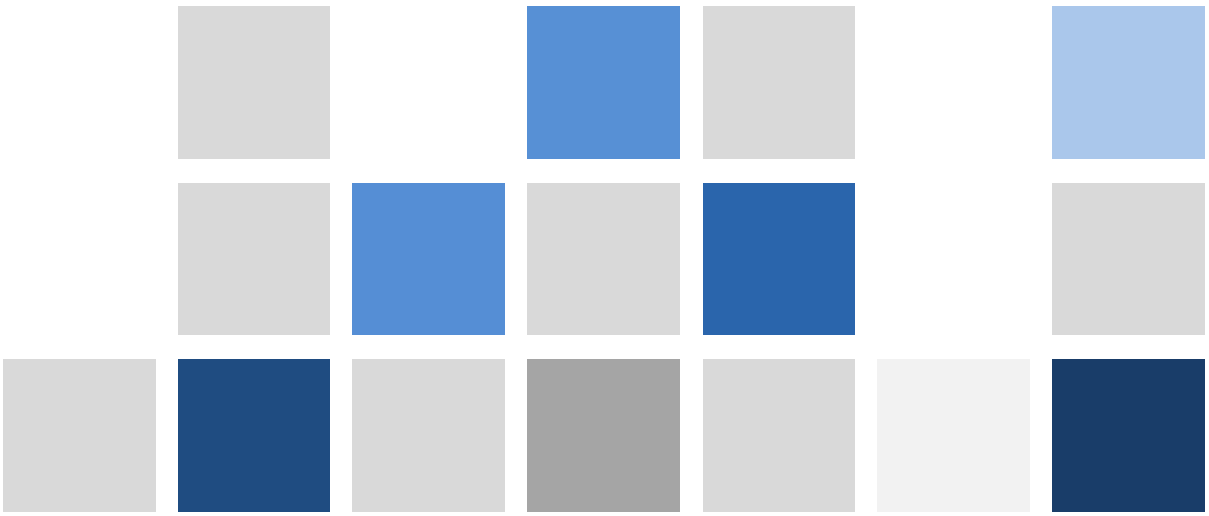
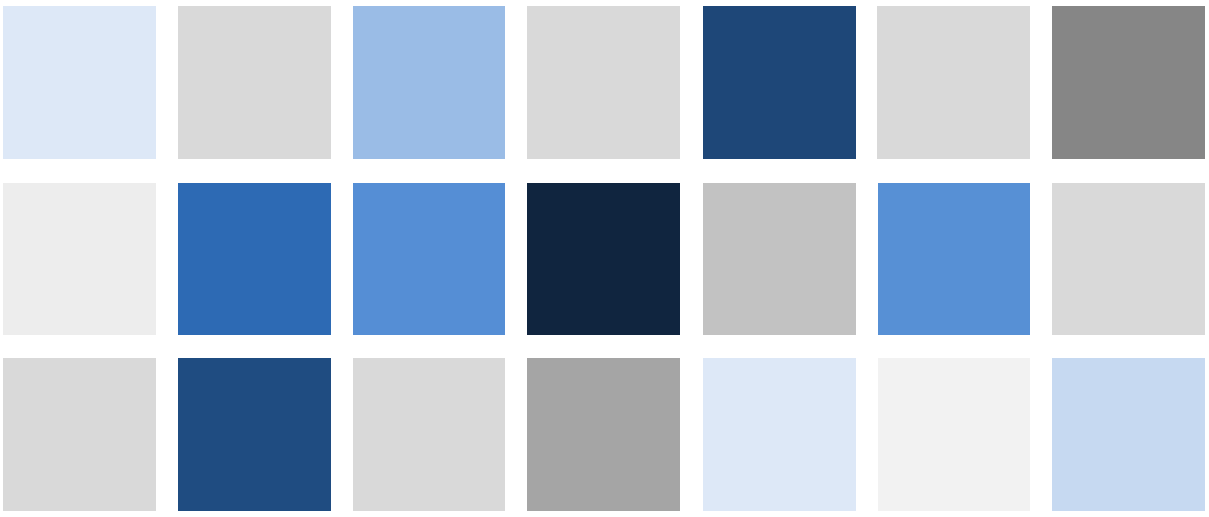


Long-term data for Europe

EURHISFIRM

D1.3: First 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

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)
Jérémy DUCROS (ÉCOLE D'ÉCONOMIE DE PARIS)
Coen FIERST VAN WIJNANDSBERGEN (ERASMUS UNIVERSITEIT ROTTERDAM)
Nathalie GIRARD (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Renata GWOŹDZIEWICZ-PĘCHERZEWSKA (UNIWERSYTET EKONOMICZNY WE WROCŁAWIU)
Stefan HOUPT (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 KUZIAK (UNIWERSYTET EKONOMICZNY WE WROCŁAWIU)
Aurélie LEMAITRE (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
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)
Emmanuel RAVIART (ÉCOLE D'ÉCONOMIE DE PARIS)
Yann RICQUEBOURG (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Uwe RISCH (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Angelo RIVA (ÉCOLE D'ÉCONOMIE DE PARIS)
Andres ROJAS CAMACHO (UNIVERSITÉ DE ROUEN NORMANDIE)
Gabriel SCHNEIDER (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Joanna SŁAWATYNIEC (ERASMUS UNIVERSITEIT ROTTERDAM)
Wassim SWAILEH (UNIVERSITÉ DE ROUEN NORMANDIE)
John TURNER (THE QUEEN'S UNIVERSITY OF BELFAST)
Lana YOO (ÉCOLE D'ÉCONOMIE DE PARIS)

APPROVED IN 2019 BY:

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

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

List of terms and acronyms

CDM	Common data model
CESSDA	Consortium of European Social Science Data Archives
CRSP	The Center for Research in Security Prices
DDI	Data Documentation Initiative
DFIH	Données Financières Historiques
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
TOGAF	The Open Group Architecture Framework
WGIS	Work Group on Identification and Standardisation

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

Table of Contents

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

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**

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

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.**

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

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**.

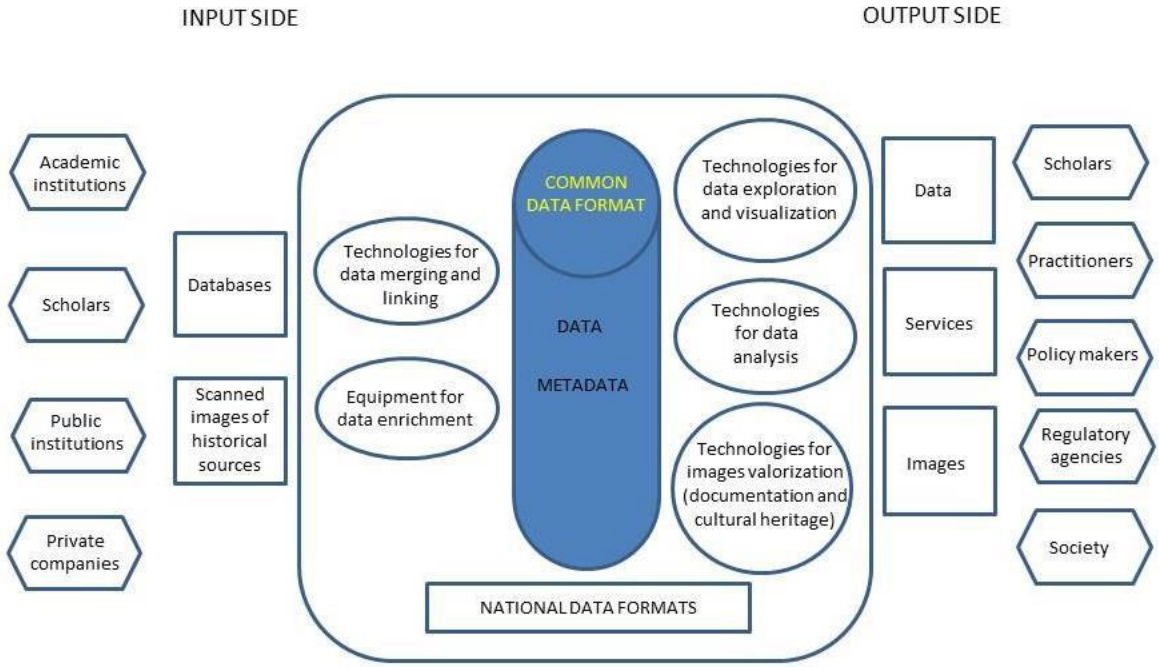


Figure 1 : The EURHISFIRM Research Infrastructure Concept

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.

