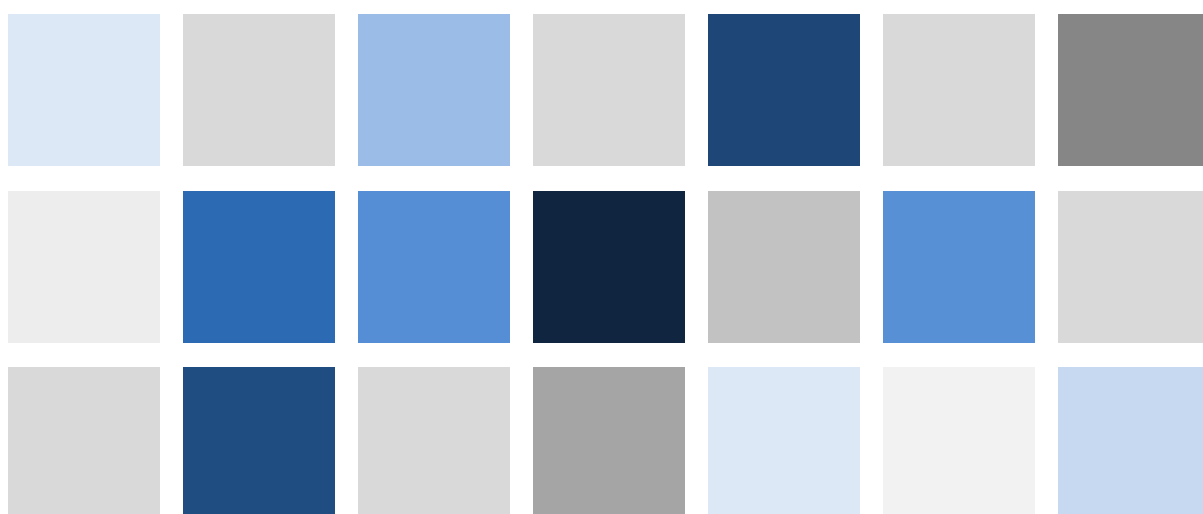


Long-term data for Europe

# EURHISFIRM

D1.10: Second yearly progress and strategy report  
to the General Assembly



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>

**AUTHORS:**

Sébastien ADAM (UNIVERSITÉ DE ROUEN NORMANDIE)  
Robin ADAMS (THE QUEEN'S UNIVERSITY OF BELFAST)  
Jan ANNAERT (UNIVERSITEIT ANTWERPEN)  
Miguel ARTOLA BLANCO (UNIVERSIDAD CARLOS III DE MADRID)  
Stefano BATTILOSSI (UNIVERSIDAD CARLOS III DE MADRID)  
Simon BOUVIER (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)  
Frans BUELENS (UNIVERSITEIT ANTWERPEN)  
Gareth CAMPBELL (THE QUEEN'S UNIVERSITY OF BELFAST)  
Bertrand COÜASNON (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)  
Christopher COYLE (THE QUEEN'S UNIVERSITY OF BELFAST)  
Achille FEDIOUN (UNIVERSITÉ DE ROUEN NORMANDIE)  
Coen FIERST VAN WIJNANDSBERGEN (ERASMUS UNIVERSITEIT ROTTERDAM)  
Nathalie GIRARD (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)  
Renata GWOŹDZIEWICZ-PĘCHERZEWSKA (UNIwersytet Ekonomiczny we Wrocławiu)  
Stefan HOUP (UNIVERSIDAD CARLOS III DE MADRID)  
Krzysztof JAJUGA (UNIwersytet Ekonomiczny we Wrocławiu)  
Abe de JONG (ERASMUS UNIVERSITEIT ROTTERDAM)  
Joost JONKER (KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN - KNAW)  
Pantelis KARAPANAGIOTIS (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)  
Wolfgang KÖNIG (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)  
Katarzyna KUZIĄK (UNIwersytet Ekonomiczny we Wrocławiu)  
Iwan LE FLOCH (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)  
Aurélien LEMAITRE (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)  
Jesús MARTINEZ CASADIEGO (UNIVERSIDAD CARLOS III DE MADRID)  
Thierry PAQUET (UNIVERSITÉ DE ROUEN NORMANDIE)  
Alexander PEUKERT (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)  
Johan POUKENS (UNIVERSITEIT ANTWERPEN)  
Lukas Manuel RANFT (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)  
Yann RICQUEBOURG (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)  
Angelo RIVA (ÉCOLE D'ÉCONOMIE DE PARIS)  
Andres ROJAS CAMACHO (UNIVERSITÉ DE ROUEN NORMANDIE)  
Joanna SŁAWATYNIEC (ERASMUS UNIVERSITEIT ROTTERDAM)  
John TURNER (THE QUEEN'S UNIVERSITY OF BELFAST)  
Lana YOO (ÉCOLE D'ÉCONOMIE DE PARIS)

**APPROVED IN 2020 BY:**

Jan ANNAERT (UNIVERSITEIT ANTWERPEN)  
Wolfgang KÖNIG (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)  
Angelo RIVA (ÉCOLE D'ÉCONOMIE DE PARIS)



## List of terms and acronyms

CDM	Common data model
CESSDA ERIC	Consortium of European Social Science Data Archives
CRSP	The Center for Research in Security Prices
DDI	Data Documentation Initiative
DFIH	Données Financières Historiques
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
EUROFIDAI	Institut Européen des données financières
FAIR	Findability, accessibility, interoperability, and reusability (data principles)
FIBO	Financial Industry Business Ontology
GLEIF	Global LEI Foundation
IPR	Intellectual property rights
LEI	Legal Entity Identifier
RDF	Resource description framework
RI	Research infrastructure
SCOB	Studiecentrum voor Onderneming en Beurs
SSHOC	Social Sciences & Humanities Open Cloud
TOGAF	The Open Group Architecture Framework
WGIS	Work Group on Identification and Standardisation



## Table of Contents

Introduction.....	5
Executive summary .....	10
Work Package 1: Project management .....	15
Work Package 2: Dissemination and communication.....	18
Work Package 3: Legal and ethical issues .....	20
Work Package 4: Data and sources inventory and documentation .....	23
Work Package 5: Common data model .....	27
Work Package 6: Data connecting and matching.....	30
Work Package 7: Data extraction and enrichment system .....	33
Work Package 8: Interaction with users.....	67
Work Package 9: Infrastructure policy and architecture .....	69
Work Package 10: Business model and governance .....	71
Work Package 11: Cultural heritage .....	74
Conclusions.....	76



## Introduction

*EURHISFIRM will design a world-class research infrastructure (RI) to connect, collect, collate, align, and share detailed, reliable, and standardized long-term company-level data for Europe to enable researchers, policymakers and other stakeholders to analyse, develop, and evaluate effective strategies to promote investment and economic growth. To achieve this goal, EURHISFIRM develops innovative tools to spark a “big data revolution” in the historical social sciences and to open access to cultural heritage in close cooperation with existing RIs.*

### A. Background and rationale

#### A.1 The need for scientific evidence

With **economic growth** still slow in some parts of Europe, the key societal challenges facing the European Union are investment, growth, and job creation. Unstable capital markets had undermined corporate investments and had led to increased unemployment and social inequality, harming citizens' well-being and sowing mistrust of public decision-makers and academic experts. To address these challenges, the European Commission has been promoting policy initiatives (such as EU capital markets and a Banking Union) to improve business access to capital, ensure financial stability, and boost investment and innovation. The European Union's Horizon 2020 Programme addresses inclusive long-term growth and social inequality to foster a social and economic framework that promotes **sustainability** in Europe. In order to promote strong, sustainable growth and to meet these urgent social and economic challenges, **the European Union needs sound scientific evidence.**

**Big data are promising tools in science today.** However, in spite of the crucial advantages offered by “born-digital” big data, they still lack the historical depth that “born-on-paper” long-term data can provide. Scientific research, government policy, and society as a whole must explore the historical data necessary to understand the dynamics of the past and how these structure the present and the future. As Mark Twain once remarked, “**History** is a boundless laboratory for real-size natural experiments: history does not repeat itself but it does rhyme”. Yet, because we lack these empirical foundations, this crucial historical understanding of our society remains unfulfilled.

IT research must therefore develop innovative models and technologies that push forward the technological frontier and spark a **big data revolution in historical social sciences**: the scaling up of the variety, quantity, and quality of available long-term data. Digitalized historical sources as part of the **European cultural heritage** represent a shared wealth in terms of citizenship, cultural growth, and economic potential.

#### A.2 The European empirical shortage

**Europe's huge research potential in the social sciences has not been entirely realised due to a lack of empirical works.** The scarcity of long-term data is particularly notable at the European level.

So far, only a very few large stand-alone European long-term databases have been built by both the **academic community** (e.g. the London Share Prices Database of the London Business School) and **private**



**companies** (e.g. Datastream (by Thomson Reuters)). Interoperability, if any, remains low among these databases.

Within **academia**, considerable resources have been devoted to construct historical datasets, often with limited aims, to study specific issues. Moreover, such datasets are scattered and dispersed and do not satisfy the FAIR data principles (Findable, Accessible, Interoperable and Re-usable): they do not permit systematic comparisons or analyses of changes over time. Moreover, access can be limited at the owners' discretion. Consequently, due to the lack of permanent infrastructures, harmonization, and universal access, these data's potential value is lost to the public.

On the other hand, the very few historical series in some **commercial databases**—despite the fact that they are used daily in business and academia—are sometimes unsuitable for research. They can lead to serious errors due to poor documentation; additionally, the foundation may have been built upon easy-to-find but inappropriate sources.

**The USA has been investing enormous resources** to build and link long-term databases suitable for research. The Collaborative for Historical Information and Analysis (CHIA) links academic and research institutions to sustain a Human System Data Resource. The Wharton Research Data Services (WRDS) provides the user with one location to access over 250 terabytes of data across multiple disciplines including accounting, banking, economics, healthcare, insurance and marketing. The Center for Research in Security Prices (CRSP), the most widely used financial database, contains prices and dividends for shares listed on the New York Stock Exchange from 1926. The recent merge between the CRSP and Compustat have expanded the research possibilities.

Because of the USA's dominant position in data production, **American companies** are frequently and implicitly deemed **"representative" or "the norm"**. Lessons are consequently drawn from their behaviour that are supposedly—but are not—applicable everywhere (including Europe), generating many biases and possibly incorrect conclusions.

To summarise, **the current lack of high quality long-term empirical European data prevents the usage and testing of models for analysing structural and cyclical changes, which are crucial for understanding the interactions between financial, economic, and social evolutions.** Creating sound future policy requires the understanding of both past and current dynamics. Creating the data to develop this knowledge requires sharp interdisciplinary skills, some of which are specific to a country, or even to a region, because of the heterogeneity of historical business rules and practises. These peculiarities call for an ad hoc **Research Infrastructure** that can also connect to other existing systems.

The **EURHISFIRM project** meets the need for such a **benchmark research infrastructure** in Europe. It will design **the most comprehensive long-run economic and financial database in the world.** It will handle data on European companies such as accounting, funding and investment, stock exchange data, governance rules, directors, patents, and headquarter locations. The creation of a **vibrant European community** will support the project's development based on **innovative technologies**, which will **connect, collect, collate, align, and share detailed, reliable, and standardized long-term company-level data for European stakeholders: policy makers, scholars, and private companies.**



## B. The project

### B.1 The foundations

This project stems from **the EURHISTOCK research group** which has been gathering specialists in economic and financial history every year since **2009**. This group has acknowledged the existing datasets' lack of completeness, the lack of coordination among the initiatives, and the heterogeneity of European data collection practices. This observation has led some countries, such as Belgium and France, to initiate coordinated efforts to build long-term **structured data with digital techniques**. Other countries in the consortium have started to collect data or are exploring their datasets' comparative issues.

### B.2 The concept

The **EURHISFIRM project** relies on innovative technologies to collect, merge, extract, collate, align and share **detailed, high-quality historical firm level data for Europe** (Figure 1).

Concerning the **inputs**, EURHISFIRM is developing innovative technologies to 1) to **merge** existing high-quality historical data; 2) to **link** them to other historical and contemporary databases; 3) to **enrich** existing data **with web-based open resources**.

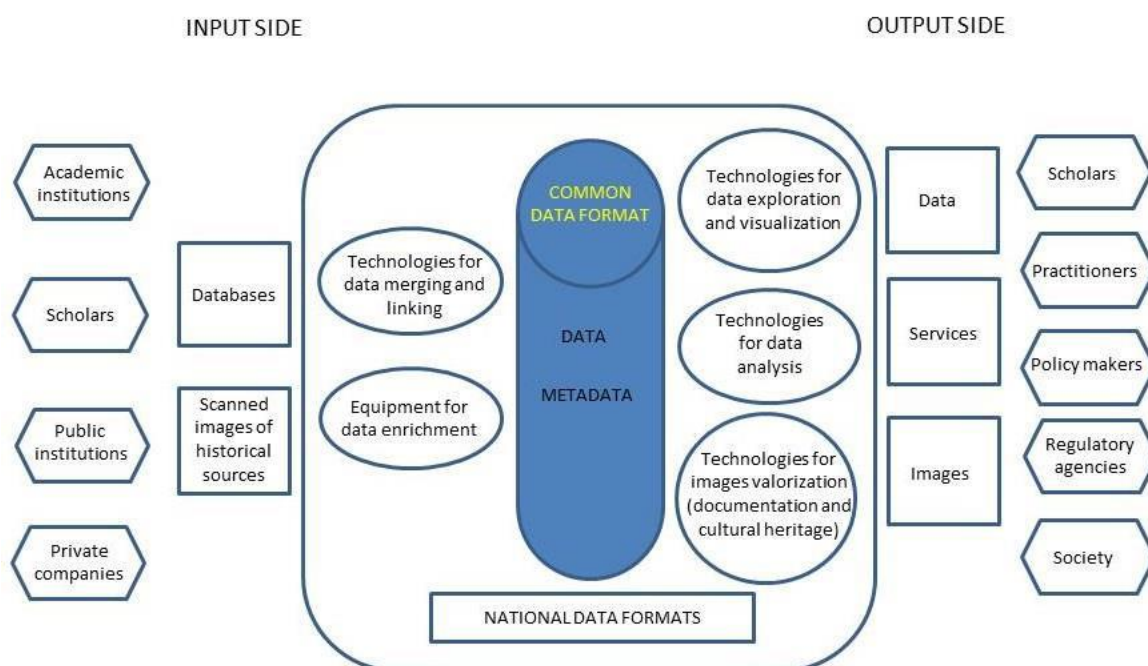


Figure 1 : The EURHISFIRM Research Infrastructure Concept

**Common format and semantics** will ensure the coherence of the data. These require a harmonization process that will gradually transform local and national heterogeneities (resulting from institutional differences or different data ownerships) into common standards. The data formats and semantics will be first set up at the country level by the consortium's national coordinators in close cooperation with national communities; it will then be reiterated towards European standards.

Concerning the **outputs**, EURHISFIRM will offer the stakeholder community with **data, services and images** for contribution to the **European cultural heritage**. The project is developing **technologies to explore and visualize** large and complex amounts of financial data in a user-friendly way, making information easily accessible for both experts and citizens. It is developing **technologies for data analysis and mining**. It will make available expertise, data-connection and data-extraction technologies in order to inspire new data collections (particularly from **young scholars**) and will create an expanding community. It will provide **images of historical sources** to provide high-quality historical data **documentation** and to preserve the **European cultural heritage**.

The principles of data merging, collating, and collecting, data standards, and services to users will be jointly determined with the **community of stakeholders**.

### B.3 Methodological approach

The **methodological approach for the RI's design** will integrate the development of its two logical parts: the data design and the platform design (Figure. 2).

The **data design** is based on an in-depth **survey and assessment of both the available data and the companies' historical sources (Work Package 4)**. To make the work manageable, the survey will be limited to 19<sup>th</sup>- and 20<sup>th</sup>-century historical printed serial sources on publicly traded companies. Accordingly, **Work Package 5** will develop **European common standards** and a **process to normalize and map data** collected from local sources using those standards. This convergence will encourage the technological development **to spark a "big data revolution" in the historical sciences and to push the technological boundaries**. **Technologies for merging** high-quality historical data **and for linking** them to other historical and contemporary databases will be developed by **Work Package 6**.

European archives and libraries have preserved a wealth of serial printed sources on companies. Work Package 7 will design a set of tools to extract high-quality data from these sources at low costs. Additionally, the web is a mine of scattered and dispersed information on European companies over the long run, and an algorithm will extract and collate this information.





