Madrid (1931), and Paris "Le Parquet" (1961-1962). To save space, the results of the recognition processes are illustrated in Internet Appendix Figures A1 to A6. In total, 2,507 pages were processed and information on 308,024 stocks was extracted.

4. Matching algorithms

Besides ensuring that data extracted within the platform are associated with the corresponding securities, the EurHisFirm research infrastructure should be able to connect both to previously extracted data and to external data. Several data matching or record linkage experiments were conducted as part of the EurHisFirm design study to identify the most suitable methodology.

The data matching process can be separated into two steps: schema matching and record matching. The goal of the schema matching procedures is to identify which tables in various databases contain similar information, and then to identify which columns in those tables can be matched. The goal of record matching is to ascertain which actual data items in various databases represent the same real-world entities. Our focus was primarily on record matching, as schema matching proved to be relatively straightforward and could be performed manually, given the readily available expert knowledge about the available databases. Record matching is much more complex and requires automated procedures. This stems of course from the fact that the number of tables and columns in a typical database will be in the order of magnitude of tens and hundreds, while the number of records will be in the thousands. Moreover, when two datasets originate in different countries and, for example, contain data from different stock exchanges, most records in each dataset will not have a match in the other dataset. We therefore rely on algorithms to find these proverbial needles in the haystack.

The first experiment focused on identifying corporate securities which were cross listed on the Paris Stock Exchange and the Brussels Stock Exchange between 1890 and 1906. The data on the listed securities and their issuers comes from the existing DFIH and SCOB (Annaert et al., 2012 ; Ducros et al., 2017). Because of the hierarchical nature of the data, whereby each security is issued by one issuer, we can first match corporations with securities on both exchanges and then match the stocks and bonds issued by those corporations. This two-step matching process significantly reduces computation

time. Corporation matching was performed on the basis of the string edit distance between the names of the corporations stored in both databases. Edit distance metrics quantify the (dis)similarity between two strings by counting the minimum number of operations required to transform one string into the other. In our experiment, the performance of two commonly used edit distance metrics, the Jaro-Winkler similarity (Winkler, 1990) and the Levenshtein distance (Levenshtein, 1966), is compared. The main difference between these metrics is the set of allowed operations (respectively transpose and insert, delete or substitute a letter). Also, Jaro-Winkler similarity gives more weight to differences at the start of the strings than to those near the end, while the Levenshtein distance gives equal importance to differences anywhere within the strings. To assess the performance of each metric, we presented potential matches to a human expert for validation. Based on the expert's assessment, we computed the true positive rate. True positives are matches that are proposed by the algorithm and have been verified to be true by the expert (as opposed to the false positives which after verification have proven to be false matches). The true positive rate is the ratio of true positives to positives (true and false). It ranges between 0 and 1, whereby a rate of 0 indicates that all matches are false positives and a rate of 1 indicates that all matches are true positives. The true positive rate is a measure of *precision* which evaluates how well a metric is performing in terms of avoiding false positives (Powers, 2011). This is most appropriate in our case as false negatives (i.e., an unidentified match) will be less harmful than a false positive. Tables 7 and 8 report the true positive rates for different thresholds of Jaro-Winkler similarity and Levenshtein distance.

Similarity score	Pos. match. (cumulative)	True pos. rate (cumulative)	Pos. match. (new)	True pos. rate (new)
1 (exact match)	56	0.98 (55/56)	56	0.98
> 0.95	108	0.83 (90/108)	52	0.67 (35/52)
> 0.90	870	-	762	-

Table 7: Results of corporation name matching from the SCOB and D-FIH databases with the Jaro-Winkler algorithm

First, we looked for exact matches (an exact match equals a Jaro-Winkler similarity of 1 or a Levenshtein distance of 0). Both algorithms yielded 56 positive matches, one of which proved to be false on account of two corporations sharing the name *Société Métallurgique de Couillet*. Due to variations in names, languages, spelling, or even simple typos, exact matching can only reveal the proverbial tip of the iceberg. We therefore experimented with different thresholds of similarity or distance to identify non-exact matches. At a filtering threshold of 0.95, the Jaro-Winkler similarity yielded 108 positive matches (including the 56 exact matches). One-third of the newly found positive matches, however, were false positives. The railways *Compagnie des Chemins de fer de l'Est* and *Compagnie des Chemins de Fer de l'Est Algérien*, for instance, have a similarity score of 0.96, but are not related and the *Compagnie des Tramways de Reims* was falsely matched to the *Compagnie des Tramways de Nantes* and the *Compagnie des Tramways de Rouen*, a similarity score of 0.96 notwithstanding. At threshold 0.9, the number of positives returned by the algorithm had become too large for expert verification. The poor performance of the Jaro-Winkler algorithm, even at high thresholds, stems from the emphasis put on similarity at the beginning of the string. Since corporation names in French, the principal language of our data, typically start with *Compagnie* or *Société*, the Jaro-Winkler similarity increases overall, resulting in many false positives. In English and Dutch where

these words (e.g. *Company* and *Maatschappij*) typically appear at the end of the corporation name string, the Jaro-Winkler might produce more satisfactory results at lower thresholds.

The experiment was continued with the normalized Levenshtein distance. This is a variant of the Levenshtein distance which takes into account the length of the input strings because a distance of 1 in a string of 4 characters is more significant than a distance of 2 in a string of 40 characters. The Levenshtein distance can be normalized with respect to longest or shortest string. We chose to normalize using the longest string, the normalized Levenshtein distances in de example above in this case being 0.25 and 0.05 respectively. From our first run of the algorithm at a distance threshold of 0.05, it immediately became apparent that the Levenshtein was much more effective than Jaro-Winkler at identifying true positive matches. The only false positive was between the *Compagnie des chemins de fer de l'Ouest de l'Espagne* and the *Compagnie des Chemins de fer de l'Est de l'Espagne* (the distance being 0.038). Further iterations of the algorithm with increasingly higher distance thresholds confirmed the effectiveness of the Levenshtein distance. Even at threshold 0.20, the number of new positives was still within the limits for expert verification, although around three-quarters turned out to be false.