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

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 19th- and 20th-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.

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

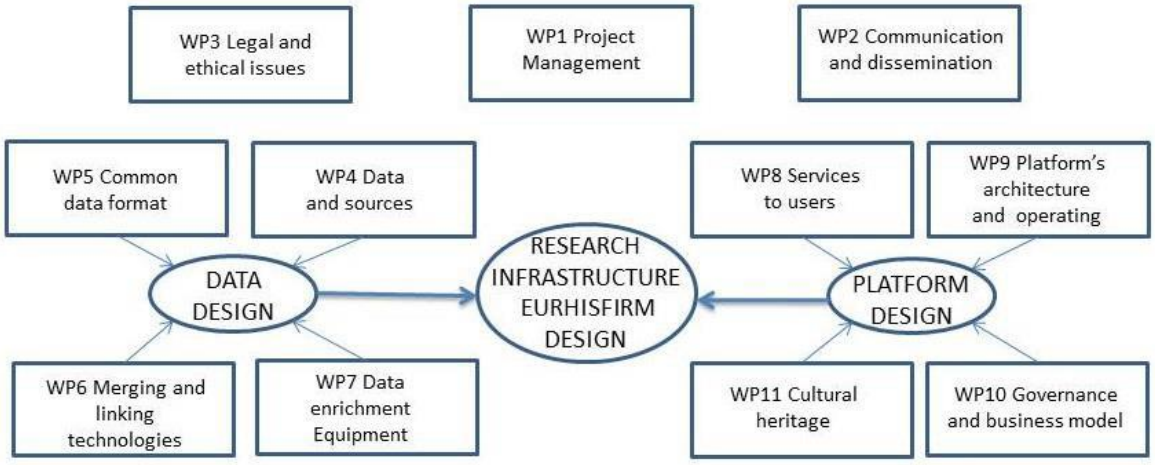


Figure 2 : Methodological Approach to the Concept

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 ERUHISFIRM (**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.

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

Executive summary

The EURHISFIRM project is on-schedule according to the project proposal. The accomplishments for each Work Package, as well as the cross-institutional Work Group on Identification and Standardisation (WGIS), are summarised below.

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 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. The priority for the first year was to establish the foundations for these parts.

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. All required deliverables and milestones to-date have been submitted (9/54 [17%] and 3/21 [14%], respectively). Regarding scientific coordination, Work Package 1 is responsible for the Data Management Plan by aligning with other Work Packages on these points. To coordinate technical milestones, Work Package 1 tracks the technical Work Packages' progress and facilitates strategic discussions as necessary. These tasks are done in collaboration with all other Work Packages.

Work Package 2: Dissemination and communication

Work Package 2 centres around the communication about the project and dissemination of its results, both internally, but more importantly, also to the broad stakeholder community. The Work Package started by creating an identity for the project, by designing a logo, templates, and crucially a website www.eurhisfirm.eu that will be one of the centrepieces for the project's communication. In addition, the first steps on the social networks were set. In order to get more focus, a dissemination and communication plan was developed to define the target audiences and the channels by which to communicate with them, as well as what to disseminate to whom. To this end, with input of all the national teams, an inventory was made of the potential stakeholders, which include policy makers, academia, business and society.

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

Work Package 3: Legal and Ethical Issues

Open data and open access is a key requirement of the European Commission for developing a research infrastructure that will become part of its infrastructure roadmap. Nevertheless, the collection and sharing of data and images can be is linked to ownership and property right issues, which need to be identified and possibly solved. Ownership rights on images and data may place restrictive conditions on access and refuse data linkage. The aim is to design rules and policies for handling ownership rights with individual researchers, research institutes, data vendors and owners of historic paper sources. This task will also propose a policy for the data access in close collaboration with Work Packages 9, 10 and 11.

The preliminary study of German copyright and unfair competition law showed that much of the raw data is indeed in the public domain and can thus be used for any lawful content. There are, however, exceptions to this rule, in particular, if longer text passages are used. It is also evident that the relevant laws of copyright and – even more so – unfair competition differ in the EURHISFIRM countries. It will thus be necessary to study the legal situation in each of these seven jurisdictions. As regards future topics of consideration, the different possible alternative uses of EURHISFIRM database can be used as scenarios to discuss legal matters within task 3.1.

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) provides the most appropriate one, namely the DDI-Lifecycle standard, which can be produced and edited by the Colectica Designer software. In addition, thanks to the inputs of all participating teams, the Work Package was able to compile a comprehensive inventory of the data and their sources for the participating countries. During the reporting period, it also started working on documenting the semantics and how these evolved over time.

Work Package 5: Common data model

Work Package 5 focuses on the development of concepts, architecture, and design of an overarching European level, Common Data Model (henceforth CDM). 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 firstly document models available within, as well as outside, the institutions of the consortium and evaluates their strong points and their weaknesses. Secondly, based on the observations of the first step, it designs an initial model of historical, European firm-level data with information that is spanned in three dimensions: financial information, accounting information and management information.

We have started Work Package 5 earlier than originally planned. The approach so far indicates that previous work on model designs done in institutions of the consortium such as SCOB and DFIH provides solid building blocks for the design of the CDM. There are various challenges when adapting these national

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

models into a European level and at this point characteristics that lead to the success of established systems, such as CRSP and EUROFIDAI can be incorporated into the CDM. The overarching identification system and the semantic structure are future focal points of the CDM design.

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 has just begun. For the first test case, 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, will be linked and integrated. To this end, the two teams involved have started to look into the technical details to achieve this. At the same time, preparatory investigations have started to single out other databases, both internal and external to the project, for testing purposes.

Work Package 7: Data extraction and enrichment system

The aim of Work Package 7 is to develop an intelligent and collaborative system for the extraction of structured information from images of historical documents related to companies' financial and economic activities. The sample document datasets for testing and validating the recognition system has been defined and selected. Based on one of the sources included in the sample dataset, Work Package 7 has started to specify and design elements of the prototype: a library of document component detectors for structure recognition, a general purpose text recognizer, a grammatical description of the source structure, a named entity extraction module for yearbooks, and a man-to-machine interface to correctly browse and validate the information. An automated linking service has been started to enrich the extracted named entities and to link them to trusted external knowledge sources.

Work Package 8: Interaction with users

The objective of Work Package 8 is 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, has been developed in order to identify the preferences of potential end-users and key stakeholders of the EURHISFIRM project. The survey has been conducted and its results were analysed. The Work Package is now identifying qualified people to conduct semi-structured interviews on their perspectives and preferences for the design of EURHISFIRM.