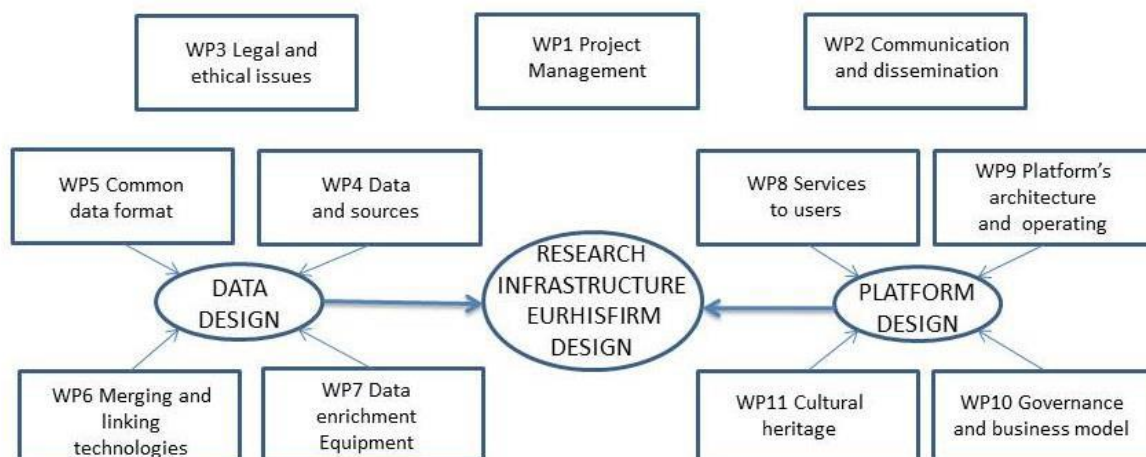


Figure 2 : Methodological Approach to the Concept

The **platform design** focuses on EURHISFIRM's future **services** to the community: the services are conceived and designed in close cooperation with the stakeholders (**Work Package 8**). The service design will guide the **platform's architecture and operations** (**Work Package 9**). Tight interconnections with the community and the analysis of related initiatives will drive both the **governance and business model** designs of EURHISFIRM (**Work Package 10**). The images produced within the Research Infrastructure will also serve as high-quality sources of data documentation and as valuable contributions for preserving the **European cultural heritage** (**Work Package 11**).

This approach is supported by the **project management** (**Work Package 1**) in charge of both the overall coordination of the project and the **final design study**; the **communication and dissemination Unit** (**Work Package 2**) for establishing and expanding a **vibrant stakeholder's community**; and the **legal and ethical unit** (**Work Package 3**) for exploring issues related to the dissemination and use of data and images, partnerships, contracts and the consortium agreement.

## Executive summary

The progress for each Work Package, as well as the cross-institutional Work Group on Identification and Standardisation (WGIS), is summarised below. Due to the complications related to the COVID-19 health crisis, this report was delivered with delay; the reporting dates for each Work Package varies between March and September 2020.

### Working Group on Identification and Standardization (WGIS)

Standards are indispensable fundamentals of any ICT-based (information and communications technology-based) system integration and system interaction design. In this environment, the Working Group on Identification and Standardization (WGIS) of EURHISFIRM aims to increase the communication and collaboration between the different work packages to facilitate the implementation of standards in the project by exchanging problem descriptions and offering solutions regarding their individual Work Packages as well as for the overarching EURHISFIRM goals and objectives. The standardization of the common data model progresses according to the roadmap of WGIS. We have decided upon the two increments of Entity data (identifying) and Entity data (historical). In the future, WGIS will focus on standardizing additional model elements such as Entity data (staging/integration/access) and information on financial instruments and financial statements.

The Architecture Team, as a subgroup of the WGIS, discusses certain technical topics in more depth. These topics include the Common Data Access Service of Work Package 5, the use of identification regimes and data matching methodologies, and the consideration of potential technology platforms and architectures for network data management.

### Work Package 1: Project management

Work Package 1 is responsible for all coordination tasks for the project in: (1) overall strategy and administration, (2) scientific aims, and (3) technical milestones. In terms of the overall strategy and administration, the project continues its execution, keeping in mind FAIR (findability, accessibility, interoperability, and reusability) data principles and European-level collaborations. In terms of project outputs, most required deliverables and milestones to-date have been submitted (29/54 [54%] and 9/21 [43%], respectively). Following the COVID-19/coronavirus pandemic that began to spread in Europe in the early months of 2020, Work Package 1 was in touch with the Project Officer in order to discuss the consequences and arrangements for possible delays in the deliverables; it was agreed that these would be noted in the deviations sections in the reporting platforms.

In the second part of the project, the Work Package focused on increasing the coordination of the Work Packages, in particular the technical work within the Work Packages 5, 6, 7, and 9. Additionally, a key focus was on the community building and to improve the project's visibility to partners and potential stakeholders. Work Package 1 also completed the third version of the Data Management Plan with other Work Package members as scheduled in the project planning.

### Work Package 2: Dissemination and communication

The strategic aim of Work Package 2 is on active communication and dissemination to the widest possible audience. The first steps focused on creating the identity of the project as a basis for further identification during all promotional activities. Based on these tools, Work Package 2 works intensively on the wide, active and proper communication of all Work Package Teams' activities directly related to the project, as well as those that can positively affect the perception of the project. For this purpose, in addition to the particular care for the timeliness and attractiveness of the website, Work Package 2 has included selected social media in the active promotion. Using these information distribution channels, the content is sent to stimulate a group of stakeholders on an ongoing basis.

### **Work Package 3: Legal and Ethical Issues**

The final report on intellectual property rights related issues will be finalized according to the schedule. It will lay out the relevant property rights and what criteria have to be fulfilled for those such rights to persist in any of the source materials. Furthermore, it will examine which actions in the context of the EURHISFIRM project might require an authorization by rightholders or instead fall under the scope of an exception to copyrights or related rights. In general, an authorization will not be necessary for reproductions in the EURHISFIRM context in a scientific setting. Making actual protected subject matter available to the public will often require an authorization. For details refer to the final report.

The report on ethical issues, namely on the protection of privacy and personal data, is on track.

### **Work Package 4: Data and sources inventory and documentation**

The tasks set in Work Package 4 are key inputs for the later works of several other Work Packages. In short, it draws up an inventory of the sources that are available for building the databases our research infrastructure aims to do, describes their contents and semantics, assesses their quality, and ultimately defines the project's documentation standard. Having appropriate metadata standards is crucial for these tasks. After a review of several standards, it was concluded that the Data Documentation Initiative (DDI) family provides the most appropriate front- and back-end metadata standards for the project. It was also proposed to produce and edit metadata with the Harvard Dataverse and the Colectica Designer software. A preliminary manual (protocol) for uploading datasets has been approved by the Steering Committee and a scientific paper on sources had been published online. Thanks to the inputs of all participating teams, the Work Package has been completed on time.

### **Work Package 5: Common data model**

Work Package 5 focuses on the development of concepts and the design of an overarching European Common Data Model with interfaces to the process architecture. It progressively sets standards, identifies best practices and facilitates harmonization processes for the integration of European, long-term, firm-level data from heterogeneous, historical, national sources.

Towards this goal, the team has firstly documented models available from within as well as outside the institutions of the consortium and evaluated their strong points and their weaknesses. Work Package 5 concluded the analysis of the back-end elements of the common model and introduced two central



preliminary design principles that support and evolvable implementation of the common data model. Currently, the team focuses on the analysis of the front-end design elements of the model.

### **Work Package 6: Data connecting and matching**

The key idea behind the research infrastructure (RI) is that users should be able to query it without needing to know in which databases the required information is to be found. Behind the screens, the infrastructure therefore needs to be able to locate and connect the relevant databases and to retrieve information from them in a consistent way. In Work Package 6, the technologies that allow this will be developed and tested. As envisaged, this work is currently about halfway. A first conceptual report on data matching has been completed, including possible solutions for major issues that can be expected when matching long-term data originating in different European countries. For the first case currently being tested, the two most advanced databases hosted by member-institutions, the Paris and Antwerp databases (DFIH (Données Financières Historiques) and SCOB (Studiecentrum voor Onderneming en Beurs), respectively) covering the Paris and Brussels exchanges, are linked and integrated. To this end, the two teams involved have worked out the technical details to achieve this. At the same time, other databases, both internal and external to the project, have been singled out for testing purposes.

### **Work Package 7: Data extraction and enrichment system**

After building the library of document components detectors for structure recognition and the general-purpose text recognizer (OCR) during the first year, Work Package 7 has built on top of them a first version of the data extraction system on yearbooks (section 3) and on price lists (section 4). These two prototypes have been designed on the French Desfossés Yearbook 1962 and the German Yearbook 1913-1914 (Handbuch der deutschen Aktiengesellschaften) for the yearbook extraction system and on the official price lists for Brussels 1912 and Paris 1961-1962 for the price lists extraction system prototype.

Work Package 7 is currently working on the Spanish yearbook 1929-1930 by building a more generic extraction system for yearbooks, combined with a new design of the interface for the active learning process, which both improve the ability of the system to be applied to a new yearbook. On official price lists, Work Package 7 is working on the cross-validation module for the stock prices and on an interface for expert's interaction with the extraction system. Additionally, work on the Brussels and Paris official price lists are underway. We will then apply and adapt the system to the official price lists for Madrid 1934. In the same way as for yearbooks, the extraction system for price lists will be improved with a generic approach. Web-linking of some extracted information will be developed mainly on the French documents. We will also continue to evaluate the two systems on a larger dataset extracted from the sample dataset: Belgian, French, German, and Spanish documents.

### **Work Package 8: Interaction with users**

The objective of Work Package 8 was to determine the optimal design of the data and services that EURHISFIRM RI should provide by gathering and analysing the preferences of potential end-users and key stakeholders (academics, practitioners, regulators, etc.). A large-scale survey, via an online questionnaire, was developed in order to identify the preferences of potential end-users and key stakeholders of the EURHISFIRM project. The survey was conducted and its results were analysed. The Work Package then



identified qualified people and conducted semi-structured interviews on their perspectives and preferences for the design of EURHISFIRM. The insights from these interviews were then synthesised with the results of the survey and recommendations were shared with EURHISFIRM partners.

### **Work Package 9: Infrastructure policy and architecture**

Work Package 9 designs the architecture and the operation of the RI, with regards to access, security, support and maintenance, in cooperation with ESFRI (European Strategy Forum on Research Infrastructures) Landmark CESSDA (Consortium of European Social Science Data Archives) and its Data Service Provider GESIS. Users' preferences on data and service design guide the platform's architecture and operation. Accordingly, the security system, the maintenance and the desk management of the platform are designed and estimated. The platform's architecture and operation are compatible with the National Focus Points' and site's policies.

The Work Package also assesses the optimal level of integration of EURHISFIRM with existing RI such as CESSDA and DARIAH, following the user requirements' specification and RI policies. Most efforts have been put into designing policies and requirements for usage of the future system. Aspects such as user types; access rights; GDPR consequences; classifications for Confidentiality, Integrity and Accessibility have been proposed. The basis for this work comes from materials from CESSDA, standards such as ISO 25010 (software quality), and reference materials for Research Infrastructures such as EOSC and FAIR.

Work Package 9's output will contain a description of required functionalities with as much detail as possible based on the input from other Work Packages.

### **Work Package 10: Business model and governance**

The objective of Work Package 10 is to develop a business and governance model that contributes to the articulation of the EURHISFIRM's platform design by drawing on best-practices as developed by existing RIs. The first report of WP10 (D10.1) articulates the key steps in the process of formulating the Value Proposition for the EURHISFIRM Research Infrastructure. It compares alternative models, paying special attention to the implications of EURHISFIRM's adherence to Open Science principles, and its integration with EOSC (European Open Source Cloud) and SSHOC (Social Sciences and Humanities Open Cloud). The report focuses on the identification of stakeholders, the development of services, the identification of revenue sources and the demonstration of value. Its conclusions suggest that EURHISFIRM in its operational stage is likely to be a prevalingly Public RI providing both basic and value-added services, and operating as an ERIC service provided within CESSDA.

### **Work Package 11: Cultural heritage**

Work Package 11 explores concepts and tools to stimulate the lasting conservation of the digitized material and provides guidelines for making those materials publicly accessible. It also explores innovative ways to use digitized images as documentation for the data extracted from them and evaluates alternative strategies to use digitized material. More specifically, Work Package 11 has three main objectives:



- ▶ The use of digital images to document data and inspire further research and Identify sources of interest for cultural heritage;
- ▶ The promotion of Europe's cultural heritage by facilitating digital preservation and online accessibility of sources with a unique historical value;
- ▶ The mobilization of digitized images of historical sources as an exceptional additional documentation for the data (including the exploration of ways to make materials accessible and connected to EURHISFIRM data).

Logistical and operational progress have been carried out to support and develop future goals and milestones. Over the coming weeks, work will be carried out to finalize the first milestones and deliverables (i.e. D11.1 Strategies and practices to value cultural heritage).



## Work Package 1: Project management

### I. Introduction

Work Package 1 is responsible for all coordination tasks for the project in: (1) overall strategy and administration, (2) scientific aims, and (3) technical milestones to successfully complete all aspects of the project according to the project proposal.

During the second year of the project, WP1 concerned with the project's long-term strategic and operational goals described in the first version of the yearly progress and strategy report (D1.3 [https://eurhisfirm.eu/wp-content/uploads/2019/04/EURHISFIRM-D1.3\\_FirstYearlyReport.pdf](https://eurhisfirm.eu/wp-content/uploads/2019/04/EURHISFIRM-D1.3_FirstYearlyReport.pdf)):

- ▶ Coordination among Work Packages: in the first year of the project, the goal was to establish coordination among the Work Packages, both operation- and content-wise. It was then identified that for the second part of the year, tighter coordination among the Work Packages, particularly among those that concern the technical work (such as Work Packages 5, 6, 7, and 9).
- ▶ Community building: Once the project framework was established, EURHISFIRM recognised the importance of building a solid research community in order to ensure the infrastructure's utility and sustainability. The challenge includes considerations such as:
  - ▶▶ How do we continuously reach out to the different stakeholders so that they know what we are doing and give us feedback on what we should be doing (in terms of scope, service providing, technology)?
  - ▶▶ How do we enable and stimulate Pan-European research using data spanning long time periods?
  - ▶▶ How do we stimulate and help the construction of new databases?

### Coordination of project logistics and of overall strategy

To ensure sound execution of the project, Work Package 1 handles the overall project logistics and strategy in cooperation with the partner institutions' administration teams, the Executive Committee, the Steering Committee, and all other EURHISFIRM team members. The main tasks include creating and monitoring timelines for inter-Work Package projects, quality check of deliverables, coordinating the progress reports (annual reports, midterm and final reports), aligning the overall strategy with the Work Packages, and handling administration (budgets, compliance with European Commission (EC) project policies, coordinating official documents and changes, and establishing internal project protocols).

In the second part of the year, Work Package 1 has also dedicated itself to increasing knowledge of the overall research landscape and developments concerning the European research infrastructure ecosystems and of the research community in the economic, financial, and historical fields.

### Coordination and development of scientific work (economic history research)

Two long-term national databases currently exist in the consortium: SCOB (Studiecentrum voor Onderneming en Beurs) of the University of Antwerp and DFIH (Données Financières Historiques) of the





Paris School of Economics. The experiences from these projects demonstrate that in addition to technical competences, substantial economic history knowledge is necessary to build database infrastructures compatible with the data's historical nuances and research needs. In other words, a close collaboration between technology and field-related research is of paramount importance. In order for EURHISFIRM to take advantage of these experiences and knowledge, Work Package 1 collaborates with other consortium members to ensure 1) the RI's scientific relevance in economic history research and 2) the project's overall scientific coherence.

### **Coordination and development of technical (information systems) work**

Work Package 1 is also involved with the technical coordination of the project with the other Work Packages to ensure technical consistency and alignment with the project's scientific aims.

## **II. Progress: Logistics and Resources**

All of the required positions for Work Package 1 were filled by July 2018. The team members are based at the Paris School of Economics. In addition, two subcontracting firms have been engaged for the technical work in order to reinforce the specialised technical work needed to complete the project. As the coordinating institution, Work Package 1 also remains in contact with the other Work Packages and their administration teams to collectively monitor the project's overall human resource needs.

Between 2019 and 2020, two project amendments were carried out, consisting of various changes such as modifications in deadline dates, budget changes among institutions, person-months, among other changes. These changes are noted in the history of changes log in the amendment documents.

Following the approval of the second project amendment, CESSDA ERIC has joined as the twelfth institution of the EURHISFIRM consortium. CESSDA ERIC will also contribute to some of the Work Package tasks, as noted in the person-months allocated.

## **III. Progress: Project Achievements**

In terms of project outputs, most required deliverables and milestones to-date have been submitted (29/54 [54%] and 9/21 [43%], respectively). Following the COVID-19/coronavirus pandemic that began to spread in Europe in the early months of 2020, Work Package 1 was in touch with the Project Officer in order to discuss the consequences and arrangements for possible delays in the deliverables; it was agreed that these would be noted in the deviations sections in the reporting platforms. Additionally, for each deliverable or milestone submitted with delay, we have provided comments in the continuous reporting platform.

Specific achievements, including those related to the points mentioned in part I (introduction), include:

- **Coordination among Work Packages:** As mentioned above, in the second year of the project, we focused on increasing the coordination among the Work Packages, particularly among those that concern the technical work (such as Work Packages 5, 6, 7, and 9). We built upon the foundations built during the first year to increase the cooperation among these particular Work Packages. The WGIS group (see the description in the Executive Summary), as well as other key large meetings that have





been held, play a key part in this goal. In January 2020, a large inter-Work Package meeting was held at the Paris School of Economics, which also helped solidify the goals and coordination moving forward, among which included the agreement to increase visibility of the work in progress in different institutions by sharing the deliverables with all inter-Work Package members (which are normally only sent to the Executive Committee for approval).