Distance score	Pos. match. (cumulative)	True pos. rate (cumulative)	Pos. match. (new)	True pos. rate (new)
0 (exact match)	56	0.98 (55/56)	56	0.98
< 0.05	84	0.98 (82/84)	28	0.96 (27/28)
< 0.10	95	0.94 (89/95)	11	0.64 (7/11)
< 0.15	121	0.86 (104/121)	26	0.58 (15/26)
< 0.20	182	0.65 (118/182)	61	0.23 (14/61)

Table 8: Results of corporation name matching from the SCOB and D-FIH databases with the Levenshtein algorithm

The increasing number of false positives at higher distance thresholds was addressed in a second experiment involving corporation name matching between the SCOB database and the London Share Price Database (LSPD). The latter is a database of securities listed on the London Stock Exchange from 1955 to the present. Its master index file contains the names of all UK listed companies during this period (Staunton, 2019). As in the previous experiment, we systematically increased the distance threshold and presented the positive matches suggested by the algorithm to a human expert for verification. The true positive rate is reported in table 9. It drops sharply at threshold 0.20 and at threshold of 0.25, the number of positives (472) became too large for human evaluation. Closer inspection of the false positives at threshold 0.20 revealed that the algorithm struggles with short names. *NDS Group Plc* and *RHJ International* from the SCOB database, for instance, were matched to any corporation in the LSPD whose names consist of a three-letter combination containing respectively an *N*, *D* or *S* followed by *Group Plc* (31 instances) or an *R*, *H*, or *J* followed by *International* (8 instances). We therefore decided to generate just one possible match per company name, namely the best match (or several best matches in case of ties), and ignore all others. Concretely, for *NDS Group Plc* we would propose *IDS Group Plc* at a threshold of 0.10, and then never propose another match at any other threshold. The inherent risk in this strategy is that we might miss out on some true positives if the correct match is, for whatever reason, not the best match present in the data. On the other hand, by eliminating hundreds (and at lower thresholds thousands) of spurious matches from

the output, we could dig deeper and discover matches that would have otherwise remained undiscovered.

Distance score	Pos. match. (cumulative)	True pos. rate (cumulative)	Pos. match. (new)	True pos. rate (new)
0 (exact match)	11	1.00 (11/11)	11	1.00
< 0.05	12	1.00 (12/12)	1	1.00 (1/1)
< 0.10	15	0.93 (14/15)	3	0.67 (2/3)
< 0.15	26	0.81 (21/26)	11	0.64 (7/11)
< 0.20	102	0.30 (31/102)	76	0.13 (10/76)

Table 9: Results of corporation name matching from the SCOB and LSPD databases with the Levenshtein algorithm

The results of the best match approach are reported in table 10. We again normalized the Levenshtein distance using the longer string and both produced the best match in LSPD for every company in the SCOB database and vice versa. In cases when two best matches were found, and one of them turned out to be correct, we count this as 0.5 true positive and 0.5 false positive. The results show quite clearly both the importance of the direction of the matching and the value of the best-match approach. With the best-match approach we were able to dig deeper into the data by raising the distance threshold, thus discovering matches that would have been left undiscovered using the previous approach, due to the unmanageable quantity of false positives in the output. An example of such a true positive match at a high distance threshold is *African Lakes Corporation* and *African Lakes Corp.* In terms of the direction of the matching process, it seems sensible to generate the best match for every company in the smaller database (SCOB), rather than the larger (LSPD). This choice is not only supported by the results, but is also intuitive, as, regardless of the actual number of discovered matches, the larger database will, per definition, always have a greater number of remaining unmatched entries.

Distance score	SCOB – LSPM	LSPM – SCOB
< 0.05	1.00 (12/12)	100 (12/12)
< 0.10	0.93 (14/15)	0.93 (14/15)
< 0.15	0.81 (21/26)	0.84 (21/25)
< 0.20	0.55 (28.5/52)	0.29 (32/111)
< 0.25	0.34 (39.5/115)	-
< 0.30	0.19 (41/211)	-

Table 10: True positive rate of best match approach to corporation name matching from the SCOB and LSPD databases

We also tried to match cross listed securities on the Paris and Brussels stock exchanges which were issued by the 108 corporations found in the first experiment using the Levenshtein distance, albeit unsuccessfully. Even at a threshold of 0.40, the algorithm could only suggest 11 possible matches. This result was not entirely unexpected, however, as security names in the respective stock exchanges official pricelists are heavily abbreviated. Also, in the French price lists of this period, additional information such as par value is included in the name column whereas in the Belgian price lists, this is listed in a separate column. The following example illustrates these differences between the SCOB and D-FIH database: the shares of the *Compagnie française des Mines et Usines d'Escombrera-Bleyberg*

were respectively listed as *Escombrera-Bleyberg (Comp Franc des Mines et Usines d') (1 a 40.000)* and *Escombrera-Bleyberg (Cie Françse des Mines et Usines d'), act. 350 fr., t. p.* (distance 0.286). We therefore turned to matching securities based on their prices on both exchanges.

Security price matching was done based on relative differences between prices, not absolute prices, because prices could be very different in scale, ranging from only a few *francs* to thousands of *francs* (in the period under study, Belgium and France were part of the Latin Monetary Union and one Belgian *franc* equaled exactly one French *franc*). Prices were matched monthly because the frequency with which prices are recorded in both databases differs. If more than one price per month was available, we took the average of all prices within a month. Concretely, in each month in which we found a price in both databases, we divided the smaller average price with the larger average price to compute the ratio for that month. After doing this for all months, we computed the average of all these monthly ratios as the final similarity score for the given pair of securities. Only pairs of securities for which at least 12 matched months are available were included in the analysis.