Work Package 9: Infrastructure policy and architecture

Work Package 9 designs the architecture and the operation of the RI, with regard 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). Users' preferences on data and services design guide the platform's architecture and operation. Accordingly, the

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

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 RI such as CESSDA and DARIAH. Following the user requirements' specification and RI policies. As per EURHISFIRM agenda, the Work Package is to officially commence late 2019, so it is in a very preliminary stage at this point. Preliminary architecture design was developed and ideas regarding the frontend model were proposed. Other issues addressed include, among others, maintenance, functionality and database administration design.

Work Package 10: Business model and governance

The aim of Work Package 10 is to develop a business and governance model that contributes to the articulation of the EURHISFIRM's platform design. The first deliverable of Work Package 10 is due on March 2020, so at this stage the Work Package is in a very preliminary stage. Nevertheless, significant progress has been made concerning two tasks: 1) the definition of alternative business model concepts, and 2) a preliminary assessment of business and governance model alternatives. While business and governance model can vary across countries, the main source of revenues for data repository is large structural public funding.

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.

As per EURHISFIRM agenda, Work Package 11 is to officially commence late 2019, so at this stage the Work Package is in a very preliminary stage. Steps taken so far are all in preparation for the official start of the project sections. Logistical and operational progress has thus been carried out to support and develop future goals and milestones. Over the coming months, work will be carried out to prepare the first milestones and deliverables (i.e. D11.1 Strategies and practices to value cultural heritage (M25)).

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

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.

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).

Coordination and development of scientific work (economic history research)

Two long-term national databases currently exist in the consortium: SCOB (**S**tudie**c**entrum voor **O**nderneming en **B**eurs) of the University of Antwerp and DFIH (**D**onnées **FI**nancières **H**istoriques) 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. The core part of this work includes, with PSE economic historian Jérémy Ducros (see Work Package chapters 4 and 7 for further details), coordinating the technical and the scientific teams and ensuring alignment between the data management plan and the work throughout the consortium.

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. This includes, with PSE IT expert Emmanuel Raviart (see Work Package chapters 4 and 6 for further details), coordinating the technical Work Packages (5, 6, 7) and participating in the cross-institutional Work Group on Identification and Standardisation (WGIS), as well as aligning with other relevant Work Packages on data standards (such as Work Package 4 on the metadata work).

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. 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.

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

To facilitate the work in the consortium, we have implemented tools such as file sharing technologies, video-conferencing tools, and instant messaging systems provided by and hosted on Huma-Num [https://www.huma-num.fr], the French national services provider of DARIAH.

Additionally, other required administrative procedures have been completed, such as the finalisation of the Consortium Agreement and budget distributions.

The project manual (D1.1) also outlines the established protocols on common procedures such as deliverable approvals and financial reporting. This manual, as well as other master documents (such as the project directory and meeting minutes) are available on an internal repository accessible to all project members. Work Package 1 also continues to monitor the project's compliance to the Grant Agreement and the Consortium Agreement.

III. Progress: Project Achievements

In terms of project outputs, all required deliverables and milestones to-date have been submitted (9/54 [17%] and 3/21 [14%], respectively). Each deliverable undergoes a rigorous check and approval process based on scientific soundness, organisation and communication, and overall alignment with the project strategy.

In addition to the logistical foundations described above in section II, Work Package 1's priority for the first year was to collaborate with the other Work Packages to ensure the project's scientific and technical coherence and to coordinate the overall scientific direction and strategy based on the decisions of the Steering Committee and the Executive Committee. These achievements include:

- Creation of the data management plan according to the Horizon 2020 Programme Guidelines on FAIR Data Management and by integrating the relevant work taking place in the other Work Packages.
- With PSE economic historian Jérémy Ducros (see Work Package 7 for more details) to coordinate other consortium members on the development of their specifications for treating their countries' sources for Work Package 7. (Definition of "specifications": guidelines on how to read, interpret, extract the raw information (i.e. paper sources of historical stock data) into digital forms.)

With Work Package 2 when needed, Work Package 1 has also executed the planned meetings outlined in the project proposal.

- These include: project kick-off meeting, Executive Committee meetings, Steering Committee meetings, General Assembly [including the Project Advisory Board meeting]).
- Work Package 1 integrates the key decisions and strategies from these meetings into the overall coordination of the project. The future deliverables of Work Package 1 (such as the midterm and final reports and future versions of the annual reports) will report on these strategic, scientific, and technical developments.

In terms of technical coordination, notable progress has been made, particularly in the common data model and the data integration with the existing systems (namely the SCOB and DFIH databases). For

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

further information, please consult the Work Package 5 and Work Package 6 chapters of this report. In brief, these achievements consist of:

- Creation of a Wikibase-based data repository by PSE IT expert Emmanuel Raviart to import long-term data: to test data import from relational databases into this web-semantic based technology, the data from the French DFIH project (Equipment of Excellence Data for Financial History) have been first used. In the following months, other tests will be run.
- Collaboration with the relevant teams (Work Package 5, 6, 7, 4, and the cross-institutional WGIS group) to begin implementing data management and storage standards, such as the Data Documentation Initiative Lifecycle (DDI 3.2), the Legal Entity Identifier (LEI) system, and the Financial Industry Business Ontology (FIBO), within the framework of The Open Group Architecture Framework (TOGAF) methodology.
- Ensuring compliance to FAIR data practices in the project developments and aligning the technical work with the overall project goals, vision, and outcomes.

IV. Conclusion

As mentioned above, the first year was dedicated to setting the logistical, scientific, and technical foundations for the project. Going forward, Work Package 1 envisions elevating the project's direction and execution to the next level, especially by sharpening its strategic vision and aligning this to the project execution. These include:

1. Updating the Data Management Plan for the subsequent deliverables
2. Continuing to increase our project's collaboration with and utility to the European community and commitment to FAIR data via our interactions with the relevant scientific, business, and public communities
3. Staying up-to-date with the developments in financial data and technology to understand our project in the larger global context
4. Redefining key performance indicators, where applicable
5. In terms of logistics, continuing the documentation of the project developments and decisions and ensuring that the project members' resource needs are met
6. Updating our risk assessments in line with our project progress and developments.

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

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.

The first three tasks were completed during the first year of the project:

- D2.1: Developing the project's identity and brand. Start date: M1; Duration: 3 months
- D2.2: Development of a dissemination and communication plan. Start date: M1; Duration: 3 months
- D2.3: Inventory of European and national distribution networks. Start date: M1; Duration: 6 months

II. Progress: Logistics and Resources

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

III. Progress: Project Achievements

The first three tasks were completed during the first year of the project. Here is the description of particular activities.

D2.1: Developing the project's identity and brand.

This task was aimed to create a unique identity for the project, including the project website and identity material. The following activities were conducted in the first year:

Creation of recognizable Project Identity materials (logo, templates):

- Development of a unique brand for the project, taking into account its characteristic features;
- Adaptation to both substantive requirements and project guidelines;
- Setting guidelines and consistency for the Project;
- Elaboration of the identity of EURHISFIRM;
- Graphic design;
- Indication of guidelines for a coherent brand image (fonts, colours, backgrounds);
- Development of ready-to-use templates for Word and PowerPoint.

Creation of website for external communication of the Project:

- Creation of Create website layout
- Creation of subpages dedicated to relevant project requirements and subjects

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

- Regular implementation of information connected with the Project
- Creation of Facebook page – tool for social networking in the Project
- Profile creating
- Implementing information connected with the Project.