- ▶ **Community building:** Work Package 1 engages with Work Package 2 in order to improve the project's visibility to partners and potential stakeholders. Additionally, the theme of community building has been a key strategic goal throughout the second year, which has thus been added and discussed within the Steering Committee and Executive Committee meetings. Continuous activities within the project members to spread awareness of the project is ongoing. The project also endeavors to become an active player in the European research infrastructure ecosystem and has accepted the invitation to join the SSHOC project in the framework of the EOSC. CESSDA ERIC (coordinator of SSHOC) has also joined EURHISFIRM as the newest consortium member. EURHISFIRM also participated in various relevant conferences of interest and intends to expand its activities and sustainability by engaging further interested research institutions, which we also realized through our application to the INFRAIA-02-2020 (Integrating activities for starting communities) project call under the application title EurHisCom.
- ▶ **Data management plans:** In line with the project schedule, EURHISFIRM released its third data management plan in spring 2020. This version incorporates the relevant inputs following the project advancements in the second year and allows us to stay on track to follow the Horizon 2020 Programme Guidelines on FAIR Data Management. The final version will be released in spring 2021.

#### IV. Conclusion

In the second year of the project, Work Package 1 worked on elevating the project's direction and execution to the next level, which focused on key topics such as enhancing the coordination among the Work Packages, enriching the project's community, and staying up-to-date with the data management plan creation.

In the third and final year of the project, the key issues to focus on will be:

- ▶ Sharpening the long-term vision of EURHISFIRM in terms of utility and sustainability
- ▶ Continuing to build the project's community
- ▶ Continuing to engage with the European research community and staying up-to-date with the latest developments in policies and technologies
- ▶ Redefining key performance indicators, where applicable.



## Work Package 2: Dissemination and communication

### I. Introduction

Work Package 2 is responsible for disseminating and communicating the project in order to promote its visibility and long-term success. It also focuses on the proper long-term communication and dissemination of the project through a variety of channels.

The first three deliverables and milestones were completed during the first year of the project:

#### 1. Deliverables:

- ▶ D2.1 Dissemination and communication plan (M2)
- ▶ D2.2 Website and project identity (M4)
- ▶ D2.3 Inventory of European and national distribution networks (M6)

#### 2. Milestones:

- ▶ M2.1 Dissemination and communication plan (M2)
- ▶ M2.2 Project website (M4)

According to the project strategy, at the second year Work Package 2 aligns with the planned schedule. The project promotion and communication will integrate the community building strategy in order to promote a strong stakeholder network.

Work Package 2 is working on the two last deliverables, D2.4: Involvement of new stakeholders as participants to the General Assembly (M36), D2.5: Presentations of the project at events and conferences (M36).

### II. Progress: Logistics and Resources

Work Package 2 has fulfilled the necessary logistics and human resources requirements.

### III. Progress: Project Achievements

The first three deliverables and milestones were completed during the first year of the project (full description of activities are available in the previous edition of this report):

#### 1. Deliverables:

- ▶ D2.1 Dissemination and communication plan (M2)
- ▶ D2.2 Website and project identity (M4)
- ▶ D2.3 Inventory of European and national distribution networks (M6)

#### 2. Milestones:



- ▶ M2.1 Dissemination and communication plan (M2)
- ▶ M2.2 Project website (M4)

The most important activity within WP2 has been the organization of General Assembly Meeting, which was held on March 15-16 2019 in Wroclaw.

In the second year of the project, Work Package 2 pays particular attention to the project assumptions for 2019/2020. WP2 focuses on maintaining and carrying for communication and dissemination continuity.

The main assumption is to diversify communication channels to get the most effective results in reaching the widest possible group of stakeholders.

Work Package 2 uses with two main communication tools:

- ▶ Website:
  - ▶▶ keeping the data up to date
  - ▶▶ ongoing updates and placement of information

- ▶ Social media

The selection of available social media is primarily based on the use of those whose quality does not compromise the status of the project.

- ▶▶ Facebook
- ▶▶ LinkedIn
- ▶▶ Twitter
- ▶▶ Research Gate.

#### IV. Conclusion

Activities undertaken by the Work Package 2 team are proceeding and are being implemented according to the schedule.

The assumptions for the next months of the project are focused particularly on communication and are concentrated on reaching out to stakeholders group with information related to the project.



## Work Package 3: Legal and ethical issues

Work Package 3 (WP 3) is divided into two distinctively different sections which are treated consecutively. The first section (WP 3.1) deals with property rights issues, which need to be identified and solved. The second section (WP 3.2) deals with ethics which is mainly understood as the legal rules protecting privacy and personality rights of natural persons. However, a clear distinction between legal and ethical norms has to be made according to the legal thinking on the continent. The principle legal source is the General Data Protection Regulation (GDPR)<sup>1</sup>.

### I. Introduction

During the creation of a database huge amounts of data are collected. During this process digital reproductions of source materials will be made and shared. If the source materials are subject to ownership rights such as copyrights and/or related rights, those actions might require authorization of the holders of the copyrights and/or related rights. WP 3.1 identifies and examines such property rights and will soon deliver a report on what issues might arise in this context. It will serve as a guideline on what potential issues are and point out ways to deal with those issues.

The preliminary study of the rules protecting privacy and personal data (3.2) showed that the collection and sharing of firm data has in general little significance regarding a possible infraction of personality rights since they almost exclusively protect natural persons and, in general, only living persons. Nevertheless the cases in which they might be touched have to be identified and assessed. The main source for the relevant legal rules is the law of the EU since the law has been harmonized to a large extent.

### II. Progress: Logistics and Resources

For WP 3.1 Fabian Brandt was employed as a research assistant over the course of the last year. He worked together with Prof. Alexander Peukert to draft the report on intellectual property rights issues.

For WP 3.2 Professor Dr. Dr. h.c. Helmut Siekmann is paid by the project since November 2019. He is distinguished Professor of the Goethe University Frankfurt.

### III. Progress: Project Achievements

#### Work package 3.1

The report on IPR design has identified which rights on the EU law level are relevant in the context of the EURHISFIRM database. Those rights are copyrights in literary works or database works and the right of the maker of the database. Further, the report lays out the conditions under which source materials might be subject to those rights. It illustrates those requirements with exemplary source materials from the WP 4.2 Deliverable (Report on the Inventory of Data and Sources). The report additionally provides the respective terms of protection, i.e. states after what periods of time the ownership rights expire or already have expired. Aside from laying out which exclusive rights those copyrights and related rights grant, the report

---

<sup>1</sup> Official Journal of the EU L 119/1.

furthermore identifies what exceptions to those exclusive rights are applicable in the EURHISFIRM context. The text and data mining exceptions (Articles 3 and 4) of the new Digital Single Market Directive are particularly relevant here.

With this three-step structure for each relevant right – i.e. criteria for protection, exclusive rights granted by the protection and applicable exceptions – the report will serve as a guideline on whether source materials are protected in the first place, on which utilizations are subject to exclusive rights and what exceptions might still render an authorization of the rightholders unnecessary.

Besides potential issues with intellectual property rights the report also examined German unfair competition law. While there are few critical issues in this regard, there is little harmonization of unfair competition law on the EU law level.

Aside from identifying relevant ownership rights in regard to the source materials, the final report also lays out what rights might persist in the final EURHISFIRM database.

The report is being revised at the moment. It will be finalized in the very near future and will be finished and delivered according to schedule.

### **Work package 3.2**

Aside from material and researching the judicature, several talks on the relevant questions were already prepared and presented. One important insight is that the material scope of the General Data Protection Regulation (GDPR) is quite limited as regards EURHISFIRM: Automated processing of personal data Article 1(1), Art. 2(1) GDPR of natural persons living (recital 27 GDPR: “not apply ... to deceased persons”).

The work of EURHISFIRM is privileged as scientific or historic research when balancing infringements of general personality rights which, in addition, also serve as basis for General Data Protection Regulation. An explicit exemption is provided by Articles 5 and 89 GDPR: Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject. “Further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes”.

The substantial part of the work on the formulation of the report was – according to the time line – planned for the months following the final version of the report on WP 3.1. In so far the progress is on time.

## **IV. Conclusion**

The final version of the report on WP 3.1 will serve a guideline on which source materials might be subject to property rights and which utilizations of protected subject matter might require authorization.

The report on WP 3.2 follows closely the report on WP3.1; both in format and methodology. This way the two (partial) reports can easily be consolidated at the end. The reader, namely if coming from a non-legal field, will be aided to find a way through the host of sometimes bewildering distinctions in legal arguing. It will deliver guidelines on practical questions that are apparent. Since GDPR is fairly new and contains a



lot of vague terminology, which has so far not been concretised by court decisions, for many questions only a preliminary assessment can be delivered.



## Work Package 4: Data and sources inventory and documentation

### I. Introduction

For the protection and convenience of the public and investors, governments, stock exchanges and commercial publishers have published tremendous amounts of information on companies in general and publicly traded companies (i.e. companies whose shares and debentures are listed on a stock exchange) in particular. Some of these data have already been worked into datasets and databases by individual scholars and research groups. Work Package 4 aims to produce homogeneous documentation of these historical printed sources and datasets. It follows a logical build-up (illustrated in the diagram below), first selecting a preliminary metadata standard and software for data and sources documentation (task 4.1), then identifying and categorising existing sources and datasets (task 4.2), followed by an in-depth historical contextualisation of their contents (task 4.3) to produce metadata for the most important sources identified in task 4.2 according to the standard chosen in task 4.1 (task 4.4). Task 4.4 will also serve as a test-case for the suitability of the selected metadata standard. Task 4.5 will therefore consist of an evaluation of and a final decision on the selected metadata standard and information system.

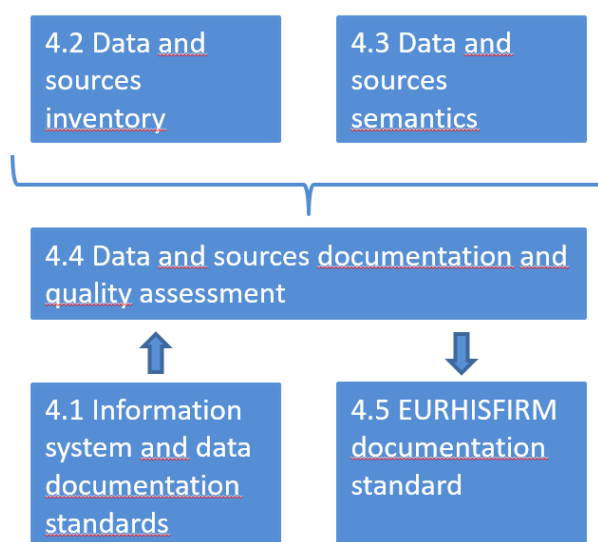


Figure: Diagram of Work Package 4

### Progress

Work Package 4 has finished on schedule in September 2019 (M18).

### II. Progress: Logistics and Resources

#### Human Resources

The lead beneficiary of Work Package 4 is the University of Antwerp. Johan Poukens has been working on Work Package 4 full time since May 1, 2018. Johan obtained master's degrees in History and Archival Sciences from the Universities of Leuven and Brussels in 2006 and 2007 and a PhD in History from the University of Leuven in 2017. Prior to joining the EURHISFIRM team at the University of Antwerp, he worked as an information specialist at Erasmus University College in Brussels and built up considerable



expertise in metadata standards, cataloguing software, database management and business process modelling.

To accomplish the objectives of Work Package 4, Johan has cooperated with information specialists from GESIS, PSE, EUR (tasks 4.1 and 4.5) and topical experts in business and financial history from his own university as well as from research institutions in Amsterdam (KNAW), Belfast (QUB), Frankfurt (SAFE), Madrid (UC3M), Paris (PSE) and Wroclaw (WUE) (task 4.2, 4.3 and 4.4).

### Tools and Technologies

Task 4.1 (see section III below) resulted in the preliminary selection of **Data Documentation Initiative Lifecycle** (DDI 3.2) and **Colectica Designer** as the metadata standard and software for the documentation of datasets and sources in task 4.4 (see section III below). Standards and software were “recommended” by the Working Group on Identification and Standardisation (WGIS) in its meeting of 27 June 2018 and “first read” by the Steering Committee on 3 July 2018. In task 4.5, it was proposed that the DDI family of standards presents the best possibilities for the EURHISFIRM infrastructure: **DDI Codebook** (DDI 2.5) for front-end purposes (i.e. documenting data and sources during upload using the **Harvard Dataverse** software) and DDI Lifecycle (DDI 3.2) for back-end purposes (i.e. data harmonisation).

In addition, the project team has assessed the suitability of FIBO (Financial Industry Business Ontology) for the infrastructure EURHISFIRM is designing. FIBO is a financial industry initiative to define financial industry terms, definitions and synonyms using semantic web principles such as RDF and OWL. This in-progress financial ontology is the output of a ten-years’ work of specialists in financial data and represents a valuable opportunity for the project as decided by the enlarged Executive Committee on 11 February 2019 in the light of the decision of the Steering Committee on 19 October 2018 on the test of RDF based web-semantic technology for the EURHISFIRM infrastructure. Work on the possible articulation of DDI Lifecycle and FIBO Ontology has been undertaken in task 4.5 and its results are included in D4.5 (see section III below).

## III. Progress: Project Achievements

### Progress

Work Package 4 consisted of five tasks in total:

1. Information system and data documentation standards (M2-M3)
2. Data and sources inventory (M3-M9)
3. Data and sources semantics (M6-M12)
4. Data and sources documentation production and quality assessment (M7-M14)
5. Defining EURHISFIRM documentation standard (M13-M18)

All tasks have been completed since September 2019.

### Submitted Deliverables

All of Work Package 4’s deliverables and milestones have currently been completed within the specified timeframe. This would not have been feasible without the indispensable input from all of the participating teams.





*D4.1: Report on the Information System and Documentation Standard (M3, June 2018)*

Having appropriate data documentation or metadata standards is crucial for the tasks of Work Package 4 and for ERUHISFIRM in general. Metadata is defined by the International Organisation for Standardisation (ISO) as "data that defines and describes other data" or, simply, "data about data". The report on the information system and documentation standard reviewed several general and social-science specific metadata standards. It was concluded that the Data Documentation Initiative (DDI) provides the most appropriate one, namely the DDI-Lifecycle standard, which can be produced and edited by the Colectica Designer software.

*D4.2: Report on the Inventory of Data and Sources (M9, December 2018)*

The data and sources inventory identifies the principal official publications of company information, stock exchange price lists and yearbooks with summary governance and financial information on publicly traded companies in each of the participating countries, as well as existing datasets and databases, and bibliographies for financial and business history. The report also provides some context on the origin of historical sources and a summary description of their contents (the exhaustive and detailed description of all information contained in the principal price lists, yearbooks, datasets and databases is reserved for D4.4).

*D4.3: Report on the Semantics of Data and Sources (M12, March 2019)*

The sources identified in task 4.2 span a period of over 200 years. During this time, company legislation, accounting and financial reporting regulations and the microstructure of markets, to name only a few, have evolved tremendously. In some cases, this results in a huge gap between our present-day conceptions of, for instance, company legal forms, governance structures and financial instruments, and the nineteenth or early twentieth century reality. The report on data and sources semantics therefore contextualises the categories of information commonly found in stock exchange yearbooks and price lists. It does so by providing clear definitions and a detailed, historical overview of legislation, regulation and customs in the fields of company identification, corporate governance, securities trading and financial reporting.

*D4.4: Report on Data and Sources Documentation and Quality Assessment (M14, May 2019)*

A selection of printed stock exchange price lists and yearbooks, one of each category for every country in the consortium, as well as several existing databases and datasets on financial history (including SCOB, D-FIH, Eurofidai and the London Share Price Database) were documented according to the DDI Lifecycle standard with the Colectica Designer software. The output has been made available both in the machine-readable DDI-XML format, as well as in a human-readable PDF file. It includes bibliographic records for each source or dataset as well as a description of their coverage (topical, temporal and geographic) and the variables contained therein (e.g. company names, prices, interest rates). For datasets and databases, it furthermore includes information on the funding of the data collection, data collection methods, data sources, data files (including file types, e.g. Excel or SQL), and data layout (i.e. how variables are organised in related tables). The report furthermore includes a ranking of source types and a methodology for assessing the quality of datasets by means of a questionnaire.

*D4.5: Report on EURHISFIRM Documentation Standard (M18, September 2019)*

Building on the results of task 4.1 and the experiences of task 4.3, the DDI family of standards was chosen as EURHISFIRM's standard of documentation. DDI is widely used in the social sciences research community



(including CESSDA). In the report, we propose different standards for different tasks: DDI Codebook (DDI 2.5) for documenting data during upload and DDI Lifecycle (DDI 3.2) for the subsequent harmonisation of data. We also evaluated two software packages for documenting datasets in DDI: Dataverse and Colectica Designer.

## Completed Milestones

### *M4.1: Protocol of Data Documentation consistent with the EC Data Management Plan (M18, September 2019)*

The protocol of data documentation is a manual for uploading data and metadata in the DDI 2.5 format with the Harvard Dataverse platform. It is intended for researchers who want to contribute their data to EURHISFIRM and facilitates meeting the FAIR (Findable, Accessible, Interoperable and Re-usable) guiding principles for research datasets.