Shares and bonds were analyzed separately. The results for shares are reported in table 11. At a similarity threshold of 0.97, we discovered 20 positive matches, 19 of which were verified as true positives by the expert. Lowering the threshold to 0.90 produced nine additional positive matches, six of which were true positives. Further lowering the threshold to 0.80 produced no additional positive matches. At a similarity threshold of 0.90, we also found 28 positive matches for corporate bonds. The true positive rate, however, was lower: 0.57 for bonds as opposed to 0.86 for shares. This poorer result was not entirely unexpected. It is the consequence of the particular challenges proposed by the bond market as many corporations typically issue many very similar bonds within overlapping periods. As such, these bonds have similar terms and, in our context more importantly, are traded at similar prices. For example, if ten bonds of a company are traded at one stock exchange, and ten at another, all with similar prices, our methods would identify 100 potential matches, of which at least 90 percent would be false positives.

"Similarity" score	Pos. match. (cumulative)	True pos. rate (cumulative)	Pos. match. (new)	True pos. rate (new)
0.97	20	0.95 (19/20)	20	0.95 (19/20)
0.90	29	0.86 (25/29)	9	0.67 (6/9)

Table 11: Results of share price matching from the SCOB and DFIH databases

Our final experiment concerned a scenario where newly collected data is added to a database. As opposed to the previous experiment, where we expected to find relatively few matches, this scenario implies that we expect to find matches for all entities from the new dataset in our database. In this case, we collected information about management from the director's names supplement (*Liste alphabétique des administrateurs*) to the 1915 edition of the Belgian *Receuil financier* (a yearbook with information on the issuers of securities listed on the Brussels Stock Exchange) and tried to match it to the SCOB database. The director's names supplement lists the names and domicile of directors (*administrateurs*) and statutory auditors (*commissaires*) in alphabetical order with reference to the board positions they held in one or more corporations. Their function is indicated by a letter (for instance, *P.* for *président*, *A.* for *administrateur* and *C.* for *commissaire*) which is followed by the name of the corporation on the board of which they exercised this function. A typical name record in the director's names supplement of the *Recueil financier* hence looks like this:

Adriaensen, Louis, Anvers. — A. Chdf. Méridionaux d'Espagne. — Crédit National Industriel. — Ghezireh Estates. — Pétroles de Boryslaw. — Westende-Plage. — C. Hauts F. Aumetz-La Paix — Hauts F. de Fontoy.

The full directors name supplement for 1915 dataset includes 1,252 corporation names. The SCOB database also includes the names of these corporations whose securities were listed on the Brussels Stock Exchange, but their names were often spelled differently or abbreviated in the directors supplement. Standard abbreviations such as "*Chdf*." (*chemins de fer*) and "*Hauts F.*" (*hauts fourneaux*) were resolved through automatic substitution before the experiment. We first tried to match both lists of names by using the Levenshtein distance, normalized with the longer string, with the threshold set to 0.1. This produced a match for 280 companies, all of them correct. Already this first experiment showed how different these results were to a setting where most of the data was expected to remain unmatched. However, successfully matching 280 out of 1252 companies was hardly satisfactory. This was mainly due to the *Receuil financier* data sometimes being severely abbreviated. However, when we raised the threshold to 0.2, the results considerably deteriorated. Of the hundreds of proposed matches, the majority was incorrect, many of whom involved the same name being matched to multiple other names.

We then attempted to use the best-match approach described above to discover exactly one match (with possible ties) for each company name in the *Receuil financier*. This, too, did not produce satisfactory results, as most best matches were clearly wrong. The reason for this was again the abbreviated nature of the *Receuil financier* data in combination with normalisation using the longer string, which often resulted in company names being matched with short names that shared some generic terms, rather than the more unique aspects of company names. For example, *Aciéries de Longwy* was matched with *Aciéries de Mons*, while the correct match would have been *Société des Aciéries de Longwy*. In conclusion, the Levenshtein distance is a good tool to identify similar names at low distance thresholds, but struggles when abbreviations are used or entire words omitted.

The intuition behind our next approach can be directly illustrated by the example above. The name *Aciéries de Longwy* is in fact entirely contained within the longer version *Société des Aciéries de Longwy*. Naturally, the condition that one name must be entirely contained within the other is far too strict. It does not allow for abbreviations (other than of the final word) or for spelling errors. We therefore decided to look for the longest common subsequence (LCS) between the two names (Hirschberg, 1977). The longest common subsequence of two strings is defined as a sequence of characters that can be found within both strings, allowing for gaps but preserving the order.

The improvement in the results was dramatic. The best match based on LCS proved correct for over 80 percent of the company names. Furthermore, for some companies for which this approach did not result in the correct match, the actual match had already been discovered using the Levenshtein distance. However, there still remained a considerable number of unmatched companies, for a variety of reasons. In some cases, it was clear what went wrong. For example, some short names were, entirely accidentally, completely contained within some very long names in the SCOB database, resulting in a larger LCS score than with their actual match, which may have differed by one or two characters. In other cases, the usage of diacritical symbols caused a mismatch.

We therefore first cleaned the two datasets by removing punctuation and diacritical marks. We then ran an approximate version of the LCS algorithm, whereby we produced the shortest match that had an LCS score of at least 90 percent of the optimal score. Using this method, in combination with the previous efforts, saw us discover a match for over 90 percent of the data, leaving just 107 companies unmatched. What stood out among the 107 unresolved cases is that they were often very short

company names. For example, *Citas* was matched to *Equitas* using the Levenshtein distance, and to *Crédit Lyonnais* using the LCS method. However, the correct match was *Compagnie industrielle et de transports au Stanley-Pool (Citas)*. Clearly, neither the Levenshtein distance nor LCS are suitable to find such well-hidden matches. We therefore turned to an even simpler method – searching for cases where one name was a substring of the other (i.e., completely contained, with no gaps). This approach yielded another 22 correct matches, bringing the total to 1,167, with 85 companies remaining unmatched.