D2.2: Development of a dissemination and communication plan.

This plan was designed to outline the dissemination and communication activities and to raise awareness amongst the different target audiences. This plan was prepared in the first months of the Project. The following elements were specified:

- All the key messages as well as the target audiences;
- Appropriate tools;
- Appropriate platforms and channels;
- Participations in conferences and workshops, to meet the information needs of the target audiences and the project's communication objectives;
- Ensuring maximum outreach of all news and project results.

D2.3: Inventory of European and national distribution networks

The task aims to complete, with the support of the members and of identified stakeholders, an overview of the stakeholders in the project. The task was completed by the following activities:

- Overview of the communities with a stake or interest in the Project, stakeholder groups from different domains: policy makers, academia, business, society;
- Contacts to key/strategic from Project point of view institutions in Poland.

The other two tasks: D2.4: Building the project's community and D2.5: Project promotion and dissemination are conducted on a continuing basis.

The most important activity within D2.5 has been the organization of General Assembly Meeting, which will be held on March 15-16 in Wroclaw.

IV. Conclusion

The progress in the Work Package aligns with the planned schedule.

Going forward, Work Package 2 will continue working on the two remaining deliverables (D2.4: Building the project's community and D2.5: Project promotion and dissemination) by integrating these two elements together. In other words, the project promotion and communication will integrate the community building strategy in order to promote a strong stakeholder network.

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

Work Package 3: Legal and ethical issues

I. Introduction

Open data and open access is a key requirement of the European Commission for developing a research infrastructure that will become part of its infrastructure roadmap. Nevertheless, the collection and sharing of data and images can be linked to ownership and property right issues, which need to be identified and possibly solved. Ownership rights on images and data may place restrictive conditions on access and refuse data linkage. This task will take stock of the ownership rights concerning data and sources in the perimeter of the research infrastructure. It will also look at the ownership rights of the processed data and planned research outcome that will result from this project. The aim is to design rules and policies for handling ownership rights and data protection rights with individual researchers, research institutes, data vendors and owners of historical paper sources. This task will also propose a policy for the data access in close collaboration with Work Package 9 on technical issues related to data collection and dissemination, with Work Package 10 on the business model related issues of data ownership rights, with Work Package 11 on digitized images collection and dissemination.

The work done so far mostly concerns the Intellectual Property Rights (IPR) status of the raw data used in EURHISFIRM. Since it has not yet been decided in which way the EURHISFIRM database will be made available to the public (open access, partly open access, commercial uses), the potential legal implications of these choices can be discussed as scenarios. In addition, these legal scenarios depend at least in part on the IPR status of the raw data (public domain or protected data). Thus, current work on Task 3.1 concentrates on the latter question (IPR status of raw data). Concerning the legal situation in Germany, a preliminary assessment (in German) has been produced. Concerning the legal situation in the other six participating countries, a research assistant will be employed in March 2019. We expect the report to be ready in March 2020.

II. Progress: Logistics and Resources

After the start of the EURHISFIRM project, Alexander Peukert attended several meetings and in addition individually talked to several members of the EURHISFIRM consortium to gather information about the scope and strategy of EURHISFIRM. Those exchanges mostly concerned legal questions surrounding the collection of data in Germany, in particular concerning the “Handbuch der Deutschen Aktiengesellschaft”.

After having received exemplary copies of the “Handbuch der Deutschen Aktiengesellschaft”, two short-term contracts (20 and 40 hours) were concluded with PhD candidates of Alexander Peukert, namely Julian Zimara and Florian Eckert. They prepared a study on possible copyright and unfair competition claims against the unauthorized use of the “Handbuch der Deutschen Aktiengesellschaft” (see below, III).

In December 2018, Alexander Peukert received Report D4.2 on the Inventory of Data and Sources, which will form the basis of the legal studies concerning IPR issues related to the raw data used in the project (input data).

In contrast to deliverable 3.2 (report on ethics), which is due only 33 months after the start of the project, D3.1 is, however, due already after 17 months, i.e. at the end of July 2019.

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

The "Report on the Inventory of Data and Sources" (D4.2) has been finalized and brought to the attention of Alexander Peukert, who is responsible for D3.1, at the end of 2018 as planned. Nevertheless, the wealth and heterogeneity of historical sources and data is higher than expected. As planned, the design of the output of the project is also still under discussion, but the variety of possible outcomes multiplies the questions accordingly. Alexander Peukert therefore asked to extend the deadline for deliverable D3.1 by eight, at least six months. This extension will not slow down or risk the success of the project as a whole but instead make sure that the IPR report covers all relevant issues.

In line with this new timeframe, the budget available for Task 3.1 will be allocated to a research assistant who will work 25% part-time for one year on the IPR study. The position has been advertised at the end of February 2019. It is expected that the research assistant position will be staffed in Mid-March or the end of March the latest, so that the final report on IPR issues is expected at the end of March 2020. The assistant will work closely with Alexander Peukert. He or she will have a permanent desk and further infrastructure at the chair of Alexander Peukert.

III. Progress: Project Achievements

- Gathering of information concerning the relevant raw data and their use in EURHISFIRM (April-December 2018);
- Preparation of a memorandum concerning the question of whether the use of the "Handbuch der deutschen Aktiengesellschaften" by EURHISFIRM triggers copyright and/or unfair competition claims by former publishers of this book in February 2019

IV. Conclusion

The preliminary study of German copyright and unfair competition law showed that much of the raw data is indeed in the public domain and can thus be used for any lawful content. There are, however, exceptions to this rule, in particular, if longer text passages are used. It is also evident that the relevant laws of copyright and – even more so – unfair competition differ among the EURHISFIRM countries. It will thus be necessary to study the legal situation in each of these seven jurisdictions.

As regards future topics of consideration, the different possible alternative uses of EURHISFIRM database can be used as scenarios to discuss legal matters within task 3.1.

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

Work Package 4: Data and sources inventory and documentation

I. Introduction: Summary

Objectives

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, 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.

Progress

Work Package 4 is on schedule as we are halfway both in terms of duration and of deliverables.

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 and will continue to cooperate with information specialists from GESIS and PSE (tasks 4.1 and 4.5) and topical experts in business and financial history from his own university as well as research institutions in Amsterdam (KNAW), Belfast (QUB), Frankfurt (GU), Madrid (UC3M), Paris (PSE) and Wroclaw (WUE) (task 4.2, 4.3 and 4.4). Particularly Johan is cooperating with Emmanuel Raviart and Jérémy Ducros based at the PSE to evaluate the interoperability of metadata and data standards.

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

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

"first read" by the Steering Committee on 3 July 2018. Meanwhile, the project team has been informed of an industry initiative to define financial industry terms, definitions and synonyms using semantic web principles such as RDF/OW, the FIBO ontology. 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 February 11 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. Currently, work on the possible articulation of DDI Lifecycle and FIBO Ontology has been undertaken (see also Work Package 5).

III. Progress: Project Achievements

Progress

Work Package 4 consists of five tasks in total:

1. Information system and data documentation standards
2. Data and sources inventory
3. Data and sources semantics
4. Data and sources documentation production and quality assessment
5. Defining EURHISFIRM documentation standard

At the time of writing, tasks 4.1 and 4.2 have been completed. The deliverables were submitted in June 2018 and December 2018 respectively. Task 4.3 is halfway. A lot of preparatory work for task 4.4 has already been done and the actual production of metadata will start shortly.

Submitted deliverables

Task 4.1 resulted in a report on metadata standards. A one-page summary of this report can be found in appendix 1.