### *M4.2: Scientific Paper on Sources and Data (M18, September 2019)*

For the final scientific paper on data and sources, we built on the information collected during tasks 4.2 and 4.3 to give a complete overview of provisions on the mandatory disclosure and publication of information on the governance and ownership of joint-stock companies in the company laws of the consortium countries from 1800. This paper has been re-written into an article and submitted to a leading journal in the field of economic history.

## IV. Conclusion

Work Package 4 has built up expertise regarding existing datasets and historical printed serial sources on publicly traded companies from 1815 onwards, as well as on data documentation (metadata) standards. Other Work Packages are drawing on its output for the elaboration of a common metadata model (Work Package 5), data connecting and matching technologies (Work Package 6), a system for automated data extraction (Work Package 7), and designing the infrastructure policy and architecture (Work Package 9). Some of the output of Work Package 4, however, also presents a valuable resource for future research in itself. The data and sources inventory (D4.2) and data and sources documentation (D4.4) can for instance guide researchers to sources of data for business and financial history. The report on data semantics (D4.3) can help them to contextualize and understand the information found in these sources, as well as in the EURHISFIRM Research Infrastructure. A scientific paper on the data and sources (M4.1) has made some of this information available to the research community. Its emphasis is on government publications (e.g. official journals and newspapers, collections of laws and decrees). A complementary scientific workshop on stock exchange price lists will be jointly organised with the *Nederlands Economisch-Historisch Archief* (NEHA) in November 2020. This workshop prepares a session on this subject during the World Economic History Conference (WEHC) of 2021 in Paris.



## Work Package 5: Common data model

### I. Introduction

Work Package 5 focuses on the development of concepts and the design of an overarching European Common Data Model with interfaces to the process architecture. It progressively sets standards, identifies best practices and facilitates harmonization processes for the integration of European, long-term, firm-level data from heterogeneous, historical, national sources. Towards this goal, we have firstly documented models available from within as well as outside the institutions of the consortium and evaluated their strong points and their weaknesses. Secondly, based on the observations of the first step, Work Package 5 proposes preliminary design principles for a common model of historical, European firm-level data with information that is spanned in three dimensions: financial information, accounting information, and management information.

### II. Progress: Logistics and Resources

To facilitate the coordination and standardization among the Work Packages of the project, Work Package 5 also coordinates the Working Group on Identification and Standardization (WGIS). Standards are indispensable fundamentals of any ICT-based (information and communications technology-based) system integration and system interaction design. In this environment, the WGIS of EURHISFIRM aims to facilitate the implementation of standards in the project, based on increased communication and collaboration between the different Work Packages. Members of the WGIS exchange problem descriptions and propose solutions regarding their individual Work Packages as well as for the overarching EURHISFIRM goals and objectives.

The Architecture Team, as a subgroup of the WGIS, currently discusses certain technical topics in more depth. These topics include the Common Data Access Service of Work Package 5, the use of identification regimes and data matching methodologies, and the consideration of potential technology platforms and architectures for network data management.

Work package 5 is supported by two young researchers. Pantelis Karapanagiotis is responsible for introducing back-end design elements and Lukas Manuel Ranft is responsible for the front-end elements of the model. In addition to this, the team is supported by Fatemeh Zare, as a student assistant, for preparatory work for Milestone 5.1 and by Jefferson Braswell, as an external consultant, for coordinating the standardization efforts of WGIS. The request to engage Braswell as subcontractor in the project is part of an amendment that has recently been submitted.

### III. Progress: Project Achievements

Report D5.1 reviewed a selection of existing micro-level data-model implementations both from within as well as outside the consortium's countries and identified best design practices. Methodologically, it introduced a conceptual separation of data models concerning the time and the cross-country domain and reviewed representative implementations from each subpart. The analysis identified the principle of maintaining the original information as the most crucial characteristic of designing evolvable models.



Report D5.2 concluded the discussion of the preliminary back-end design concepts of the common data model. The approach of the report is characterized by the principle of least intrusiveness. The proposed solutions respect national idiosyncrasies and allow national centers to advance in a collaborative but independent manner. The report outlined that the common model's implementation mostly benefits from employing both relational and non-relational technologies to address different issues. Last but not least, it highlighted the central importance of collaborative instruments in setting and maintaining common standards.

After extensive debates within and across all Work Package teams, the WGIS has introduced a one year plan that examines the most important data elements for the development of common standards. The current draft of the standard (v1-05\_202020304) describes the commonly agreed common model's standardization of legal entities. All Work Package leaders have agreed to this incremental specification of EURHISFIRM standards and, in particular, to the two first standardization steps. Future work will focus on financial instruments and subsequently on financial statements.

Report D5.3 consists of three independent streams of work, in which information is collected from different target groups. Within the framework of D5.2, a preliminary data model was created, which is being revised by WGIS. The results determined by WGIS make for the starting point of the first workstream of D5.3. Regarding the course of action linked to building the data model, all project members are to be asked in a survey for their opinion on appropriateness. In the second workstream, more than 13,000 publications from the financial sector are being analysed to extract the data used. Authors of academic publications were asked about the reason for using non-EU data. For authors indicating that insufficient quality and/or limited quantity prevented them from using EU data, use cases of their respected publications are being extracted. This shows both qualitatively and quantitatively the limiting use cases of EU data. The use cases of publications that have already used EU data will also be extracted. Thus both groups can be compared. As a third work stream, further international stakeholders, preferably policymakers and companies, are being interviewed in order to include the feedback of these groups in the process of creating the data model.

All these findings on the use and availability of the data will be used to revise the preliminary data model created in D5.2. This should be the result of Deliverable D5.4, which is the final data model of Work Package 5. The extendibility of this data model after an implementation caused by new sources or requirements is then the subject of Deliverable D5.5.

The team prepares a scientific paper on model evolvability that will apply the fundamental design principles that were identified in the Work Package's reports. Conditional on data availability, special emphasis will be given to accommodating historical sources of members without current implementations.

#### IV. Conclusion

Concerning the design and evolvability of the common data model and common data access service, Work Package 5 identified that the cornerstone characteristics can be summarized in two principles. The principle of 'maintaining the original information' and the principle of 'least intrusiveness'. When it comes to setting common standardization, the findings of Work Package 5, in line with previous international



experiences, indicate that acceptable international standards can only be set in representative collaborative instruments. In accordance with this finding and with the significant contributions of all the consortium's members, the WGIS progressively enhances the common model's standard.



## Work Package 6: Data connecting and matching

### I. Introduction

EURHISFIRM aims to develop, test and assess innovative technologies to match or connect existing historical and contemporary company-level data on European companies in order to align long-term data. This implies on the one hand that the data produced by third parties willing to deposit their data within the research infrastructure RI must be matched (i.e. identical entities such as companies, persons and securities appearing in multiple sources must be identified and linked). On the other hand, databases which cannot be integrated into the RI must be connected as well, and to do this Work Package 6 develops and tests innovative technologies to make databases interoperable. More specifically, the objectives of Work Package 6 are:

1. To develop and define the conceptual framework and technologies to match entities in the existing datasets and databases within the RI EURHISFIRM.
2. To test technologies to match national and cross-countries data.
3. To develop and define a conceptual framework and technologies for connecting data to other historical and contemporary databases deposited within other databases and infrastructures.
4. To test technologies to connect national and cross-countries data.

### II. Progress: Logistics and Resources

#### Human resources

The lead beneficiary of Work Package 6 is the University of Antwerp for which we have engaged two researchers. Johan Poukens who started working on Work Package 4 full-time since May 1, 2018 and continues to do this work for Work Package 6 as well. He was joined by a top expert in database development, Boris Cule who entered the project in October 2019. They work in close cooperation with Jailbreak (who took over Emmanuel Raviart's tasks when he left the project in the fall of 2019) and Jérémy Ducros based at PSE. The reason is obvious: Work Package 6 requires the combined input of the analysis of the content of other databases and the expert knowledge of a computer scientist. Boris Cule received his PhD in Computer Sciences at the University of Antwerp. Moreover, he was heavily involved in the past in developing the SCOB database. Jailbreak is an IT development company with extensive experience in data linking and warehouse building. Jérémy Ducros obtained a PhD in Economics from the *École des hautes études en sciences sociales* (EHESS) in Paris and he is collaborating with Work Package 4 with the aim to coordinate the specification production for Work Package 7. This team will collaborate for the accomplishment of Work Package 6 with Johan Poukens and Jérémy Ducros preparing and supporting the work of Boris Cule and Jailbreak.

#### Tools and technologies

Work Package 6 will work with databases and datasets in various data formats (e.g. simple tables and relational databases). For some tasks (e.g. task 6.1), data will be transformed to graph databases. In this case, the data will be managed with the open source Wikibase software.



Automated techniques with human supervision and verification will be used for record matching. The actual matching techniques will be determined after a comparison and evaluation during task 6.2 which is ongoing at the time of writing. Technologies for connecting data will equally be chosen during task 6.3.

### III. Progress: Project Achievements

#### Progress

Work Package 6 consists of four tasks which will result in two deliverable (reports) and two milestones:

Task	Output
6.1 Data Matching Design (M11-M23)	D6.1: Report on Data Matching Issues and Methodologies (M23, February 2020)
6.2 Test of Data Matching (M18-M24)	M6.1: Case study matching (M24, March 2020)
6.3 Data connecting design (M20-M29)	D6.2: Report on Data Connecting Issues and Methodologies (M29, August 2020)
6.4 Test of data connection (M26-M32)	M6.2: Case study connecting (M32, December 2020)

#### Task 6.1 Data Matching Design

Task 6.1 is currently completed and the deliverable has been submitted to the EC on schedule. Report D6.1 discusses the concept of data matching, as well as the issues that arise from attempting to match data from various sources within the EURHISFIRM project. Furthermore, it presents an overview of existing data matching methodologies, and discusses the extent to which they are, either directly or after suitable adaptations, applicable within our framework in terms of efficiency and accuracy.

#### Task 6.2 Test of Data Matching

In preparation of the data matching test (ongoing at the time of writing), a first meeting was held on February 11, 2019 with the participation of the IT expert Emmanuel Raviart, as well as with Boris Cule. At that meeting, it was decided to start the investigation and integration of the two most advanced databases within the project that will be linked and integrated. These are the Paris DFIH (Données Financières Historiques) and the Antwerp SCOB (Studiecentrum voor Onderneming en Beurs) databases respectively covering the Paris and Brussels stock exchanges.

The teams of Antwerp and Paris subsequently decided in their meeting of November 12, 2019 to run two tests on the same datasets in different data formats in parallel: the Antwerp team will match entities from the DFIH and SCOB databases in their native relational format (SQL), the Paris team will first transform both databases to a graph format (RDF/triples) before matching.

#### Task 6.3 Data connecting design

Like task 6.1, task 6.3 is a conceptual work that is currently still in a preparatory phase. A review of existing historical and contemporary datasets based on the data and sources inventory prepared during task 4.2 and the semantical analysis of task 4.3 is underway. For this task, the Antwerp team will also cooperate



closely with the partners involved in Work Package 5. SAFE (WU) in particular has extensive experience in identification keys (e.g. national company identifiers).

#### *Task 6.4 Test of data connection*

Candidate-databases for the test of data connection, such as the London Share Price Database (LSPD) and EUROFIDAI are already being investigated. Members of the Antwerp team have reached out to Mike Staunton of the London Business School (EURHISFIRM Project Advisory Board member) who has agreed to include the LSPD in Work Package 6 to test the matching and connecting of databases. Members of the Paris team have likewise been in contact with the CNRS about Eurofidai. Different scenarios for matching and connecting data from Eurofidai to consortium members' databases (i.e. SCOB and DFIH), from a basic exchange of identifiers to a complete merger, are being discussed. In preparation of the tasks within Work Package 6, the LSPD and Eurofidai have also been documented in D4.4 (see above). Again, the teams of Antwerp and Paris will divide most of the work amongst themselves, whereby the Antwerp team will work on connecting the Belgian and French data to the LSPD database and the Paris team to the Eurofidai.

## IV. Conclusion

From the first tasks (nearly) completed, it is evident that, the use of powerful matching algorithms notwithstanding, an amount of human interaction and effort is and will remain paramount for obtaining satisfactory results from matching historical data originating in multiple national datasets and databases. Incomplete data and language related issues (e.g. translation and transliteration of company names) are commonly encountered in these datasets and potential matches therefore require expert evaluation and verification before being promoted into the infrastructure database. This is an element that should be reckoned with in the design of the infrastructure's policy and architecture, as well as in the design of its data and metadata governance policy by Work Packages 9 and 5 respectively.





## Work Package 7: Data extraction and enrichment system

### I. Introduction

Work Package 7 develops an intelligent and collaborative system for the extraction of structured information from images of historical documents related to companies' financial and economic activities. The system focuses on printed serial sources related to listed companies such as stock exchange yearbooks and price lists. In the deliverable D7.1, we provided general software libraries, which can be used to build different prototypes of document recognition and understanding systems adapted to different kinds of documents. The deliverable D7.2 is composed of the first version of two recognition systems: one for yearbook information extraction, and one for price list data extraction. Those systems have been applied to several corpora: the German Yearbook 1913-1914 (Handbuch der deutschen Aktiengesellschaften), the French Desfossés Yearbook 1962, the official price lists for Brussels 1912 and Paris 1961-1962, which are part of the document samples dataset validated by the Steering Committee. This document samples dataset is made of three yearbooks, three stock price lists, with three different languages, on three time periods: before WWI, interwar and post WWII. The two remaining corpora are the Spanish yearbook 1929-1930 and the official price lists for Madrid 1934.

This report presents the work done in the Work Package 7 from April 1st 2019 to March 31st 2020. Details on the evaluation of the two developed systems have been presented in the deliverable D7.2.

### II. Progress: Logistics and Resources

The following people were recruited for Work Package 7, from April 1st 2019 to March 31st 2020:

Wassim Swaileh, Postdoc for Work Package 7, started on September 1st 2018, part-time (50%), LITIS, Université de Rouen Normandie;

Andres Rojas Camacho, Research Engineer for Work Package 7, started on September 15th 2018, part-time (50%) LITIS, Université de Rouen Normandie, since September 2019 he is full-time (100%) on Work Package 7.

Achille Fedioun, Research Engineer for Work Package 7, was recruited full-time (100%) on Work Package 7 starting on January 1st 2020.

Simon Bouvier, Research Engineer for Work Package 7, started on October 15th 2018, full-time (100%), IRISA, Insa de Rennes.

Iwan Le Floch, Research Engineer for Work Package 7, started on March 9th 2020, full-time (100%), IRISA, Insa de Rennes.

Jérémy Ducros, economic historian (replacing Elisa Grandi, who had started in April 2018), Paris School of Economics. Coordination and production of specifications for Work Package 7.



### III. Progress: Project Achievements

#### 1 Yearbook Information Extraction system

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

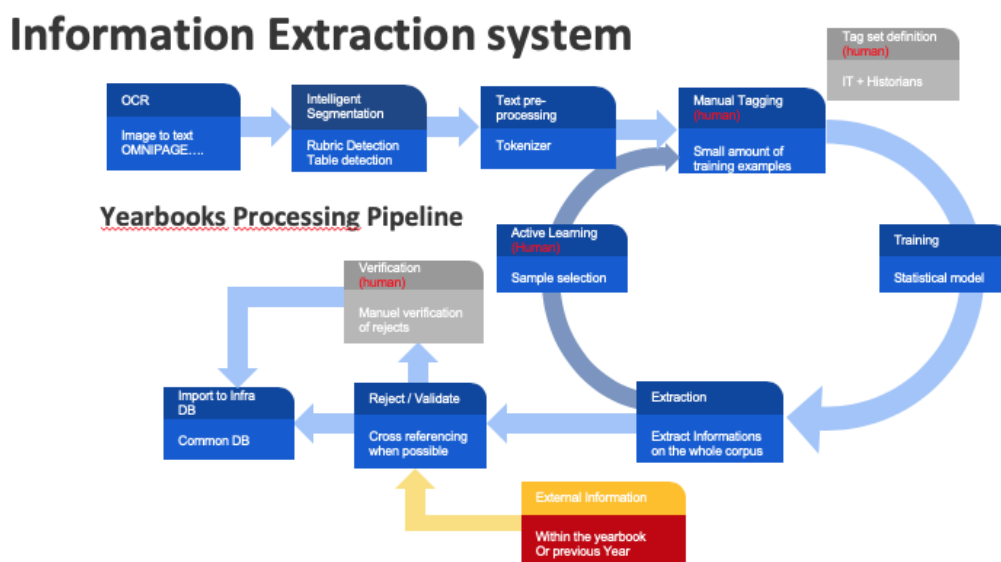


Figure 1: Overview of the generic Information extraction pipeline in yearbooks.