A further inspection of the SCOB database revealed that for 37 of those 85 there was in fact no match to be found because they were not listed on the Brussels Stock Exchange (the directors of *Brussels Motor Cab*, for instance, were included in the directors name supplement of the *Recueil financier* but the company was listed on the *London Stock Exchange*). Removing these unlisted corporations from the directors name supplement dataset reduced the original sample to 1,215. An analysis of the 48 unmatched companies revealed a variety of different reasons for the failure of the matching algorithms to find the correct match. In some cases, the abbreviations were too severe (e.g., *Automobiles SAVA* and its correct match *Société Anversoise pour la Fabrication de voitures automobiles*), in others the order of the words was different (e.g., *Chemins de fer meridienaux italiens* and *Société Italienne pour les chemins de fer méridionaux*), and some were simply too different to be matched by any algorithm (e.g., *Ciments North* and *North's Portland Cement and Brick Works*).

These cases aside, we managed to correctly match 1,167 out of 1,215 companies, a total of 96 percent, which is highly satisfactory. Nevertheless, it is important to note that we needed a variety of sometimes very different techniques to identify all these matches. The inevitable conclusion is that no technique is sufficient on its own. For nearly exact matches, Levenshtein distance performs well. For strings of considerably different lengths, LCS-based techniques give the best results, yet sometimes produce some glaring omissions, too. Finally, substring-based methods can help find very short strings in much longer strings and thus discover further matches.

By way of conclusion, we will dwell briefly on the interaction between matching algorithms and human experts. While fully automated record matching techniques have been a topic of theoretical research for a long time (Newcombe et al., 1959), in practice, such approaches have been successfully used only in very limited settings (He et al., 2018), (Abramitzky et al., 2020), relying exclusively on exact matches or on known and fixed data structures. In general terms, an automated method will evaluate each potential match using a certain similarity measure and then simply output the matches that have a similarity score higher than the predefined threshold. Clearly, in our setting, where false positives can be very harmful, this approach would be too risky. Nevertheless, even with this in mind, record matching algorithms remain crucial to our efforts, as they can be used to rank the potential matches based on a similarity or distance score and offer them to a human expert for verification. In this way, the human effort is minimized, as the algorithm can relatively quickly discard the majority of pairs of records that clearly cannot be matched. However, we cannot rely on an algorithm to automatically insert matches into the newly-built indices without expert verification.

5. Concluding remarks

Bringing together several standalone data sets and develops intelligent tools to create a Pan-European database covering historical financial data on 19th and 20th century companies, securities and persons. It is an immense and ambitious task. With its design study, the EurHisFirm team believes that it has

shown that it is nevertheless a feasible project. At least two challenges need to be met. The first is that a colossal amount of data needs to be digitized and stored in a structured and well-documented infrastructure. The second is that the data needs to be indexed in such a way that securities and persons can be linked to companies, both within an exchange but also across exchanges. This requires unique identification at the European level of firms and securities, but also of persons linked to the firms' management or shareholders. With EurHisFirm we have set important steps to meeting these challenges. The extraction platform is shown to be capable to "industrialize" data harvesting of structured information, like information stored about companies and securities in yearbooks and price information from price lists. Key features responsible for its success are not only the intelligent way it can cope with these kinds of structured information and that it can learn from its "experiences" to benefit data extraction exercises on new corpora. Equally important is how it collaborates with the experts, who are presented images of the original sources that enable them to solve the questions generated by the system. Moreover, the platform's learning capacity minimizes the number of questions generated and the time that experts need to spend in educating the system. The second main challenge is being able to retrieve the information contained in the scattered databases in a comprehensive way. In the EurHisFirm design study, we propose a common data standard based on an identification system, which was not discussed in detail here. On the one hand, it will enable the progressive unique identification of firms, persons and securities and display the level of confidence reached. On the other, it allows to cope with concepts that vary over time and across space. We have also tested several matching algorithms to find identical firms, securities and persons, not only across different historical data sets, but also in contemporary databases. Also for these applications have we shown that with limited effort from experts the success rate for finding matches is very satisfactory. To a large extent this is again thanks to the exploitation of the structured nature of the information, more specifically the price information.

EurHisFirm hopes to be able to start building the research infrastructure along the lines of the design study and to foster the construction of new databases by helping the teams with methodology, the platform and the development of the necessary user-friendly software to operate and consult the databases.

Bibliography

Journal articles, book chapters, reports, and conference proceedings

Abramitzky, R., Mill, R., Pérez, S., 2019. Linking individuals across historical sources: A fully automated approach. Historical Methods A Journal of Quantitative and Interdisciplinary History 53, 1-18

Annaert, J., Buelens, F., de Ceuster, M.J.K., 2012. New Belgian Stock Market Returns: 1832–1914. Explorations in Economic History 49, 189–204.

Chazalon, J., Coüasnon, B., Lemaitre, A., 2011. Iterative analysis of pages in document collections for efficient user interaction, in: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. pp. 503–507.

Coüasnon, B., 2006. DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way. International Journal on Document Analysis and Recognition 8, 111–122.

Ducros, J., Grandi, E., Hautcoeur, P.-C., Hekimian, R., Riva, R., Ungaro, S., 2017. Le projet DFIH: humanités numériques et histoire financière, in: Cavalié, E., Clavert, F., Legendre O., Martin D. (Eds.), Expérimenter les humanités numériques. Des outils individuels aux projets collectifs, Presses de l'Université de Montréal, Montréal, pp. 127-144

EURHISFIRM, 2021. Final Report (D1.14).