Task 4.2 resulted in an inventory, with a lot of contextual information, on the available datasets and historical printed serial sources on publicly traded companies in the post-1815 period. A one-page summary of this report can be found in appendix 2.

Future deliverables

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. Task 4.3 therefore aims to contextualise the categories of information commonly found in stock exchange yearbooks and price lists. At the time of writing, a preliminary version of the report on data and sources semantics (D4.3) has been sent out to the aforementioned topical experts. They have been asked to provide general feedback as well as additional information and examples for their respective countries by the beginning of March. The timely delivery of the final report will be

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

contingent on the timely receipt of their replies. Given the experience of task 4.2, we have no reason to doubt this will be the case.

Task 4.4 will result in the actual production of detailed metadata, up to the level of individual variables, for a selection of the most qualitative existing datasets and printed sources identified during task 4.2. The results of task 4.3, together with those of task 4.3, will be used as input into the standards development process in Work Package 5. The definition and structure of the data elements found in printed sources and datasets can contribute to the formation and adoption of more normative EURHISFIRM data standards, for instance a controlled vocabulary for the data elements. It will also further the analysis of the FIBO Ontology in regards to the contents, the interoperability with DDI and the issues linked to the historical concepts that are not included in the FIBO ontology. The analysis of FIBO standards is run by Emmanuel Raviart and Jérémy Ducros based at PSE while the interoperability with DDI will be jointly evaluated with Johan Poukens (the results of this analysis will be one of the subjects of deliverable D4.5). During task 4.1, Johan already acquainted himself with the selected metadata standard (DDI 3.2) and software (Colectica Designer). The software has also been licenced and installed after the completion of task 4.1. The documentation and samples that will constitute the basis for task 4.4 have also already been collected over the course of task 4.2 and 4.3. Task 4.4 will require only a minor involvement of project partners (for instance, providing additional scans or images of sources and examples of datasets). Therefore, there is no reason to assume task 4.4 and its deliverable will not be finished before the deadline of May 31, 2019.

IV. Conclusion

Work Package 4 builds up expertise regarding existing datasets and historical printed serial sources on publicly traded companies. Other Work Packages will draw on its output for the elaboration of a common metadata model (Work Package 5), data connecting and matching technologies (Work Package 6), and system for automated data extraction (Work Package 7). 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 RI. A scientific paper will make this information available to the research community.

We must keep in mind, that, finished deliverables notwithstanding, the work of Work Package 4 is and will remain a work-in-progress. We are confident that, through the combined expertise and experience of topical experts and extensive research in the catalogues of national, specialized and online libraries, the inventory (D4.2) includes all important longstanding serial sources. Since the completion of the report, however, a handful of previously unknown individual volumes have surfaced. The same is true for digital data collections. Apart from a few well-known databases, there is an undoubtedly high dark number of smaller datasets in the long tail of science. The amendment of the inventory (as well as the semantical documentation, for instance) will require a collective effort from all members and partners of EURHISFIRM as well as of the wider research community. We therefore also need to think about ways to capture new discoveries of sources and documentation and add them to the existing sets in a collaborative manner.

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

Appendix 1: Summary of report on metadata standards (D4.1)

In task 4.1, the Data Documentation Initiative Lifecycle standard was selected for producing documentation on data and sources (i.e. task 4.4). Data Documentation Initiative (DDI) is a set of metadata standards and controlled vocabularies for documenting studies and datasets in the social sciences and economics. It has been developed within the community of social sciences data archives and currently is the preferred data documentation standard of the Consortium of European Social Science Data Archives (CESSDA).

The DDI standard was chosen over other metadata standards for documenting research datasets. With increasing levels of detail, Dublin Core, DataCite and da|ra contain a set of elements for describing datasets and printed sources at the study level (common elements include for instance identifiers, title, creator, publisher, abstract, coverage and format). Da|ra adds elements specific to the social sciences (for instance for the description of data collections), but only DDI includes elements for a detailed documentation of the individual variables contained in a dataset. This information at the variable level is crucial to the analysis of semantics and the elaboration of data models within EURHISFIRM.

Two versions of the DDI standard are available. For EURHISFIRM, we selected the more extensive DDI-Lifecycle (currently version 3.2) over the simpler DDI-Codebook (currently version 2.5) since it is more suitable for documenting historical printed serial sources (primarily because it supports a wider range of date formats, including dates in different calendars). Furthermore, DDI-Lifecycle includes the possibility to reuse content over several instances. For instance, variables and other metadata elements can be documented once and then referenced in several studies. Finally, DDI-Lifecycle incorporates Dublin Core for citation-type metadata. This means that within the DDI standard, you have the option to use Dublin Core elements instead of DDI elements for capturing basis bibliographic information about a source. Because Dublin Core is widely used, this greatly improves the interoperability of EURHISFIRM with third-party databases.

To produce and edit data documentation in the DDI-XML format, we chose Colectica Designer. Colectica Designer is proprietary software, but there are no open source editors which support the DDI standard down to the variable level (Harvard Dataverse for instance mostly covers study level metadata). Since it is proprietary, Colectica Designer comes at a monthly licence cost of $ 59 per user. However, a licence is only required for editing. Viewing can be done with the free Colectica Reader. Furthermore, Colectica can import and export DDI-XML files so vendor lock-in is less of an issue.

For more information on the Data Documentation Initiative and Colectica, we refer to their respective websites:

- https://www.ddialliance.org
- https://www.colectica.com

Appendix 2: Summary of report on data and sources inventory (D4.2)

Task 4.2 consisted of gathering bibliographic references of historical printed sources of information and available datasets on companies. For reasons of feasibility, the emphasis was on publicly traded companies

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

in the post-1815 period and on governance, financial and geographical information. This focus on publicly traded companies notwithstanding, the inventory also includes many sources on unlisted joint-stock companies and on companies in general. After all, publicly traded companies may not have been listed directly after incorporation and many important sources (e.g. publications of constitutional documents) did not distinguish between listed and unlisted companies. In the report, these sources are categorised, contextualised and summarised.

We distinguish three categories of historical printed sources:

- Official (government) publications of governance and accounting information**:** Constitutional documents (i.e. deeds of incorporation and articles of association) were published first in official **collections of law and decrees** because, historically, the incorporation of joint-stock companies required government consent. Later, mandatory publicity of corporate events replaced government authorisation and **government gazettes** and **official newspapers** began publishing notices of, for instance, incorporation, modification of articles, nomination of directors and liquidation. Accounting data were also published in government gazettes and official newspapers as a consequence of mandatory disclosure of annual accounts.

- Financial market information (e.g. securities prices and dividends) was published by stock exchanges in **official prices lists**. (Parts of the) official price list may also have been published in newspapers which can serve as a substitute for missing official lists.

- For the convenience of investors, stock exchanges and commercial publishers also began publishing summary governance, financial market and accounting information in the form of **yearbooks**. A distinction can be made between yearbooks which focus only on companies listed on a specific stock exchange and yearbooks which publish information on all joint-stock companies in one country.

Listing the availability of all three categories for the seven countries of the EURHISFIRM consortium falls beyond the scope of this summary. In general, governance information is available form official government publications for the entire post-1815 period. Official publications of accounting data, however, only commenced at a later point in time (the first uninterrupted series started in Belgium in 1873, the last in Germany in 2007). Stock exchange official lists are available from the beginning of the nineteenth century in North-western Europe (and from the second half of the nineteenth century in Central and Southern Europe). Yearbooks, finally, are available from the last quarter of the nineteenth century, first quarter of the twentieth century at the latest.