#### 1.1 Text Blocks Selection by Document Structure Recognition (Task 7.2)

##### 1.1.1 System

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

##### A. Component detectors and tools

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

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

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

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

1 68.000	168.000	168.000	168.000	1.680.000
185.926	193.498	193.185	206.310	2.165.769
165.744	173.781	201.789	204.162	2.340.984
15.120	15.120	18.345	*10.846	*204.076
584.790	550.399	581.819	589.318	6.390.829
87.143	41.576	49.404	53.709	520.879
19.217	22.760	34.332	46.214	484.145
17.926	18.816	15.072	7.495	36.733
480.998	442.579	482.976	512.369	5.382.336
85.866	89.004	83.271	79.454	971.760
584.790	550.399	581.819	589.318	6.390.829

Figure 2: Alignments can be used as separators between columns

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

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

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

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

~~Dividenden:~~ 1897: M. 17 pro Aktie p. r. t.; 1898—1913: 4, 4<sup>1</sup>/<sub>2</sub>, 5, 5<sup>1</sup>/<sub>2</sub>, 5<sup>1</sup>/<sub>2</sub>, 6, 6, 6, 6, 6<sup>1</sup>/<sub>2</sub>, 6<sup>1</sup>/<sub>2</sub>, 6, 5<sup>1</sup>/<sub>2</sub>, 5<sup>1</sup>/<sub>2</sub>, 6<sup>0</sup>/<sub>10</sub>. Coup.-Verj.: 4 J. (F.)

~~Vorstand:~~ Franz Kügelgen, Joh. Geusgen.

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

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

## B. Structural analysis of yearbooks

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

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

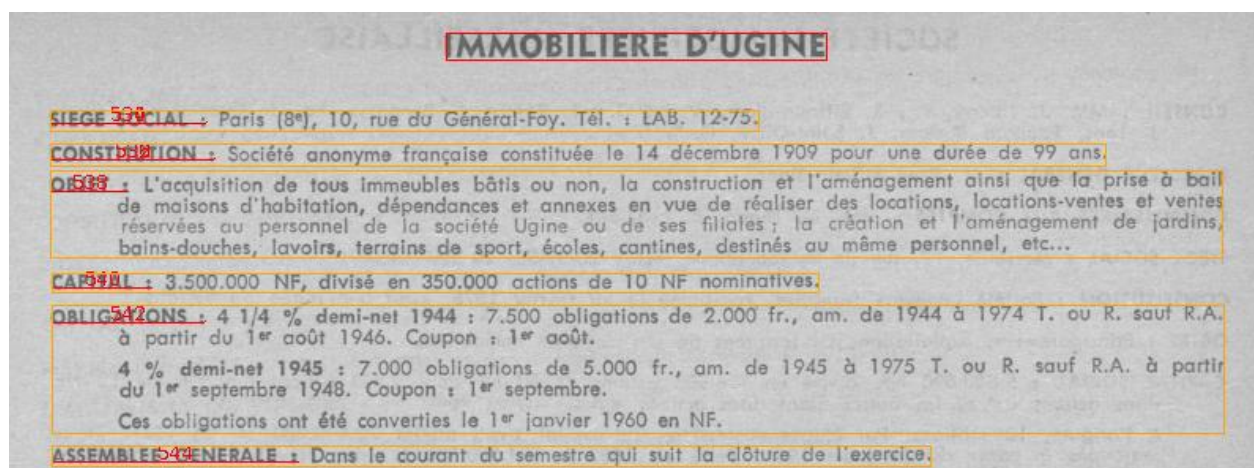


Figure 4: Rubrics detection example on the Desfossés yearbook (1962, tome 2, page 267)

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

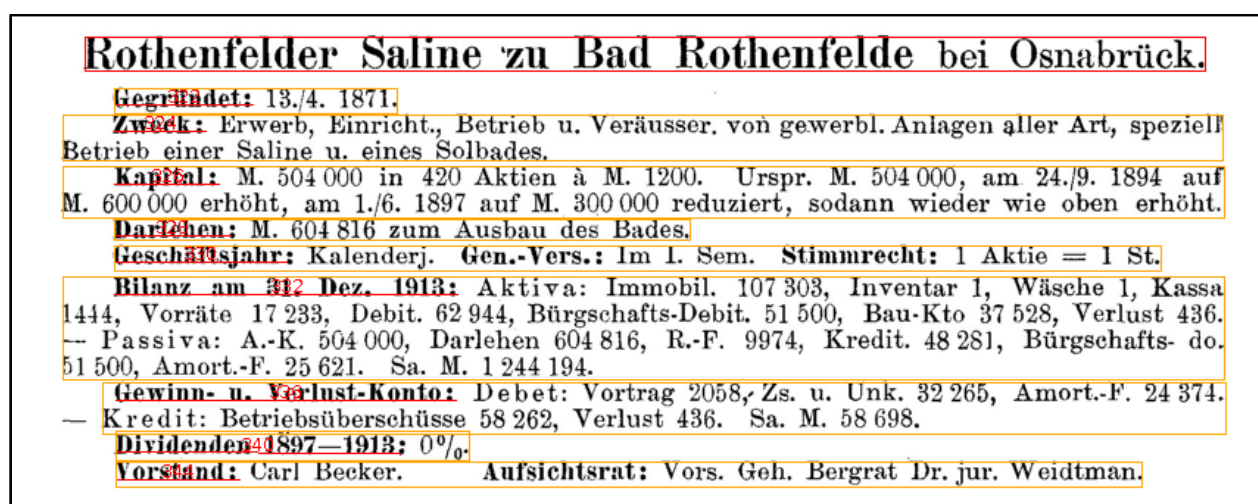


Figure 5: Rubrics detection example on the Handbuch yearbook (1914-15, page 951)



### C. The specific case of tables and balance sheets

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

	PRODUITS BRUTS	BÉNÉFICES NETS	RÉSERVES	REPORT A NOUVEAU	DIVIDENDE TOTAL	PAR ACTION	COURS EXTRÊMES DES ACTIONS	
			(En 1.000 francs)			(En francs)		
1956	6.088.542	378.682	114.000	517	248.372	500 net	17.750	13.240
1957	7.170.810	518.947	100.000	1.186	396.450	400 net	25.500	12.250
1958	8.430.411	634.516	150.000	1.454	455.823	450 net	16.450	11.930
1959	8.606.530	625.000	125.000	10.374	461.973	450 net	28.500	15.680
			(En nouveaux francs)					
1960	87.406.985	6.250.000	1.250.000	197.456	4.615.653	4,50 net	280,00	204,00
1961 (30 sept.)							289,60	210,00

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

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

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

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

	PRODUITS BRUTS	BÉNÉFICES NETS	RÉSERVES	REPORT A NOUVEAU	DIVIDENDE TOTAL	PAR ACTION	COURS EXTRÊMES DES ACTIONS	
			(En 1.000 francs)			(En francs)		
1956	6.088.542	378.682	114.000	517	248.372	500 net	17.750	13.240
1957	7.170.810	518.947	100.000	1.186	396.450	400 net	25.500	12.250
1958	8.430.411	634.516	150.000	1.454	455.823	450 net	16.450	11.930
1959	8.606.530	625.000	125.000	10.374	461.973	450 net	28.500	15.680
			(En nouveaux francs)					
1960	87.406.985	6.250.000	1.250.000	197.456	4.615.653	4,50 net	280,00	204,00
1961 (30 sept.)							289,60	210,00

Figure 7: Alignments detections in a table

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

BILANS A FIN FEVRIER		1957	1958	1959	1960	1961
ACTIF		(En 1.000 francs)			(En nouveaux francs)	
Immobilisations (nettes) .....		935.782	1.227.013	1.410.422	15.603.600	12.442.920
Autres valeurs immobilisées .....		732	886	6.917	83.326	125.203
Réalizable :						
Valeurs d'exploitation .....		1.395.754	1.968.875	2.376.917	15.242.855	41.998.609
Débiteurs .....		688.116	557.218	356.669	3.817.091	6.277.836
Titres de placement .....		172.041	104.011	97.888	1.204.126	1.092.650
Disponible .....		407.931	466.048	463.636	2.252.413	2.934.388
Résultats .....		»	»	»	»	3.114.266
PASSIF		3.600.356	4.324.051	4.712.449	38.203.411	67.985.872
Capital .....		520.000	510.000	510.000	5.100.000	5.100.000
Réserves .....		575.144	613.205	627.519	5.469.129	4.813.296
Fonds de renouvellement et provisions ..		396.197	600.900	597.429	6.364.083	9.059.571
Dette à long terme .....		22.867	22.111	21.321	190.005	178.527
Dette à court terme .....		1.824.383	2.313.580	2.700.131	20.929.655	48.834.478
Bénéfices .....	(1)	261.765	264.254	256.049	250.539	»
		3.600.356	4.324.051	4.712.449	38.203.411	67.985.872

(1) Avant impôt sur les Sociétés.

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

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

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

BILANS A FIN FEVRIER		1957	1958	1959	1960	1961
ACTIF		(En 1.000 francs)			(En nouveaux francs)	
Immobilisations (nettes) .....		935.782	1.227.013	1.410.422	15.603.600	12.442.920
Autres valeurs immobilisées .....		732	886	6.917	83.326	125.203
Réalizable :						
Valeurs d'exploitation .....		1.395.754	1.968.875	2.376.917	15.242.855	41.998.609
Débiteurs .....		688.116	557.218	356.669	3.817.091	6.277.836
Titres de placement .....		172.041	104.011	97.888	1.204.126	1.092.650
Disponible .....		407.931	466.048	463.636	2.252.413	2.934.388
Résultats .....		»	»	»	»	3.114.266
PASSIF		3.600.356	4.324.051	4.712.449	38.203.411	67.985.872
Capital .....		520.000	510.000	510.000	5.100.000	5.100.000
Réserves .....		575.144	613.205	627.519	5.469.129	4.813.296
Fonds de renouvellement et provisions ..		396.197	600.900	597.429	6.364.083	9.059.571
Dette à long terme .....		22.867	22.111	21.321	190.005	178.527
Dette à court terme .....		1.824.383	2.313.580	2.700.131	20.929.655	48.834.478
Bénéfices .....	(1)	261.765	264.254	256.049	250.539	»
		3.600.356	4.324.051	4.712.449	38.203.411	67.985.872

(1) Avant impôt sur les Sociétés.

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

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



BILANS AU 31 DECEMBRE	1956	1957	1958	1959	1960
<b>ACTIF</b>					
		(En 1.000 francs C.F.A.)			
Immobilisations (nettes) .....	313.978	291.509	272.901	300.672	303.359
Autres valeurs immobilisées .....	1.812	1.820	1.865	1.865	1.865
Réalisable :					
Valeurs d'exploitation .....	535.288	490.346	596.313	460.672	344.465
Débiteurs .....	229.208	350.011	341.206	1.251.556	601.911
Disponible .....	12.981	7.527	17.677	6.029	89.876
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476
<b>PASSIF</b>					
Capital .....	200.000	300.000	300.000	300.000	300.000
Réserves .....	102.040	122.153	202.152	307.255	70.643
Fonds de renouvellement et provisions	170.000	110.000	90.000	80.000	116.079
Dette à court terme .....	535.118	485.678	475.107	955.152	663.484
Bénéfices .....	86.109	123.377	162.703	378.387	191.270
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476

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

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

BILANS AU 31 DECEMBRE	1956	1957	1958	1959	1960
<b>ACTIF</b>					
		(En 1.000 francs C.F.A.)			
Immobilisations (nettes) .....	313.978	291.509	272.901	300.672	303.359
Autres valeurs immobilisées .....	1.812	1.820	1.865	1.865	1.865
Réalisable :					
Valeurs d'exploitation .....	535.288	490.346	596.313	460.672	344.465
Débiteurs .....	229.208	350.011	341.206	1.251.556	601.911
Disponible .....	12.981	7.527	17.677	6.029	89.876
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476
<b>PASSIF</b>					
Capital .....	200.000	300.000	300.000	300.000	300.000
Réserves .....	102.040	122.153	202.152	307.255	70.643
Fonds de renouvellement et provisions	170.000	110.000	90.000	80.000	116.079
Dette à court terme .....	535.118	485.678	475.107	955.152	663.484
Bénéfices .....	86.109	123.377	162.703	378.387	191.270
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476

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

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

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

Figure 12: Example of columns and currencies.

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

BILANS AU 31 DECEMBRE	1956	1957	1958	1959	1960
ACTIF		(En 1.000 francs C.F.A.)			
Immobilisations (nettes)	313.978	291.509	272.901	300.672	303.359
Autres valeurs immobilisées	1.812	1.820	1.865	1.865	1.865
Réalizable :					
Valeurs d'exploitation	535.288	490.346	596.313	460.672	344.465
Débiteurs	229.208	350.011	341.206	1.251.556	601.911
Disponible	12.981	7.527	17.677	6.029	89.876
PASSIF	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476
Capital	200.000	300.000	300.000	300.000	300.000
Réserves	102.040	122.153	202.152	307.255	70.643
Fonds de renouvellement et provisions	170.000	110.000	90.000	80.000	116.079
Dette à court terme	535.118	485.678	475.107	955.152	663.484
Bénéfices	86.109	123.377	162.703	378.387	191.270
	1.093.267	1.141.213	1.229.962	2.020.794	1.341.476

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

### 1.1.2 Evaluation

#### A. Evaluation process

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

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

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

#### B. Results

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



# BANQUE FRANÇAISE DU COMMERCE EXTERIEUR

**CONSEIL** : MM. G. Audouin, J. Barrot, R. Boyer, A. Coudré, P. Darnas, A. Develin, M. Duboulet, Maurice Laroche, A. Leger, J. Marlin, R. Maurin, Olivier Nouzeu-Nivet, G. Personnat, M. Plescoff, J. Poiny, Les nominations du Président et du Directeur général sont approuvées par arrêté du Ministre de l'Economie Nationale et du Ministre des Finances qui désignent également deux commissaires du Gouvernement auprès de la Société.

**COMMISSAIRES DU GOUVERNEMENT** : MM. J. Hirsch-Dantès, Directeur général au Ministère des Finances, représentant la Direction des Finances Extérieures, H. Baudouin, Chef du Service de l'Expansion Economique au Secrétariat d'Etat aux Affaires économiques, représentant la Direction des Relations Economiques Extérieures.

**DIRECTEUR** : M. J. Chénay, D.G.

**COMMISSAIRES AUX COMPTES** : MM. J. Boulenger, P. Simon.

**SIÈGE SOCIAL** : Paris (91), 21, boulevard Malesherbes, Tél. : POC 98-09.

**CONSTITUTION** : Société française, constituée le 30 septembre 1947 pour une durée de 99 ans, statuts approuvés par décret en Conseil d'Etat.

**OBJET** : La Banque Française du Commerce Extérieur a pour objet de faciliter le financement des opérations d'exportation ou d'importation avec l'Etranger, les départements et territoires d'Outre-mer, les Etats de la Communauté, soit directement, soit en collaboration avec les autres banques et organismes qui interviennent dans le financement du commerce extérieur.

**CAPITAL SOCIAL** : 15 millions de NF, divisé en 300.000 actions nominatives de 50 NF chacune, entièrement libérées, émises par les établissements suivants : Banque de France, Caisse des Dépôts et Consignations, Crédit National, Caisse Nationale de Crédit Agricole, Crédit Lyonnais, Société Générale pour favoriser le développement du Commerce et de l'Industrie en France, Comptoir National d'Escompte de Paris, Banque Nationale pour le Commerce et l'Industrie.

**ASSEMBLÉE GÉNÉRALE** : Dans les six premiers mois de l'année.

**REPARTITION DES BÉNÉFICES** : 5 % à la réserve légale, 5 % d'intérêt non cumulatif aux actions, sur l'excédent 10 % au personnel, 10 % aux autres associés de la Banque et aux œuvres de solidarité du personnel, le reliquat et suivant décision du Conseil pour réserves sur le nouveau solde subordonné à la décision du Conseil, 95 % à la disposition de l'Assemblée.

**SERVICE DES TITRES, PARLEMENT DES COUPONS** : Siège social.

**TRANSFERTS** : Siège social.

**COTATIONS** : Titres non cotés. — Notice SEF : BA 168.

**COUPONS NETS** : 120 mars 1956), 410 fr. (13 mars 1957), 320 fr. 80 (12 mars 1958), 380 fr. 90 (11 mars 1959), 390 fr. (10 mars 1960), 330 fr. (12 mars 1961), 340 NF.