EURHISFIRM, 2020. Data Matching Case Study (M6.1).

Galibert, O., Kahn, J., Oparin, I., 2014. The zonemap metric for page segmentation and area classification in scanned documents, in: 2014 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2594–2598.

Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R., 2019. A two-stage method for text line detection in historical documents. International Journal on Document Analysis and Recognition (IJDAR) 22, 285–302.

He, Z.-L., Tong, T.W., Zhang, Y., He, W., 2018. A database linking Chinese patents to China's census firms. Scientific Data 5, 180042.

Hirschberg, D.S., 1977. Algorithms for the Longest Common Subsequence Problem. Journal of the ACM 24, 664–675.

Lemaitre, A., Coüasnon, B., Camillerapp, J., 2009. Use of Perceptive Vision for Ruling Recognition in Ancient Documents, in: Eighth IAPR International Workshop on Graphics RECognition. pp. 3–12.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10, 707-710.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P., 1959. Automatic Linkage of Vital Records. Science 130, 954–959.

Powers, D.M.W., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. International Journal of Machine Learning Technology 2, 37–63.

Staunton, M., 2019. London Share Price lspm201812 & lspd201812 Reference Manual. London Business School, London.

Wilkinson, M.D., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018.

Winkler, W.E., 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage Cleaning and Analyzing Sets of Files View project.

Websites

Centre for Research in Security Prices. University of Chicago Booth School of Business. URL: http://www.crsp.org/.

Collaborative for Historical Information and Analysis. University of Pittsburgh. URL: http://chia.pitt.edu/.

Data for Financial History. DFIH. URL: https://dfih.fr/.

FAIR Principles. GO FAIR. URL: https://www.go-fair.org/fair-principles/.

Strategy Report on Research Infrastructures Roadmap 2018. ESFRI. URL: http://roadmap2018.esfri.eu/.

Study Center for Companies and Exchanges. University of Antwerp. URL: https://www.uantwerpen.be/en/research-groups/scob/

Wharton Research Data Services. Wharton - University of Pennsylvania. URL: https://wrds-www.wharton.upenn.edu/.

Figures

Figure 1: Overview of the global strategy

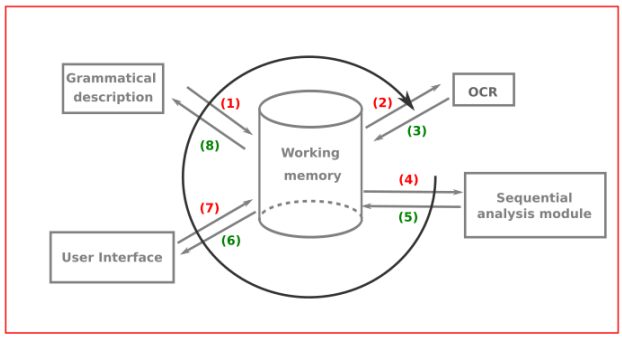
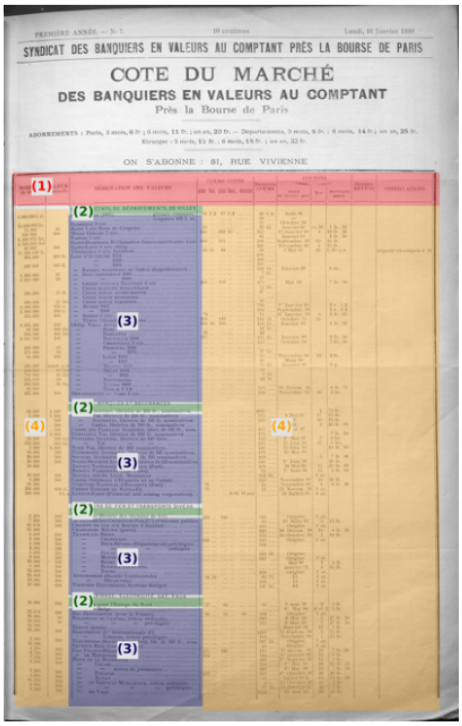


Figure 2: Example of the global strategy for price lists extraction: sequence of 4 iterations



1.Iteration 1: column

1.1. *[Grammatical description]* Extraction of columns without prior knowledge
1.2 *[OCR]* Transcription of column name text-lines
1.3 *[Sequential analysis module]* Rupture detection on columns names
1.4 *[Sequential analysis module]* Rupture detection on columns widths
1.5 *[Grammatical description]* Extraction of reliable columns with clues : columns name and columns widths

2.Iteration 2: section

2.1 *[Grammatical description]* Extraction of columns without prior knowledge
2.2 *[OCR]* Transcription of section name text-lines
2.3 *[Sequential analysis module]* Rupture detection on sections names
2.4 *[Grammatical description]* Extraction of reliable columns with clues : sections names

3.Iteration 3: stocks names

3.1 *[Grammatical description]* Extraction of stocks without prior knowledge
3.2 *[OCR]* Transcription of stocks name text-lines
3.3 *[Sequential analysis module]* Sequence alignment for stock name identification
3.4 *[User Interface]* Question on stock name apparition and errors correction

4.Iteration 4: other fields

4.1 *[Grammatical description]* Extraction of stocks without prior knowledge
4.2 *[OCR]* Transcription of other fields text-lines
4.3 *[Sequential analysis module]* Rupture detection for other fields correction
4.4 *[User Interface]* Question when ambiguities