In addition to these printed sources, scholars have also collected and published information on (publicly traded) companies in the form of printed compendia of companies and entrepreneurs and histories of individual businesses, as well as in the form of datasets and databases. These present a wealth of secondary information and the inventory therefore also includes the most important databases and compendia and bibliographies of business histories.

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

Work Package 5: Common data model

I. Introduction

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

Towards this goal, Work Package 5 firstly documents the models available within, as well as outside, the institutions of the consortium and evaluates their strong points and their weaknesses. Secondly, based on the observations of the first step, it designs an initial 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

The main tasks of Work Package 5 have been organized around two aspects. The first aspect (consisting of tasks 5.1 and 5.2) takes into account existing examples of historical financial databases whose implementations can provide requirements for the subsequent development and design of the CDM. This part focuses on the design of the attributes, structure, and standards of such databases for the purpose of developing the CDM. The second aspect (consisting of tasks 5.3-5.5) concerns the specification of the CDM from the perspective of the business requirements and needs of the end-user to access the EURHISFIRM research infrastructure in a consistent and understandable fashion that meets the needs of the user community. It proposes, receives feedback and updates accordingly the model specification based on end-user input and needs.

The Work Package 5 team consists of three persons based at Goethe University, namely Lukas M. Ranft, employed 35% by the project, Pantelis Karapanagiotis, employed 50% by the project and Jefferson Braswell as an external consultant. The first is responsible for the design of the front end part of the CDM, the second for the back end part and the third for standardisation and identification issues. This team works in close cooperation with Emmanuel Raviart (IT developer) and Jérémy Ducros (economic historian) based at the Paris School of Economics).

In terms of tools, The Open Group Architecture Framework (TOGAF) was introduced to help in the identification of key processes of the CDM. As far as it concerns technologies, Work Package 5 is focused on designing issues rather than proof of concepts and as such the need to introduce technologies is minimal.

III. Progress: Project Achievements

Based on the official project's timeline, Work Package 5 consists of five deliverables and a milestone. The first deliverable depends on the output of the specification of sources performed by Work Package 4. Subsequent deliverables depend on the output of the first deliverable. The official starting month of Work

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

Package 5 is February 2019. Nevertheless, to the extent that it was feasible, some central issues of the package's deliverables have already been tackled.

Progress related to the documentation of national models, identification of best practices and standardization of initial metadata concepts has been achieved. These issues fall into the realm of the package's first deliverable. From the countries of the consortium, the existing financial historical databases of Belgium (SCOB) and France (DFIH) have been reviewed, and documentation of the underlying model is proceeding. The documentation process has included interviews with national experts on these databases. Outside of the scope of the consortium, current work focuses on understanding the EUROFIDAI and CRSP/Compustat models. The former is a model of European, contemporary, financial data and the latter a historical, firm-level US database. With the documentation process well underway, the focus in the upcoming months will turn on analysing the strong points of the various models in order to inform the development of data standards and best practices that the CDM can potentially adopt.

In parallel with the above, progress have been achieved in the direction of the second deliverable. The second deliverable is the core of the design of the CDM. Although the final assessment requires the completion of the first deliverable, preliminary work has been done with respect to mainly three topics. Firstly, commonalities that represent best practices in the models of the consortium have already been identified as important features of the CDM. For instance, the principle of preserving the original source of information is a feature that one can find in both the SCOB and DFIH models and it is most certainly a desirable feature of the CDM. Secondly, on the topic of investigating the potential overarching identification system, the reference data model of the Legal Entity Identifier (LEI) system as developed and supported by the Global LEI Foundation (GLEIF) is examined. Thirdly, in providing evidence that the DFIH SQL relational data model is not dependent on proprietary features of the currently used Oracle database engine, a small-scale script was developed to port the database into an open source SQL relational database engine (Postgres). It should be said, however, that the development and definition of the CDM is not directly aimed at determining the technical implementation of the physical infrastructure that will eventually support the CDM, although these considerations could be taken into some account should they affect the higher-level definition of the CDM. Fourthly, there is also ongoing evaluation of the benefits and weaknesses of relational and non-relational technologies on the CDM design. To evaluate the technologies, the web-semantic based system is developed in parallel with the analysis of the RDF based FIBO ontology. Its articulation with DDI documentation standards is also under investigation.

IV. Conclusion

Consequently, the approach so far indicates that previous work on model design done in institutions of the consortium like SCOB and DFIH provides solid building blocks for the design of the CDM. There are various challenges when adapting these national models into a European level and at this point characteristics that lead to the success of established systems, such as CRSP and EUROFIDAI can be incorporated into the CDM. The overarching identification system and the semantic structure are future focal points of the CDM design.

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

Work Package 6: Data connecting and matching

I. Introduction: Summary

Objectives

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 RI must be matched. 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 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 will continue to do this work for Work Package 6 as well. He will be joined by a top expert in database development, Boris Cule who will enter the project in the course of 2019. They will work in close cooperation with Emmanuel Raviart 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. Emmanuel Raviart is an IT developer 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 and Jérémy preparing and supporting the work of Boris and Emmanuel.

Tools and technologies

At the meeting of February 11, 2019 (see further) Emmanuel Raviart has presented some new tools to investigate the matching of databases. Actually he has also access to the Antwerp database and is trying out to make some first evaluations of these tools on matching and connecting issues between the DFIH database and the SCOB database.

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

III. Progress: Project Achievements

Progress

Work Package 6 consists of two tasks:

- 6.1 Report on data matching issues and methodologies (M23)
- 6.2 Report on data connecting issues and methodologies (M29)

As indicated in the project proposal, the start date of Work Package 6 is M11 (February 2019). As a consequence, initial efforts have already been made. Johan Poukens is already preparing Work Package 6. 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: the Paris (France) and the Antwerp database (Belgium). At the same time other candidate-databases will be investigated, such as the London Share Price database and EUROFIDAI. In addition, also databases of countries such as Portugal, Sweden or Spain are potential candidates. As a first inquiry a questionnaire has been prepared and was sent to the participants in the project as well as to the owners of some external databases.

Submitted deliverables

Not applicable (as Work Package 6 has just started).

Future deliverables

All of the deliverables still remain for submission (as Work Package 6 has just started).

IV. Conclusion

Work Package 6 will profit enormously from the experience built by Johan Poukens during his work on Work Package 4 and Jérémy Ducros who is coordinating the production of specifications for Work Package 7 but also from the experience built by Boris Cule and Emmanuel Raviart in previous years. We are quite sure we will accomplish the work on that Work Package within the time set out in the project proposal.

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

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. To keep the Work Package manageable, the system only takes into account historical printed sources related to listed companies such as yearbooks and exchange lists. We first defined and selected the document samples dataset for testing and validating the recognition system we develop. We started to work on one of these sample tests: French Yearbooks (Desfossés, 1962), to specify and design elements of the prototype: a library of document components detectors for structure recognition (section III.1.), a general purpose text recognizer (section III.3.), a grammatical description of the Yearbook structure (section III.2.), a named entity extraction module for yearbooks (sections III.4.) and a man machine Interface to correctly browse and validate extraction results (section III.5.). An automated linking service has been started to enrich extracted named entities and link them to trusted external knowledge sources (section III.6).