RESERVES	PROG. BÉNÉF.	AMORTIS.	RÉSERV. NET	RESERVES	BÉNÉF. DISTRIB.	DIVIDENDES
—	—	—	En 1.000 francs	—	—	En francs
1956	?	?	439.865	381.223	200.554	450 »
1957	?	?	431.589	384.808	202.558	450 »
1958	1.617.694	38.971	165.231	392.361	250.575	500 »
1959	?	?	165.231	392.361	250.575	500 »
(En nouveaux francs)						
1960	?	?	4.115.949	1.350.000	2.580.549	500

BILANS AU 30 DÉCEMBRE	1956	1957	1958	1959	1960	En NF	
Passif							
A Capital	1.500.000	1.500.000	1.500.000	1.500.000	15.000.000		
B Réserves	1.170.000	1.415.949	1.465.969	1.540.929	75.751.465		
C Dettes	39.907.810	39.150.478	31.184.172	38.713.315	460.813.769		
D Dettes à court terme	92.615.247	107.505.786	138.135.373	145.153.020	231.648.671		
E Dettes à long terme	10.112.893	22.896.517	38.247.848	17.699.740	350.471.726		
F Dettes à court terme	997.505	515.744	501.834	795.504	6.865.701		
G Dettes à long terme	28.765.483	38.684.283	38.684.283	32.111.326	483.457.172		
H Dettes à court terme	8.119.714	2.085.950	6.451.190	10.754.222	95.456.021		
I Dettes à long terme	488.792	488.899	588.912	487.545	4.193.454		
J Dettes à court terme	1.137.870	1.363.744	1.515.558	1.812.840	92.491.488		
K Dettes à long terme	186.091.848	184.868.729	229.698.651	271.045.443	3.627.175.461		
Actif							
F Capital	243.802	254.128	460.113	514.714	6.578.703		
G Réserves	24.314	31.744	17.127	25.029	875.554		
H Dettes	7.525.847	8.393.875	9.405.497	12.609.795	164.893.547		
I Dettes à court terme	189.096.828	179.427.281	167.296.090	183.097.740	2.947.220.710		
J Dettes à long terme	28.655.864	35.883.304	22.392.106	31.692.015	448.544.846		
K Dettes à court terme	38.501.438	38.648.291	32.079.098	41.599.547	543.370.254		
L Dettes à long terme	563.106	776.114	684.818	816.729	15.256.444		
M Dettes à court terme	140.051.848	131.627.873	129.098.651	121.045.443	3.627.175.461		
N Dettes à long terme	—	—	—	—	—		

— 222 —

# COTONNIERE-FRANCO-CHADIENNE « COTONFRAN »

**CONSEIL** : MM. G. Bousquet, P.H. ; A. Aron, P. ; E. Van Gans, V.P. ; G. Boquet, A.O.G. ; S.A.I. Prince Louis-Hopfland, P. Bourcart, A. Chac, Char, Camille, Chac, Camille, H. Evreille, P. Gillereux, G. Gohr, A. Huet, M. Mahon, H. Doury, E. Senn, Calais de la Sablonnière des Rix du Coton du Tchad-Congo Afrique.

**ADMINISTRATEUR DIRECTEUR GÉNÉRAL** : M. G. Bouquet.

**SECRÉTAIRE GÉNÉRAL** : M. H. Gobelin.

**COMMISSAIRES AUX COMPTES** : M. G. Delpech, G. Goumy.

**SIÈGE SOCIAL** : Fort-Lamy (Tchad), Bureau à Paris (91) 9, avenue de Friedland, Tél. : BAL 30-50.

**CONSTITUTION** : Société anonyme constituée en 1926 pour une durée de 99 ans.

**OBJET** : Toutes opérations commerciales ou industrielles se rapportant aux plantes, produits végétaux et similaires, et notamment au coton.

**CAPITAL** : 742.500.000 fr. de C.F.A. en 199.000 actions de 3.750 fr. C.F.A. dont 178.250 actions A et 19.000 actions B.

A Targite, 500.000 fr. Par étapes successives, le capital net atteint 220 millions de fr. C.F.A. en 1954. Foré en 1955 à 300 millions par création de 44.000 actions gratuites de 2.500 fr. (1 par 2) en 1957 à 495 millions par création de 66.000 actions gratuites de 2.500 fr. (1 par 2) et 19.000 actions de 2.500 fr. C.F.A. B. Foré en 1958 à 742.500.000 fr. par élévation du nominal des actions de 2.500 fr. C.F.A. à 3.750 fr. C.F.A.

**ASSEMBLÉE GÉNÉRALE** : Dans les deux mois suivant la clôture de l'exercice.

**REPARTITION DES BÉNÉFICES** : 5 % à la réserve légale, constitution éventuelle de réserves : 5 % aux actions. Sur le surplus : 10 % au Conseil et la solde aux actions A et B.

**LIQUIDATION** : Exécution du passif, la solde aux actions.

**SERVICE FINANCIER ET TRANSPORTS** : Banque de l'Afrique Occidentale, Banque Union Française, Crédit Lyonnais, Caisse d'Epargne, B.N.C.I., Crédit du Nord.

**COTATIONS** : « Cote Diffusée ». — Actions 25. — Notice SEF : CO 217.

**COUPONS NETS** : 120 mars 1956), 410 fr. (13 mars 1957), 320 fr. 80 (12 mars 1958), 380 fr. 90 (11 mars 1959), 390 fr. (10 mars 1960), 330 fr. (12 mars 1961), 340 NF.

RESERVES	PROG. BÉNÉF.	AMORTIS.	RÉSERV. NET	RESERVES	BÉNÉF. DISTRIB.	DIVIDENDES
—	—	—	En 1.000 fr.	—	—	En francs
1956	84.697	82.500	82.500	41.500	360 »	5.880
1957	84.178	82.500	82.500	41.500	360 »	5.880
1958	86.008	79.700	107.517	81.669	387 50	5.558
1959	87.288	22.680	88.112	69.500	257 53	6.528

BILANS AU 30 DÉCEMBRE	1956	1957	1958	1959	1960
En 1.000 francs C.F.A.					
A Capital	88.606	495.000	743.500	743.500	743.500
B Réserves	617.777	643.152	801.378	375.911	572.000
C Dettes	973.516	915.568	685.000	971.378	641.553
D Dettes à court terme	615.600	517.000	618.800	682.000	446.500
E Dettes à long terme	357.916	398.568	66.200	289.378	195.053

BILANS AU 30 DÉCEMBRE	1956	1957	1958	1959	1960
En 1.000 francs C.F.A.					
A Capital	88.606	495.000	743.500	743.500	743.500
B Réserves	617.777	643.152	801.378	375.911	572.000
C Dettes	973.516	915.568	685.000	971.378	641.553
D Dettes à court terme	615.600	517.000	618.800	682.000	446.500
E Dettes à long terme	357.916	398.568	66.200	289.378	195.053
F Dettes à court terme	1.218.356	1.456.700	2.330.437	2.587.307	1.423.256
G Dettes à long terme	618.817	668.468	668.811	674.033	658.867
H Dettes à court terme	589.581	600.829	681.813	679.654	608.386
I Dettes à long terme	448.828	448.828	448.828	448.828	448.828
J Dettes à court terme	145.500	139.139	184.500	88.963	52.118
K Dettes à long terme	243.778	148.477	243.778	243.778	243.778
L Dettes à court terme	1.168.135	1.464.700	2.530.437	2.587.307	1.423.256

(\*) Bénéfice net.

— 550 —

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

Figure 15 presents examples of entire Handbuch page analysis.

**Krefelder Stahlwerk, Act.-Ges. in Krefeld-Fischeln.**  
Gegründet: 26./I. bzw. 5./4. 1900; eingetr. 29./5. 1900.  
Zweck: Errichtung u. Betrieb von Werken zur Herstellung von Stahl jeder Art. Die Ges. stellt als Spezialität hochwertige Qualitätsstähle her, welche in der Form von Roh-  
Handbuch der Deutschen Aktien-Gesellschaften 1914/1915 I. 51

42





shares with their amount. As these named entities are reported through a textual description and not placed into a table, a certain variability was introduced in phrasing the text at the time of publication. In addition, some information is sometimes partially missing. Figure 16 shows a case where the same tag set can be used for French and German. Notice that irrelevant words in the text are labelled with the tag "Other", as is the standard convention adopted for information extraction.

CAPITAL: 2.846.250 NF en 56.925 actions de 50 NF A l'origine 300.000 fr. Par étapes successives le capital avait atteint 6.900.000 fr. en 1943. Transformé en piastres en 1946 et porté à 1.035.000 piastres par création de 34.500 actions nouvelles de 10 piastres réparties gratuitement (1 pour 2). Porté en 1950 à 6.210.000 piastres par élévation du nominal à 60 piastres; en 1952 à 7.762.500 piastres par élévation du nominal de 60 à 75 piastres puis titres regroupés en 150 piastres à partir du 19 janvier 1953. Porté en 1954 à 15.525.000 piastres par élévation du nominal à 300 piastres; en 1955 à 28.462.500 piastres par élévation du nominal à 500 piastres et création de 5.175 actions gratuites de 500 piastres (1 pour 10), puis capital transformé en 1956 en 284.625.000 francs. Converti le 1<sup>er</sup> janvier 1960 en 2.846.250 NF

**Capital:** M. 4 000 000 in 4000 Aktien à M. 1000 Urspr. M. 1 000 000, erhöht zur Verstärkung der Betriebsmittel lt. G.-V. v. 21.10. 1909 in M. 500 000 mit Div.-Ber. ab 1./1. 1910, begeben an die alten Aktionäre zu 130% franko Zs. Agio mit M. 125 070 in R.-F. Mit Rücksicht auf die stetige Entwickl. u. den erheblich gesteigerten Auftragsbestand der Ges. beschloss die a.o. G.-V. v. 27.5. 1911 weitere Erhö. um M. 500 000 in 500 Aktien mit Div.-Ber. ab 1./7. 1911, übernommen von G. Fromberg & Co. zu 200%, angeboten den alten Aktionären v. 14./6. – 27./6. 1911 zu 220%. Agio mit M. 500 000 in R.-F. Nochmals erhöht lt. G.-V. v. 16.2. 1912 um M. 2 000 000 (auf M. 4 000 000) in 2000 Aktien mit Div.-Ber. ab 1./7. 1912, übernommen von einem Konsort. (G. Fromberg & Co. etc.) zu 200% franko Zs. zuzügl. aller Kosten bis zum Betrage von M. 170 000, angeboten den alten Aktionären im Febr.-März 1912 zu 220%. Agio mit M. 2 000 000 in R.-F.

**Legend:**  
 Last amount  
 Share amount  
 Nb shares  
 Chg-date  
 Chg-amount  
 Init-amount  
 Currency  
 Ini-date

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

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

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

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

CAPITAL: 2.846.250 NF en 56.925 actions de 50 NF A l'origine, 300.000 fr. Par étapes successives le capital avait atteint 6.900.000 fr. en 1943. Transformé en piastres en 1946 et porté à 1.035.000 piastres par création de 34.500 actions nouvelles de 10 piastres réparties gratuitement (1 pour 2).

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

### B. Statistical models for information extraction

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

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

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

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

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

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

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

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

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

where  $\vec{h}_t$  represents a view of the LSTM of the past sequence of characters of character  $x_t$  while  $\overleftarrow{h}_t$  represents the model parameters. Similarly, a backward language model computes the probability in the reverse direction as follows:

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

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

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

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

### C. Active learning scheme

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

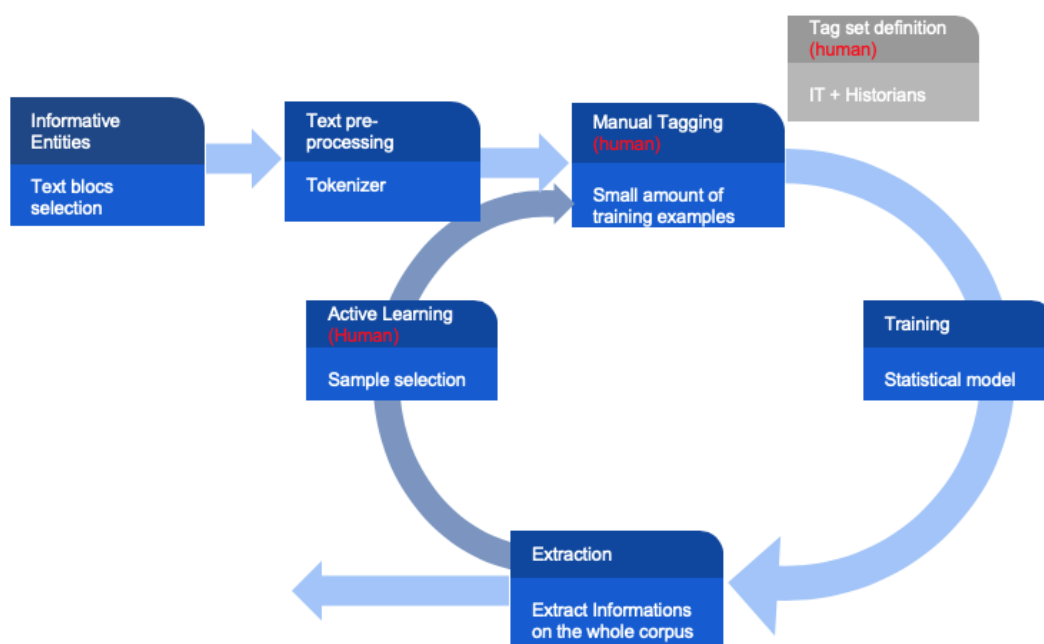


Figure 18: General architecture of the named entity extraction system, with human interaction in an active learning loop.

### 1.2.2 Evaluation

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

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

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

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

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

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

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

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

## Entity linking

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

We experimented two methods for linking the entities into chunks. The minimum distance method regroups entities with their closest neighbour entity. Using the link tag, we are able to learn how to link the entities. We then consider a sequence of linked entities, as entities of the same chunk. The two methods have been evaluated on the CAPITAL rubric using the CRF model, see deliverable D7.2 - table 5. The results show better performance when using the learned tag link.



## Active learning

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

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

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

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

### 1.2.3 Specification of tags for each rubric

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

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

This process takes many iterations, mostly because of the size a yearbook and the quantity of pages. This means that even if a specification document was done, there have always been some special cases that were hard to find in the yearbook and so extra consultation was needed in order to know how to proceed with these new cases.

The specification documents for the two Yearbooks have been written:





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

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

#### 1.2.4 Labelled Datasets

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

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

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



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

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

#### 1.2.5 Annotation Interface

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

#### Search and Filter specific documents

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

### Find a document

#### by Collection

1. FRYB

#### by Category

FRYB ▾

Administrators ▾

Not validated● Validated●

Search

#### by Status

FRYB ▾

Not validated● Validated●

Search

*Figure 20: Search and Filter feature of the interface*

### Visualisation of a document

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

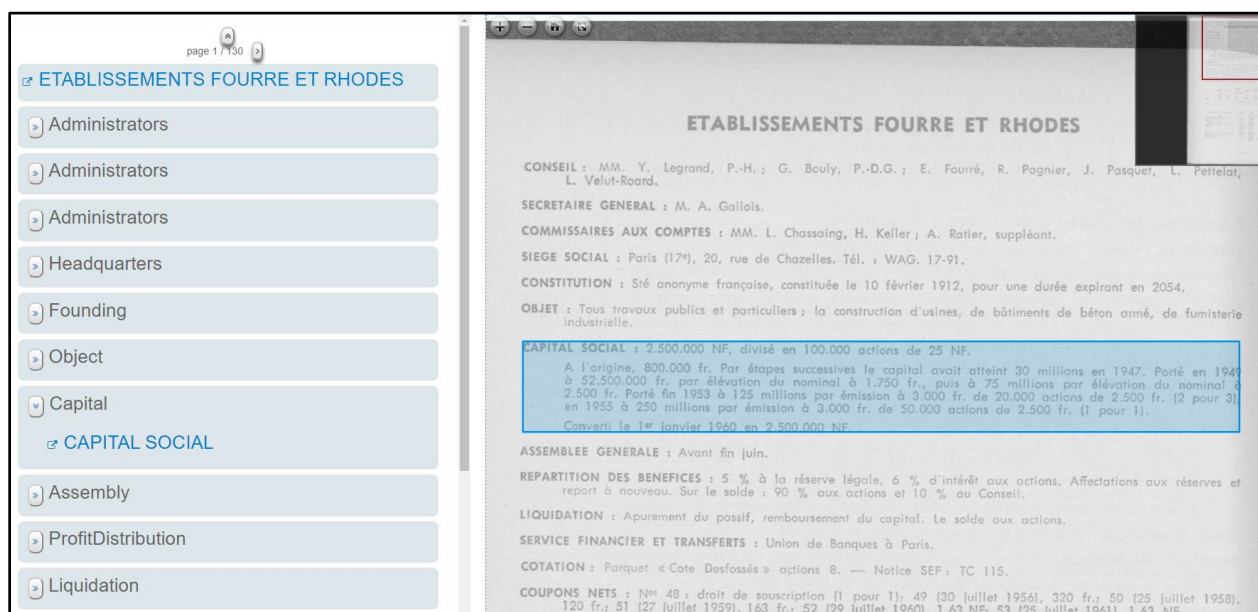


Figure 21: Visualisation of the sections of a document in English

### Visualisation of a textual section

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

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

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

CAPITAL SOCIAL : 2.500.000 NF, divisé en 100.000 actions de 25 NF.

A l'origine, 800.000 fr. Par étapes successives le capital avait atteint 30 millions en 1947. Porté en 1949 à 52.500.000 fr. par élévation du nominal à 1.750 fr., puis à 75 millions par élévation du nominal à 2.500 fr. Porté fin 1953 à 125 millions par émission à 3.000 fr. de 20.000 actions de 2.500 fr. (2 pour 3); en 1955 à 250 millions par émission à 3.000 fr. de 50.000 actions de 2.500 fr. (1 pour 1). Converti le 1<sup>er</sup> janvier 1960 en 2.500.000 NF.

You must be signed in to annotate.  
Click on a tag to select it and apply it on the tokens.