Figure 3: User interface - case of ambiguous interpretation

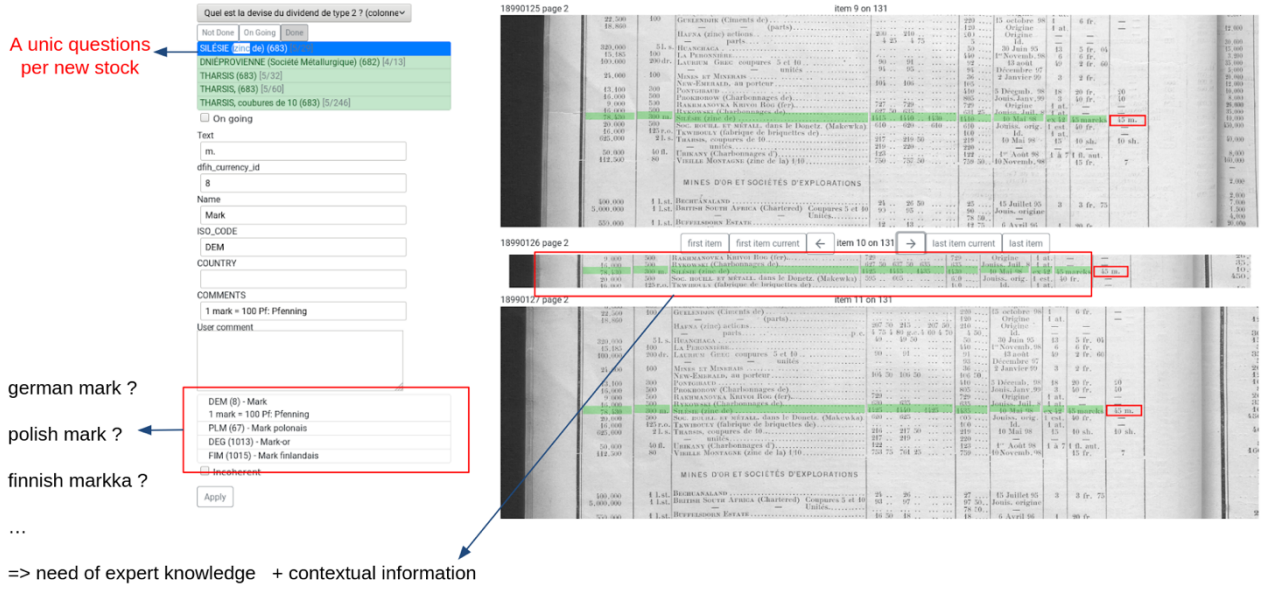


Figure 4: Brussels 1912 Price Lists: Global meta table recognition with reading order recognition