II. Progress: Logistics and Resources

II.1. Human Resources

The following people were recruited for Work Package 7:

- 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;
- Simon Bouvier, Research Engineer for Work Package 7, started on October 15th 2018, full-time (100%), IRISA, Insa de Rennes.
- Jérémy Ducros, economic historian (replacing Elisa Grandi, who had started on April 2018), Paris School of Economics. Coordination and production of specifications for Work Package 7.
- Gabriel Schneider, assistant for the work with deliverable 7.4, at the Hessische BibliotheksInformationsSystem (HeBIS), c/o Universitätsbibliothek Frankfurt am Main.

II.2. Document Dataset

A document samples dataset has been defined to build, validate and evaluate the recognition system which will be designed in Work Package 7. With the work of Work Package 4, examples of document images from more than 30 yearbooks and stock price lists, from 6 different countries, from different periods, have been proposed by partners. A first selection has been done on scan quality constraints. The Steering Committee did the last selection of the dataset. It follows the constraints proposed at the kick-off meeting, and is made of three yearbooks, three stock price lists, with three different languages, on three time periods: before WWI, interwar and post WWII (see Table 1). It has been decided to start working

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

on the French documents because of their availability and quality. First specifications on these documents have been defined.

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

Table 1: Document samples dataset for Work Package 7

III. Progress: Project Achievements

III.1. Specification and Design of a Library of Document Components Detectors for Structure Recognition – Task 7.2 (IRISA)

We developed a library containing some tools usable for the structure recognition on different kinds of corpuses, yearbooks and stock price lists, including:

- the recognition of table rulings;
- the recognition of logical separators of columns and rows in tables which do not contain physical rulings;
- the localization of text lines and segmentation of these lines depending on the tabular structure;
- the reconstruction of fragmented text lines inside a same column.

For the recognition of tables rulings, we use a segment extractor based on Kalman filtering. Due to the scan quality and the deteriorated state of some of the documents, this tool needs more context to recognize all the physical elements in a page. We have to work on imperfect detection such as fragmented rulings or false positives. To this end, we use the DMOS-PI method to define a grammatical description in EPF (Enhanced Position Formalism) that combines the line segments extracted by the Kalman filtering in order to reconstruct the whole rulings.

In tables lacking physical rulings, we are able to build logical separators based on structural elements such as the alignment of text lines, or blank spaces.

In order to localize text lines within the page, we combine an existing system based on deep learning, called dhSegment [Ares Oliveira 2018], with a grammatical description in EPF. Indeed, dhSegment offers a good overall efficiency but still has some flaws that we have to deal with. For example, when columns are too close to each other in a table, text lines are often sub-segmented, with one line for multiple columns. Once again, a combination with a different grammar in EPF is used to split contextually text lines that are cut by a vertical ruling belonging to a table structure.

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

This library is shared with a French national project focusing on the lists of the Paris OTC market: HBDEX (Exploitation of Historical Big Data for the Digital Social Sciences: application to financial data), and thus, part of this work is done in this context. It is planned that this joint library will keep evolving throughout both of the projects.

III.2. Specification and Design of a Prototype of Yearbook Structure Recognition System – Task 7.2 (IRISA)

We started the work on the yearbooks structure to extract information from the administrators lists (Desfossés Yearbook, year 1962, tome 1).

These tables are composed of 3 columns containing personal information about all of the administrators: Names, addresses, companies and positions in which the administrator is involved.

The structural grammar used in the extraction of these data describes an administrator as a single object with a name, address and a list of positions and companies. This description is built on the elements detected by the library (section III.1.). The analyser is able to locate each piece of information in the page with the coordinates of the bounding box (Figure 1).

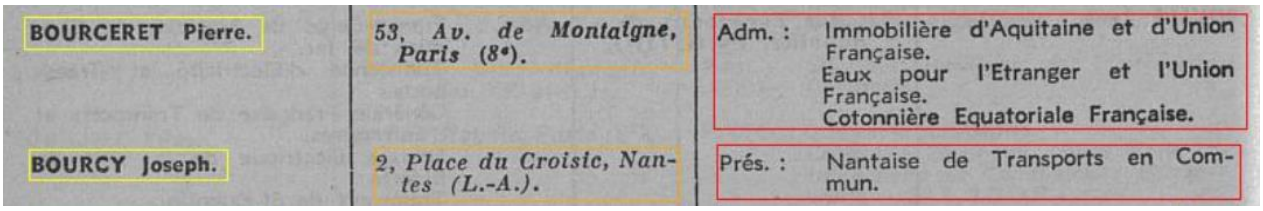


Figure 1: Example of structure extraction in tables of administrators in yearbook Desfossés, cell level.

The column on the right (companies and positions) is subdivided in two columns: one for the positions, and one for the affiliated companies. This logical separator is constructed following the alignment of the company's names, and is used to split the text lines that would otherwise cover both fields (Figure 2).

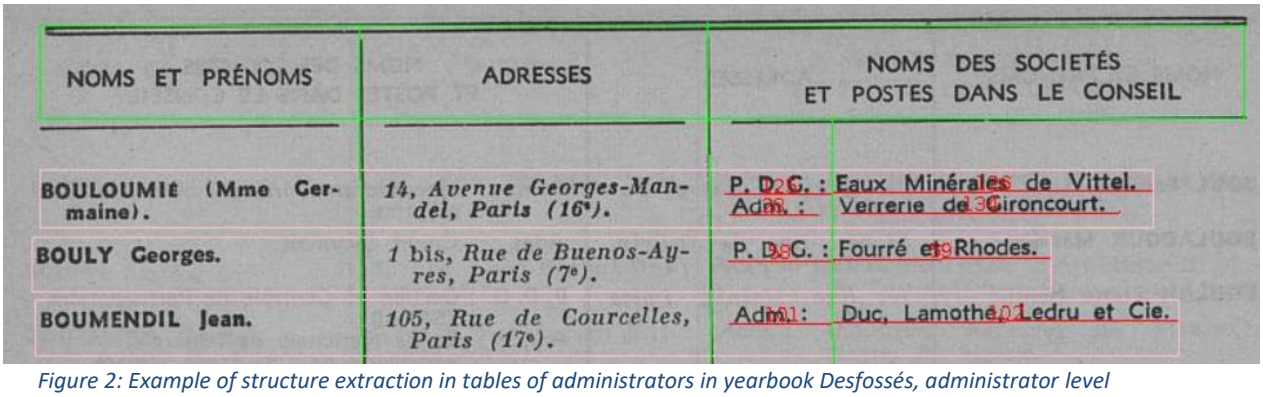


Figure 2: Example of structure extraction in tables of administrators in yearbook Desfossés, administrator level

We also started working on a second grammar with the intention of extracting the information from company sheets (Desfossés Yearbook, year 1962, tome 2).

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

The first step consists in locating the blocks of text composing a company: titles, sections and paragraphs, and extracting their content (Figure 3). This detection of the organization of the different blocks of text of a company, will be later linked to the work on the general-purpose OCR (section III.3.) and the named entities extraction module (section III.4.).