Capital (Category)

link currency ini-date chg-date last-date ini-amount chg-amount last-amount O

Validated ☐

CAPITAL SOCIAL

Validated ☐

: 2.500.000 NF divisé en 100.000 actions de 25 NF. A l'origine, 800.000 fr. Par étapes successives le capital avait atteint 30 millions en 1947. Porté en 1949 à 52.500.000 fr. par élévation du nominal à 1.750 fr., puis à 75 millions par élévation du nominal à 2.500 fr. Porté fin 1953 à 125 millions par émission à 3.000 fr. de 20.000 actions de 2.500 fr. (2 pour 3); en 1955 à 250 millions par émission à 3.000 fr. de 50.000 actions de 2.500 fr. (1 pour 1). Converti le 1 janvier 1960 en 2.500.000 NF.

Type	Date	Amount	Currency
Last		2.500.000	NF
Initial	A l'origine	800.000	fr
Change	1947	30 millions	
Change	1949	52.500.000	fr
Change		75 millions	
Change	1953	125 millions	
Change	1955		
Change		250 millions	
Change	1 janvier 1960	2.500.000	NF


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

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

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

1) link currency ini-date chg-date last-date ini-amount chg-amount last-amount O

2) en 1949 à 52.500.000 fr. pa  
, puis à 75 millions par él  
é fin 1953 à 125 millions par

3) 

4)

Type	Date	Amount	Currency
Last		2.500.000	NF
Initial	A l' origine	800.000	fr
Change	1947	30 millions	
Change	1949	52.500.000	fr
Change		75 millions	
Change	1953	125 millions	
Change	1955	250 millions	
Change	1 janvier 1960	2.500.000	NF

Figure 23: Use of the correction interface

### Visualisation of a tabular section

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

BILANS AU 31 DECEMBRE		1956	1957	1958	1959	1960
ACTIF		(En 1.000 francs)				
Immobilisations (nettes) .....		194.071	175.455	177.706	188.965	1.856.500
Autres valeurs immobilisées .....		19.657	17.222	22.454	26.637	320.555
Réalizable :						
Valeurs d'exploitation .....		323.700	150.823	148.077	1.189.060	11.865.016
Débiteurs .....		289.162	435.129	350.643	358.929	3.552.724
Titres de placement .....		3.583	3.644	3.606	3.001	35.685
Disponible .....		33.403	27.731	108.710	245.616	3.381.985

You must be signed in to annotate.  
Click on a tag to select it and apply it on the tokens.  
BalanceSheet (Category)  
part assets den year amount currency part liabilities part unknown O

Validated ☐

Immobilisations (nettes)

Validated ☐

ACTIF Immobilisations (nettes) 1956 194.074 En 1.000 francs

Validated ☐

ACTIF Immobilisations (nettes) 1957 175.455 En 1.000 francs

Validated ☐

ACTIF Immobilisations (nettes) 1958 177.706 En 1.000 francs

Validated ☐

ACTIF Immobilisations (nettes) 1959 188.965 NF

Validated ☐

ACTIF Immobilisations (nettes) 1960 1.856.500 NF

Unknown	Liabilities	Assets	Year	Item	Amount	Currency
		ACTIF	1956	Immobilisations (nettes)	194.074	En 1.000 francs
		ACTIF	1957	Immobilisations (nettes)	175.455	En 1.000 francs
		ACTIF	1958	Immobilisations (nettes)	177.706	En 1.000 francs
		ACTIF	1959	Immobilisations (nettes)	188.965	NF
		ACTIF	1960	Immobilisations (nettes)	1.856.500	NF

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

### Towards a more friendly interface

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

## 2. Price list data extraction system

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

### 2.1 Document Structure Recognition (Task 7.2)

#### 2.1.1 System

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

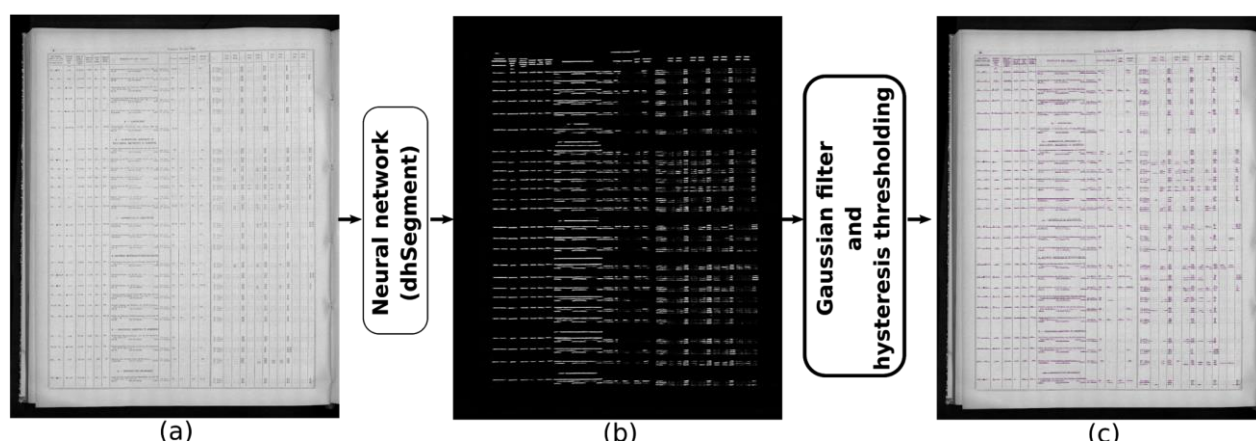


Figure 25: Example of text line extraction in a stock price lists document. (a) Image to be processed - (b) Probability map for text-line in this image - (c) Extracted text-lines

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



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

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

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

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

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

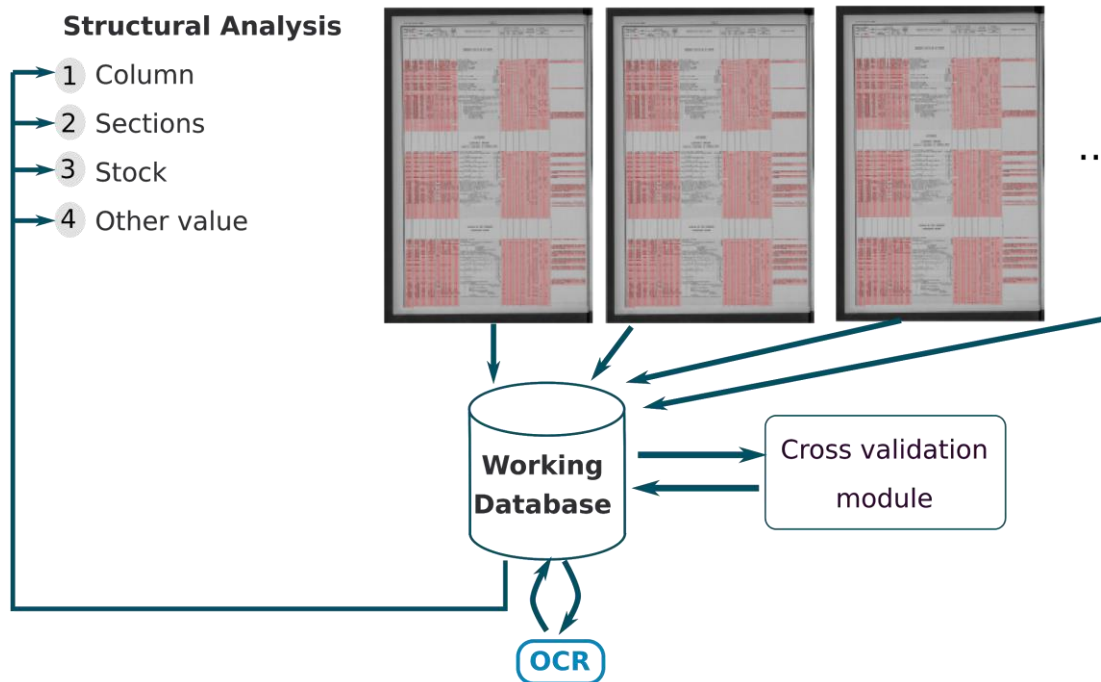


Figure 28: Overview of the global strategy

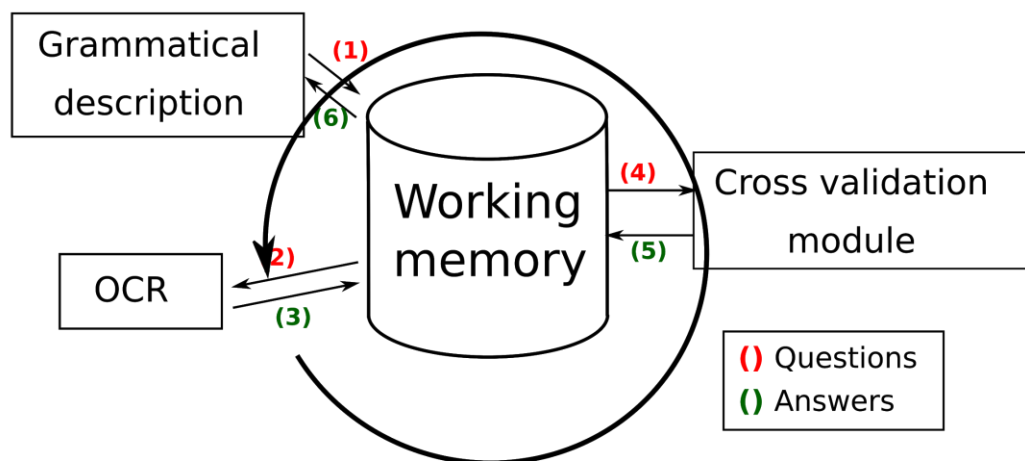


Figure 29: Interaction between the modules of the strategy

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

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

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

### 2.1.2 Evaluation

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

#### A. Evaluation on “La Coulisse”

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





Figure 30(a) displays two side-by-side images of a financial table. The left image shows a table with a missing column on the right side. The right image shows the same table with the missing column restored by the strategy.

(a)

Figure 30(b) displays two side-by-side images of a financial table. The left image shows a table with two missing columns on the right side. The right image shows the same table with the missing columns restored by the strategy.

(b)

Figure 30: Qualitative results: Improvement obtained with our strategy

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

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



## B. First evaluation on Paris and Brussels official lists

(a)

(b)

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

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

## 2.2 General-purpose text recognizer (OCR) (Task 7.3)

### 2.2.1 System

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

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

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

### 2.2.2 First instance of the system

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

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

### 2.2.3 Towards a more generic system

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

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

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

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

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



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

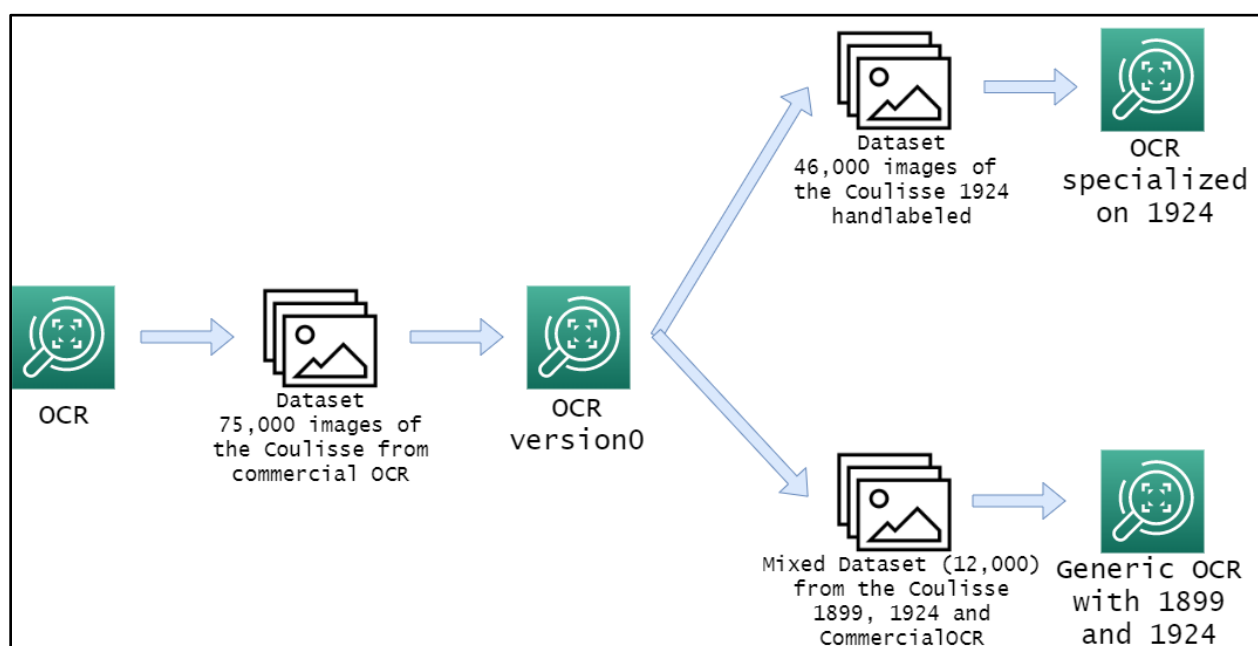


Figure 32: Learning strategy for the OCR with the new datasets

### Size of data

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

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

### Conclusion

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

#### 2.2.4 Annotation Interface

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

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

- ET TAVERNES ZIMMER §
- - - Jouissance
CAFÉ-RESTAURANT AMÉRICAIN
DOCKS RÉMOIS (Compt. Gén.Alim. et App.) Unités.
Eco (Société Technique Alimen.), série A §
- - série B §
*ERIDANIA (Société Industrielle) Unités§

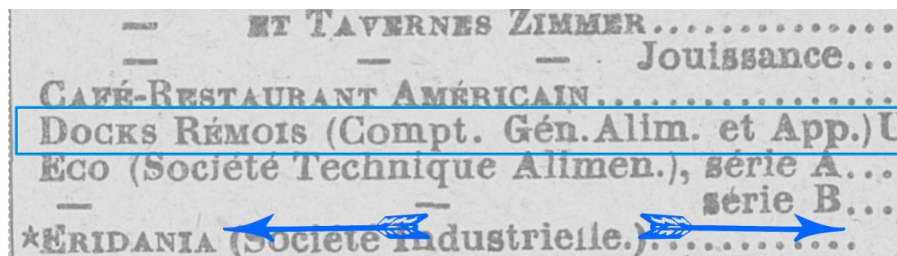


Figure 33: The correction interface for hand labelling data

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

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



Figure 34: Day after and day before feature

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



#### IV. Deliverables and Milestones

The work presented in sections 3 and 4 corresponds to the deliverable D7.2 (First version of the data extraction system), which contains more detailed results, and to the milestone M7.1 at M24.

#### V. Conclusion

After building the library of document components detectors for structure recognition and the general-purpose text recognizer (OCR) during the first year, we have built on top of them a first version of the data extraction system on yearbooks (section 3) and on price lists (section 4). These two prototypes have been designed on the French Desfossés Yearbook 1962 and the German Yearbook 1913-1914 (*Handbuch der deutschen Aktiengesellschaften*) for the yearbook extraction system and on the official price lists for Brussels 1912 and Paris 1961-1962 for the price lists extraction system prototype.

We work on the Spanish yearbook 1929-1930 by building a more generic extraction system for yearbooks, combined with a new design of the interface for the active learning process, which both improve the ability of the system to be applied to a new yearbook. On official price lists, we are working on the cross-validation module for the stock prices and on an interface for expert's interaction with the extraction system. We are currently working on the Brussels and Paris official price lists. We will then apply and adapt the system to the official price lists for Madrid 1934. In the same way as for yearbooks, the extraction system for price lists will be improved with a generic approach. Web-linking of some extracted information will be developed mainly on the French documents. We will also continue to evaluate the two systems on a larger dataset extracted from the sample dataset: Belgian, French, German, and Spanish documents.

#### VI. References

- [Akbik 2018] Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: *Proceedings of the 27th International Conference on Computational Linguistics*. pp. 1638-1649 (2018)
- [Ares Oliveira 2018] S. Ares Oliveira, B. Seguin, and F. Kaplan, "dhSegment: A generic deep-learning approach for document segmentation," in *Frontiers in Handwriting Recognition (ICFHR)*, 2018 16th International Conference on, pp. 7-12, IEEE, 2018.
- [Chazalon 2011] J. Chazalon, B. Couasnon, and A. Lemaitre. Iterative analysis of pages in document collections for efficient user interaction. In *International Conference on Document Analysis and Recognition*, 2011
- [Diem 2017] Diem, M., Kleber, F., Fiel, S., Grüning, T., & Gatos, B. (2017, November). cbad: Icdar2017 competition on baseline detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 1355-1360). IEEE.
- [Galibert 2014] O. Galibert, J. Kahn, I. Oparin. (2014, October). The ZoneMap metric for page segmentation and area classification in scanned documents. IEEE.
- [Grover 2008] Grover, C., Givon, S., Tobin, R., Ball, J.: Named entity recognition for digitised historical texts. In: *LREC* (2008)



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

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

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

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

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

[Swaileh 2020] W. Swaileh, T. Paquet, Sebastien Adam<sup>1</sup>, and Andres Rojas Camacho<sup>1</sup>, A Named Entity Extraction System for Historical Financial Data, Document Analysis Systems (DAS), accepted, 2020.