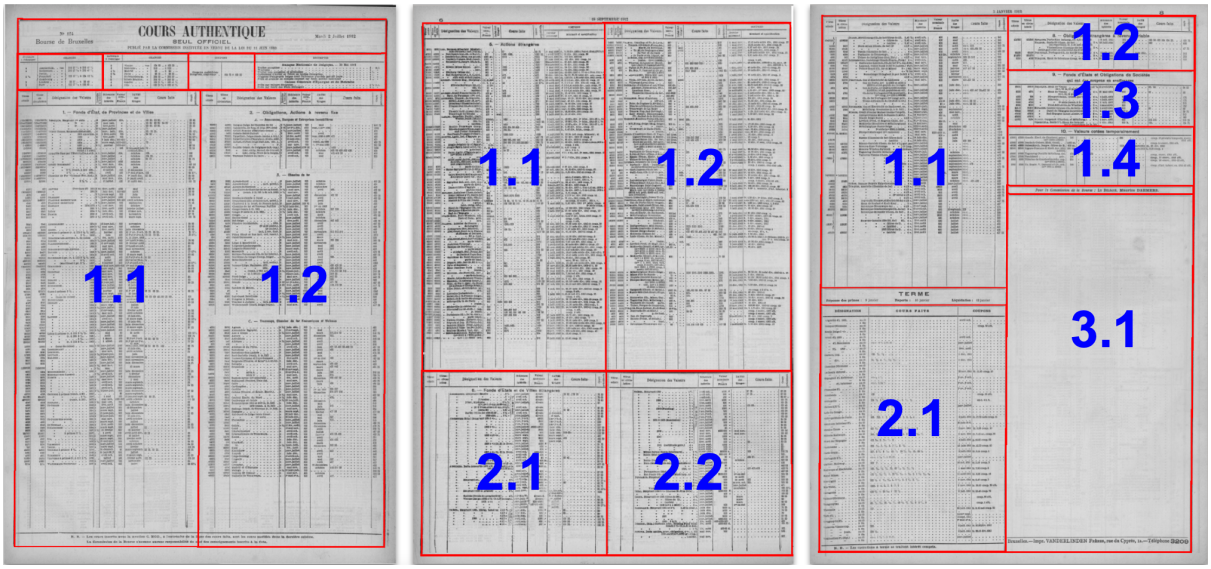


Figure 5: Meta-structure extracted on a page from Parquet price lists

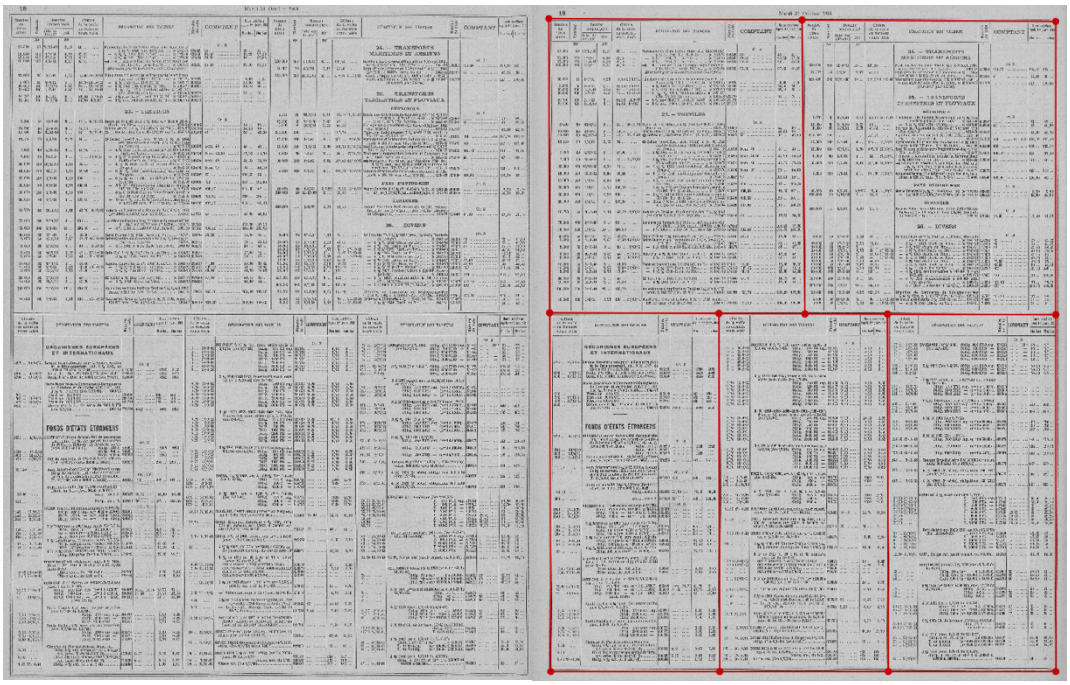


Figure 6: Example of text line extraction in a stock price lists document. (a) Image to be processed - (b) Probability map for text-line in this image - (c) Extracted text-lines

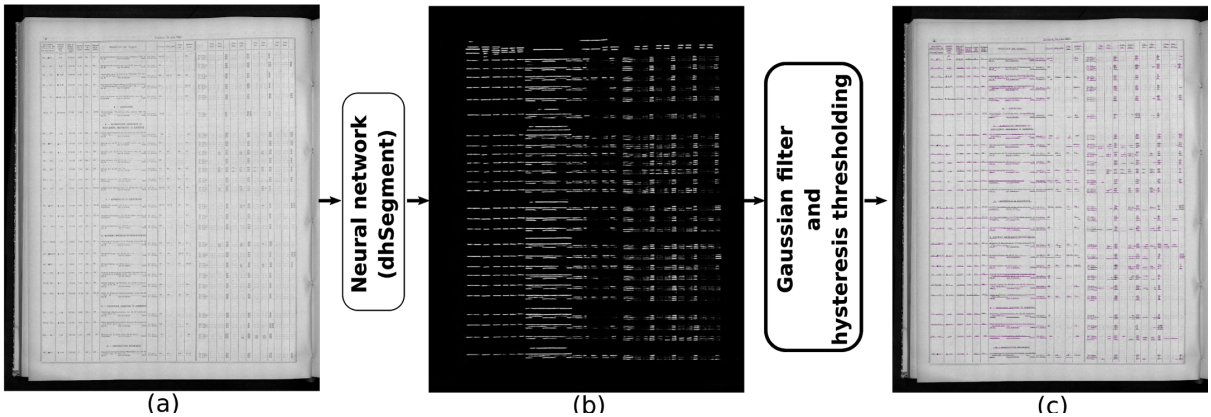
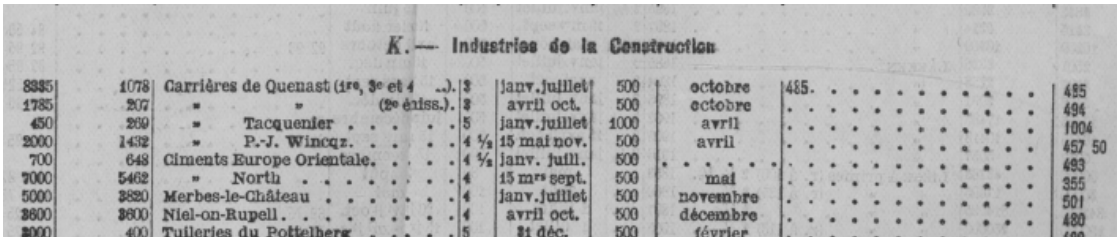


Figure 7: Example of a section title outside of columns in the Brussels price list

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) *livrables*		89234	....	.....	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	37,50	2738 ..	....	Grand Hôtel, act. (ex-c. 10)	×	563579	2710	.....	3298	2155
59.115	100	5/6/61	B 7,35	836 ..	....	Lutétia (Hôtel), act. (ex-c. 14)	×	3777	350	.....	385 ..	284 ..

Figure 8: Example of a section title inside of a column in the Parquet 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 éniss.)	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	3600	Niel-on-Rupell	4	avril oct.	500	décembre	. . . . .	. . . . .	480
2000	400	Tuileries du Pottelberg	5	31 déc.	500	février	. . . . .	. . . . .	400

Figure 9: Example of a staggered line from the Parquet price list

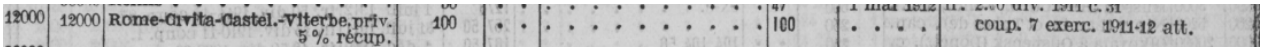


Figure 10: Example of a stock on multiple line from the Brussels price list

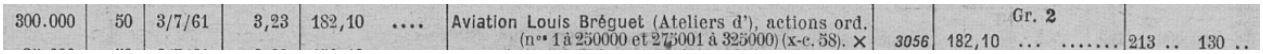


Figure 11: Example of an 'idem' symbol (dash) after a stock on multiple lines from the Parquet price list

Figure 12: Example of a title identified in blue from the Madrid price list

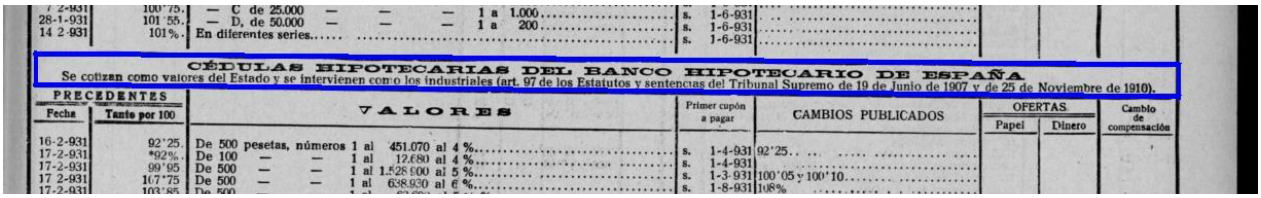
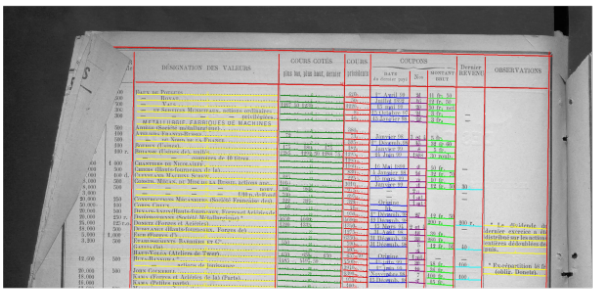


Figure 13: Example of improvement obtained with our strategy, thanks to the context of the collection (left: before validation; right: after validation)



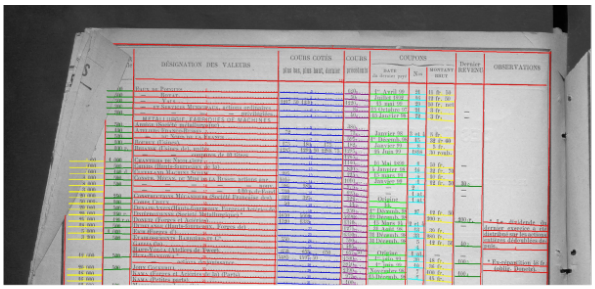
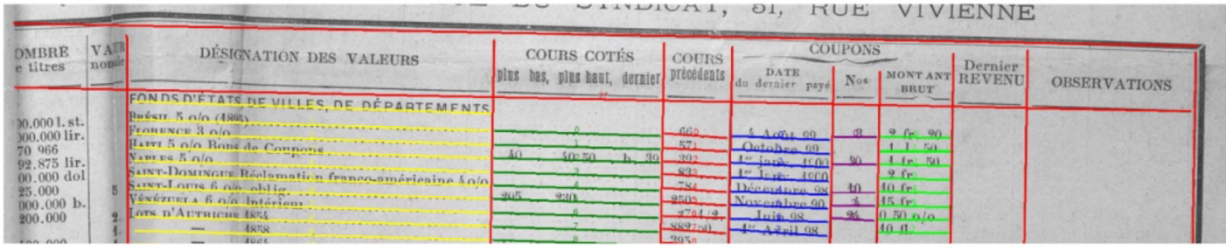
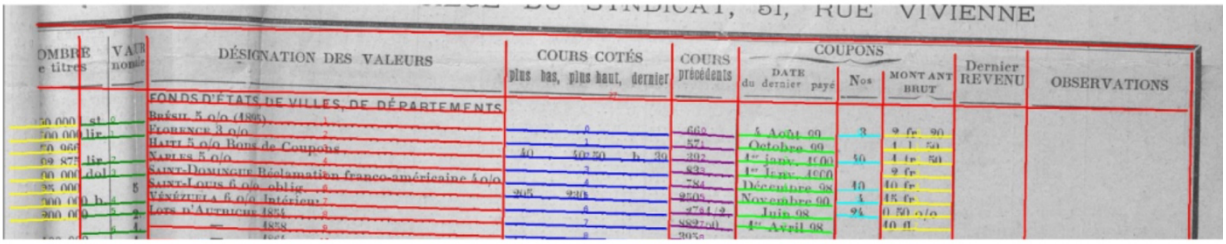


Figure 14: Example of a limit case of our strategy (only 18 errors for 4, 055 pages), when the document degradations are too strong to be corrected, even with the context of the collection



(a) Grammatical description alone (before validation)



(b) Global strategy (after validation)

Figure 15: organization of the recognition pipeline

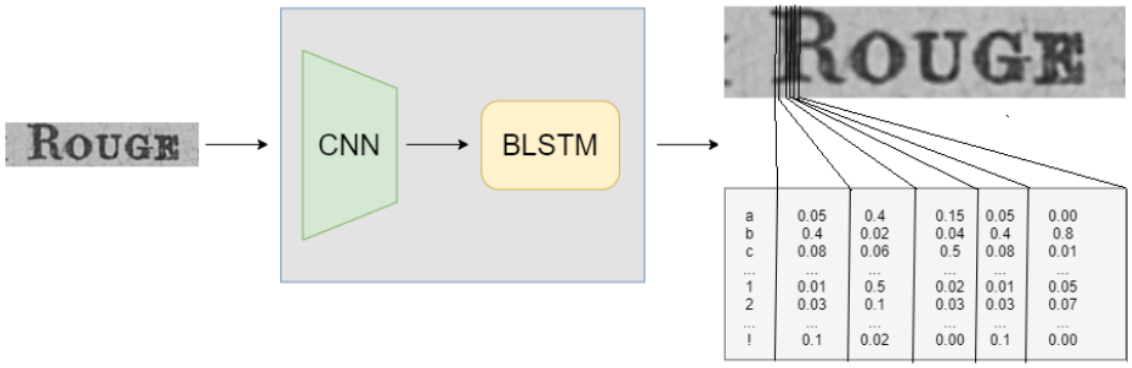


Figure 16: Example of columns with fixed (black) and variable content (red)