## Work Package 8: Interaction with users

### I. Introduction

The aim of Work Package 8 was to determine the optimal design of the data and services that EURHISFIRM RI should provide, by gathering and analysing the preferences of potential end-users and key stakeholders (academics, practitioners, regulators etc.). In order to do this, the specific objectives of Work Package 8 were:

- ▶ To develop a large-scale survey, via an online questionnaire, on stakeholders' perspectives and preferences for the design of EURHISFIRM;
- ▶ To conduct the survey and analyse the results;
- ▶ To identify qualified persons and conduct semi-structured interviews on their perspectives and preferences for the design of EURHISFIRM;
- ▶ To produce recommendations that will guide the design and data policy of the EURHISFIRM RI.

We have completed all four objectives.

### II. Progress: Logistics and Resources

There have been no additions to the resources of Work Package 8 since the last annual report.

### III. Progress: Project Achievements

The first deliverable of this Work Package (D8.1) was to develop a large-scale survey, via an online questionnaire, in order to identify the preferences of potential end-users and key stakeholders of the EURHISFIRM project. This was completed and submitted on 31 August 2018 and the survey can be found at <http://www.eurhisfirmsurvey.eu/>.

The second deliverable of this Work Package (D8.2) was to conduct the survey and analyse the results. This was completed and submitted on 31 January 2019.

The third deliverable of this Work Package (D8.3) involved identifying and interviewing stakeholders and potential end-users in order to clarify the findings of D8.2 and to produce recommendations for the design of the EURHISFIRM RI. This was completed and submitted on 26 August 2019.

The fourth deliverable of this Work Package (D.8.4) was to make specific recommendations for the optimal design of the data and services that the EURHISFIRM RI should provide. This was completed and submitted on 30 October 2019, along with a summary in Milestone 8.1. Both documents were published on <https://eurhisfirm.eu> and uploaded to EURHISFIRM's repository on OpenAIRE.



#### IV. Conclusion

Analysis of the online questionnaire (D8.2) and qualitative interviews (D8.3) yielded valuable insights relating to the content and usability desired by potential end-users and stakeholders of the EURHISFIRM RI. Based on this research, we recommend that resources should be focussed on data relating to the twentieth century, as this was the most popular time period among our respondents. Regarding forms of company data, we recommend that resources should be focused on ordinary equity market data because it was by far the most popular with respondents. There was also a desire expressed for accounting data, specifically, total assets, total debt, revenues, and profits. Respondents were less keen on data related to government and corporate bonds. In terms of frequency of share prices, we recommend daily and monthly data. As regards geography, the United Kingdom was the most popular country, followed by Germany and France. Turning to usability, we recommend that the EURHISFIRM platform should as far as possible enable users to manipulate the data themselves. Specifically, users should be able to download the data in bulk, in MS Excel format, and with minimum restrictions on downloads per day. We recommend that the EURHISFIRM platform use Wharton Research Data Services as an example of best practice in this area. However, for less popular data, we recommend that EURHISFIRM provide simple, non-tabulated PDF scans of the original source document. For reassurance as to the accuracy of data, users should be able to 'click through' to a scan of the original document and EURHISFIRM should provide a full citation of any source material. We also recommend that EURHISFIRM provide an explanation of the methodology and rationale for any interpretation or manipulation of data carried out by EURHISFIRM researchers. Finally, regarding feedback, an email address would be sufficient.



## Work Package 9: Infrastructure policy and architecture

### I. Introduction

Work Package 9 designs the architecture and the operation of the RI, with regard to access, security, support and maintenance, in cooperation with ESFRI Landmark CESSDA. Users' preferences on data and services design guide the platform's architecture and operating. Accordingly, the security system, the maintenance and the desk management of the platform are designed and estimated. The platform's architecture and operating are made coherent through the National Focus Points and site's policies. The Work Package also assesses the optimal level of integration of EURHISFIRM with existing RIs such as CESSDA and DARIAH. Following the user requirements' specification and RI policies, the users' development unit is designed. This unit support users in accessing data and services. More specifically, Work Package 9 aims to design:

- ▶ The RI national and European policies taking into account user requirements' specifications to design a federated and distributed RI;
- ▶ The technical architecture functionality of the RI;
- ▶ The system, database and network administration of the RI;
- ▶ The users project development unit of the RI

### II. Progress: Logistics and Resources

WP9 has started in part-time sooner than planned in the proposal, to allow for building relationships with and receiving input from stakeholders, like CESSDA and DARIAH. It has been proven to be useful to take into consideration the overall architecture of the RI right from the start of the project. The earlier start will not lead to overly depletion of the WP9 budget.

The structure of the WP9 deliverables will be slightly different compared to the description in the project proposal (page 89 in the final version). The proposal mentions a "conceptual design report" with deliverables D9.1 and D9.4, where at the same time D9.4 is intended to be a report that describes technical aspects of the RI. By carefully comparing the WP9 task descriptions and report descriptions in the proposal, we've proposed to setup this structure for the WP9 deliverables:

- ▶▶ D9.1 Report on RI policy specifies direct and indirect functional requirements for a future system implementation of EURHISFIRM, based on input from the proposal and reports from other Work Packages;
- ▶▶ D9.3 and D9.4 reports will be combined to a single report that describes the business architecture, functional architecture and technical architecture aspects of the RI (not including WP7 infrastructure). This structure complies to the open standard of The Open Group Architecture Framework (TOGAF). This report will be aligned with the D9.1 report.
- ▶▶ D9.2 report will follow the setup of the proposal.



First deliverable (Report on RI policy as part of the conceptual design report) due in M27 is finalized.

Other steps taken so far to ensure logistical and operational progress include:

- ▶ A new hire for the project: Joanna Kinga Sławatyniec started as a research fellow for the project at Erasmus Rotterdam University (simultaneous hire for Work Package 9 and Work Package 11, and help for Work Package 4 on UK data knowledge). She joined on December 1, 2018.
- ▶ General meetings with all group members: To assure transparency and an awareness of EURHISFIRM goals, aims and deadlines face to face catch up meetings have taken place
  - ▶▶ The overarching aim of the meetings is to ensure that all the project requirements are being met and EURHISFIRM deadlines will be fulfilled.
- ▶ In addition to the face to face meetings with all team members, regular Skype sessions take place as well as other meetings, including:
  - ▶▶ Monthly catch-up session between Joanna Kinga Sławatyniec and Coen Fierst van Wijnandsbergen
  - ▶▶ Bi-weekly Skype sessions between Joanna Kinga Sławatyniec and Joost Jonker
  - ▶▶ Occasional Skype calls between all team members
  - ▶▶ A meeting covering technical aspects of the project (time writing, etc.) between Joanna Kinga Sławatyniec, Marlies Vreeswijk, and Juan de Weger that took place on mid-December in Rotterdam.
- ▶ Lastly, there has also been regular participation in WGIS Zoom meetings on behalf on the team members.

### III. Progress: Project Achievements

As discussed above, Work Package 9 is progressing steadily. First deliverable has been finalized. Everything is on track keeping in mind the Work Package's milestones and deliverables.

### IV. Conclusion

Over the coming months, work will be carried out in preparation of the remaining milestones and deliverables. To ensure the milestones are reached and deliverables are submitted on time, team members will collaborate with Coen Fierst van Wijnandsbergen, who has been leading Work Package 9.





## Work Package 10: Business model and governance

### I. Introduction

The objective of Work Package 10 is to develop a business and governance model that contributes to the articulation of the EURHISFIRM's platform design (jointly with Work Package 8-Interaction with users, Work Package 9-Infrastructure policy and architecture, and Work Package 11-Cultural heritage). The governance model will be developed in constant interaction with the community of EURHISFIRM's stakeholders to ensure a balanced composition of its governing and supervisory bodies. The business model will focus on different users' profiles to define the modalities of their access to EURHISFIRM's data and services (taking into account issues such as data ownership and intellectual property rights). It will also design a revenue model that will guarantee the long-term sustainability of EURHISFIRM. The sequence of tasks to be achieved by Work Package 10: T1) definition of alternative business model concepts; T2) preliminary assessment of business and governance model alternatives; T3) assessment of stakeholders' preferences and feedback from experts; T4) detailed business and governance model design. They will lead to three deliverables: D1) a preliminary report on business model and governance assessment (M24); D2) a report on preferences expressed by stakeholders and qualified experts (M29); D3) a final report on business model and governance as part of the conceptual design report (M36).

### II. Progress: Logistics and Resources

In September 2019 we finalized the recruitment of strategic human resources necessary for the development of WP10. Starting on October 1<sup>st</sup>, 2019, we hired Jesús Martínez Casadiego, a telecom engineer with more than 10-year professional experience in project management and the coordination of international partners and clients in the consulting company Everis (<https://www.everis.com/global/en>). Jesús is also attending a MOOC Master programme in Data Science, which enhances his full understanding of EURHISFIRM's objectives and challenges. Since October 2019 we have worked at Tasks 10.1 (Survey of alternative business model and governance concepts) and 10.2 (Preliminary assessment of business model and governance concept alternatives), in preparation of Deliverable D10.1 (Preliminary Report on Business and Governance Model Assessment) and related Milestone M10.1 (M10.1 Preliminary business and governance model design).

### III. Progress: Project Achievements

EURHISFIRM's business and governance model will draw on best-practices as developed by existing RIs. Our main reference are the OECD Principles and Guidelines for Access to Research Data from Public Funding, published in 2007. More specifically, we draw from the two reports published in December 2017 by the OECD Global Science Forum and its partners in the framework of the Open Data for Science project. These identify the development of a sustainable business model for research data as a high priority, and provide a systematic analysis of income streams, costs, value propositions and business models for data repositories, based on structured interviews with repository managers from 18 countries



and a broad range of research areas. The studies provide a comprehensive summary of key issues, which include: how existing data repositories currently are funded, what are their key revenue sources, which additional innovative revenue sources are being explored, how revenue sources fit together into sustainable business models, what incentives and means for cost optimization are available, and which revenue sources and business models are most acceptable to stakeholders. The experience of existing RIs suggest that data ingest & curation, and system development & maintenance are usually the largest cost drivers. In turn, technological development and automation, learning by staff & users, and shared services for administration and management are identified as the main sources of cost optimization. In terms of revenue sources, structural funding (long-term contract with a research or infrastructure funder, usually a public one) and host-institution funding (from a research performing host institution, such as universities or research centres) are prominent in the business models of existing RIs. In turn, the diversification of resources is generally limited, as a majority of RIs report funding from one source only.

Deliverable D10.1 (Preliminary Report on Business and Governance Model Assessment) was completed and submitted to the approval of the Executive Committee of EURHISFIRM on March 26, 2020. A Business Model is a plan for the successful operation of a business by identifying sources of revenue, the intended stakeholder base, products, and funding details. The development of a Value Proposition for the RI is a necessary pre-requisite for the formulation of a business model. For that purpose, the report focuses on the identification of stakeholders, the development of services, the identification of revenue sources and the demonstration of value. The report assesses alternative business and governance models to be implemented when EURHISFIRM will enter into its operational stage. In the process of model selection, we follow CESSDA (Consortium of European Social Science Data Archives) 's approach, based on an adaptation of the BMC (Business Model Canvas) model, and use CESSDA's "Archive Development Canvas" (from CESSDA SaW – Cost-Benefit Advocacy Toolkit) in order to approximate EURHISFIRM's business model. The report compares different models (public RI, public RI with value-added services, public-private RI, private RI) with a special focus on their different revenue sources. Special attention is paid to the implications of EURHISFIRM's adherence to Open Science principles, and its integration with EOSC (European Open Source Cloud) and SSHOC (Social Sciences and Humanities Open Cloud). An important aspect of the report is a comprehensive mapping of stakeholders' profiles, role and influence, as a preliminary step in the launching of a systematic surveying of stakeholders. They will provide a critical feedback that will inform the final choice of a business and governance model. The report also analyses the range of possible services that EURHISFIRM could provide, maps their potential beneficiaries and identifies the related revenue streams. Finally, the report discusses a set of indicators that approximate the economic benefits of EURHISFIRM (centred on the cost of not-having EURHISFIRM as a FAIR research infrastructure) as well as the socioeconomic benefits in terms of scientific and socio-economic impact, tailoring to the case of EURHISFIRM the best practice established by the European Commission and the OECD. Finally, the report discusses the alternatives available in terms of legal forms (with a special attention to the possibility to operate as a service-providing ERIC within CESSDA) and reviews the governance structure of existing RIs.

In preparation of future deliverable and milestones, we started working at Task 10.3 (Characterizing the stakeholders) and Deliverable D10.2 (Report on preferences expressed by stakeholders and qualified experts) by mapping EURHISFIRM's potential stakeholders with the objective of preparing personalized



surveys, tailored around the different stakeholder profiles to be contacted. Surveys will be sent to each EURHISFIRM's partners, who will be in charge of forwarding them to national stakeholders. Meetings/workshops will be organized with strategically relevant stakeholders.

#### IV. Conclusion

The work done for Deliverable D10.1 allowed us to identify clearly a number of key characteristics that EURHISFIRM's business and governance model is likely to develop when it turns into an operative RIs:

- ▶▶ it will be a prevalingly Public RI providing both basic and value-added services
- ▶▶ it will endorse the Open Science Principles by integrating into EOSC and SSHOC
- ▶▶ it will be an ERIC service provider within CESSDA.



## Work Package 11: Cultural heritage

### I. Introduction

Work Package 11 explores concepts and tools to stimulate the lasting conservation of the digitized material and provides guidelines for making those materials publicly accessible. It also explores innovative ways to use digitized images as documentation for the data extracted from them and evaluates alternative strategies to use digitized material. More specifically, Work Package 11 has three main objectives:

- ▶ The use of digital images to document data and inspire further research and Identify sources of interest for cultural heritage
- ▶ The promotion of Europe's cultural heritage by facilitating digital preservation and online accessibility of sources with a unique historical value;
- ▶ The mobilization of digitized images of historical sources as an exceptional additional documentation for the data (including the exploration of ways to make materials accessible and connected to EURHISFIRM data).

### II. Progress: Logistics and Resources

WP 11 is composed of two tasks : (1) Task 11.1: Evaluating strategies and practices to value cultural heritage and (2) Task 11.2: Elaborate criteria of partnership to value cultural heritage. The timing of the two tasks, as per the Official EURHISFIRM deadlines, is as follows:

- ▶ Task 11.1: Evaluating strategies and practices to value cultural heritage.
  - ▶▶ Start date: 1 November 2019
  - ▶▶ End date: 20 April 2020
- ▶ Task 11.2: Elaborate criteria of partnership to value cultural heritage.
  - ▶▶ Start date: 1 May 2020
  - ▶▶ End date: 21 October 2020

As per EURHISFIRM agenda, Work Package 11 officially commenced in November 2019. Planning and preparation have already taken place for Task 11.1, due end of April 2020. No deviation from the original schedule is expected. Logistical and operational progress was thus carried out in support and development of future goals and milestones. These include:

- ▶ A new hire for the project: Joanna Kinga Ślawatyniec started as a research fellow for the project at Erasmus Rotterdam University (simultaneous hire for Work Package 9 and Work Package 11, and help for Work Package 4 on UK data knowledge). She joined on December 1, 2018.
- ▶ General meetings with all group members: To assure transparency and an awareness of EURHISFIRM goals, aims and deadlines face to face catch up meetings have taken place



- ▶▶ The overarching aim of the meetings is to ensure that all the project requirements are being met and EURHISFIRM deadlines will be fulfilled.
- ▶ In addition to the face to face meetings with all team members, regular Skype sessions take place as well as other meetings, including:
  - ▶▶ Monthly catch-up session between Joanna Kinga Sławatyniec and Coen Fierst van Wijnandsbergen
  - ▶▶ Bi-weekly Skype sessions between Joanna Kinga Sławatyniec and Joost Jonker
  - ▶▶ Occasional Skype calls between all team members
  - ▶▶ A meeting covering technical aspects of the project (time writing, etc.) between Joanna Kinga Sławatyniec, Marlies Vreeswijk, and Juan de Weger that took place on mid-December in Rotterdam.

### III. Progress: Project Achievements

As discussed above, the first Task of Work Package 11 is due end of April. Everything is on track keeping in mind the Work Package's milestones and deliverables. Slight amendments have to be incorporated given the Covid-19 epidemic, in particular pertaining to archival visits.

### IV. Conclusion

Over the coming weeks work will be carried to finalize the first milestones and deliverable (i.e. D11.1 Strategies and practices to value cultural heritage (M25)). To meet the deadlines, Joanna Kinga Sławatyniec will work closely with Joost Jonker and Abe de Jong.



## Conclusions

Within the one-year period since the previous reporting (March 2019-March 2020), the project progressed satisfactorily. Since March 2020, due to the current ongoing global health crisis (COVID-19), some delays have been encountered; nevertheless, the project does not expect any significant difficulties or deviation from the original plans.

As the project enters in its third and final year, the key priorities will be to continue completing the deliverables and milestones and to ensure the good production of the final report for the RI design. The upcoming General Assembly meeting (which has been postponed to autumn 2020 from March 2020 due to the health crisis) will focus on these priorities for the final report.

Additionally, the third year will prioritise the following key goals:

- ▶ Ensuring the thematic and technical coherence of the final deliverables towards the overall global vision of the EURHISFIRM project and its future
- ▶ Sharpening the long-term vision of EURHISFIRM in terms of utility and sustainability, taking into account the current developments in the European research infrastructure and research communities
- ▶ Continuing to build the project's community and continuing to become involved in the wider European research infrastructure community.