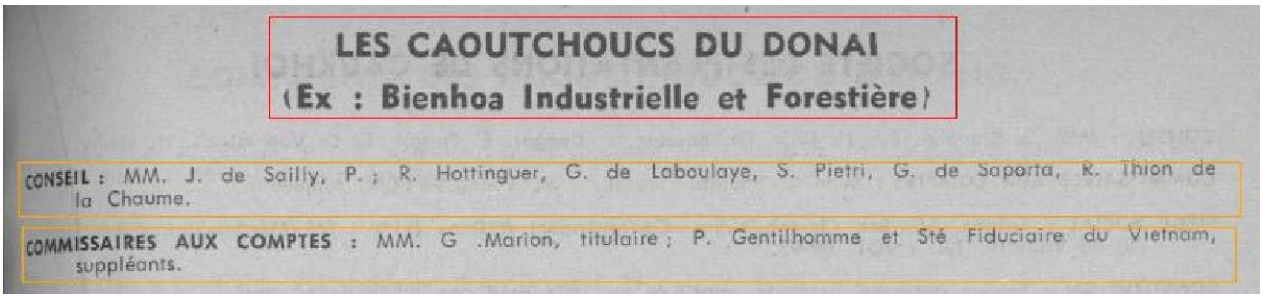


Figure 3: Example of structure extraction in company lists in yearbook Desfossés

The next task will be to analyse the structure of the different balance sheets, with a major difficulty coming from the variable structure of the tables, and the lack of any physical rulings.

The combination of all these descriptions and named entities extraction will allow to generate all the extracted information from the yearbook.

III.3. Specification, Design, and Training of a general-purpose text recognizer (OCR) – Task 7.3 (LITIS)

We design our own Deep Learning based OCR platform. The architecture is composed of 5 layers of convolutional neural networks (CNN) with a 3 X 3 kernel followed by a RELU activation function. Between each layer max-pooling with a 2 X 2 Kernel is applied. Inspired by the VGG-16 architecture the number of convolutional units increases when passing through the network, starting with 32 units for the first two layers and ending with 128 units for the last layer. The output of the last CNN layer is flattened to get a Vector that is fed to a dense layer composed of 256 units. Then the architecture ends with two bilateral recurrent layers composed of 256 units followed by a dense layer and a softmax activation function. The final output is a sequences of character hypothesis together with their respective probabilities.

Training this architecture is performed using the CTC loss function [Graves 2006]. The software has been developed using Python language and the Keras / Tensorflow platforms using a specific implementation of the CTC layer developed at LITIS (https://git.litislab.fr/TextRecognition/CTCModel).

The OCR outputs can be parsed using language models depending on the context of use. Language models can encode lists of possible stock names, or the syntactic rules used to write specific information such as prices, dates etc… The illustration in Figure 4 gives an example of a typical output of our general-purpose recognizer without the introduction of any lexicon post correction. Further extensions will concern extending the character set in order to include specific symbols such as stars, abbreviations etc…

Figure 4: Example of Deep Learning based OCR output, without any lexicon post-correction

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

III.4. Specification, design, and first evaluation of a named entity extraction module for yearbooks – Task 7.3 (LITIS)

We designed a prototype for named entity extraction module based on regular expressions. Starting with the specifications of the French yearbook 1962 (Desfossés), we defined three interdependent levels of information (Figure 5). Level 1 is related to the company names and every related information to each company. Level 2 is related to the description of each sub-structures related each company such as governing board etc… Level 3 is related to each named entity associated to each sub-structure of the company.

We designed three extraction components dedicated to the extraction of each type of entity.

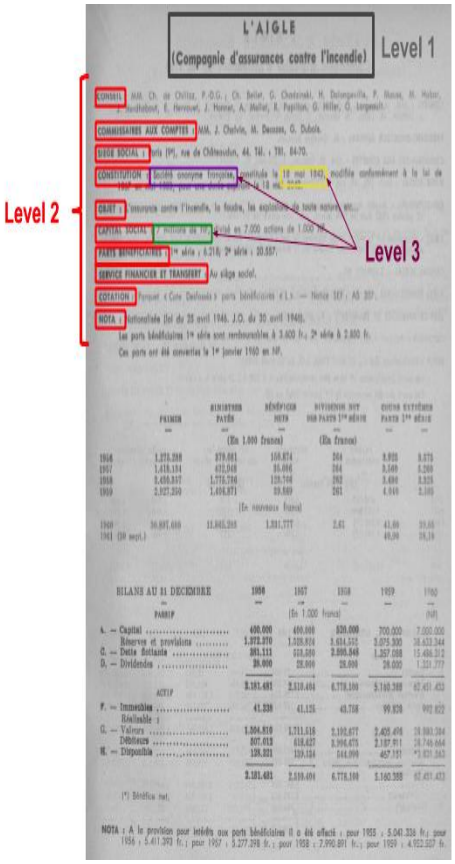


Figure 5: Named entities extraction in Yearbook Desfossés 1962: three interdependent levels of information

Level 1 information can be extracted by detecting the company names which are usually localized within the first paragraph of the page and represents the root for the next extraction level.

Level 2 information can be extracted by detecting the set of paragraphs that starts with specific keywords that are the names of the entities related to the company, such as “conseil d’administration” etc….

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

Level 3 information comprises the various named entities that describe the composition of the entity concerned. Here we need to extract person names, dates, amounts etc… and their meaning (date of creation, capital, …). We designed a rule-based approach using regular expressions that model the organisation of words and their specific typographical properties.

The extraction pipeline depicted in Figure 6 consists of applying an OCR on the document image to obtain the Unicode transcriptions at first. Then, the extraction system operates sequentially from level 1 to level 3 to produce the set of information that describes each company.

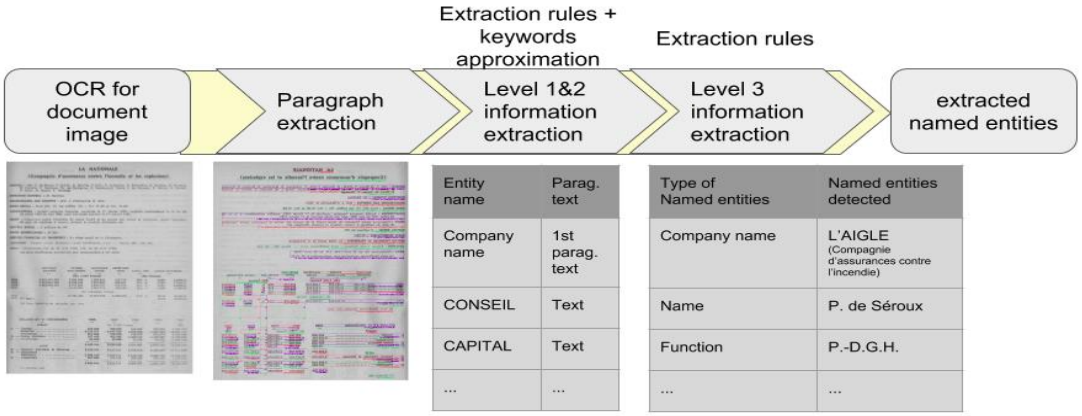


Figure 6: Named entities extraction pipeline

The targeted yearbook consists of 2376 pages, only 1689 pages contain the information to be extracted. Other pages are composed of illustrations, table of content, publicity, introduction pages etc… There are in total 18211 paragraphs that contain possible keywords to be extracted at level 2. In total, 608 possible keywords have been identified in these paragraphs, only 57 possible keywords occur more than 10 times and represent 95.45% of the whole set of paragraphs. Some of the 57 most frequent possible keywords contain misspelling errors that prevent them from being merged with other correct possible keywords. In order to correct misspelling errors, we use a lexical approximation method to match the possible misspelled keywords to a predefined list of verified and correct keywords. The other less frequent (551) unverified keywords represent a mixture of correct and misspelled possible keywords that cover 4.55 % of the paragraphs. Currently, we get the specifications from PSE for 19 keywords occurring in 7835 paragraphs that cover 43.02% of the whole set of paragraphs (excluding tables in the page). The table 2 gives the list of the 19 verified keywords. An additional list of keywords is currently under specification in close interaction with PSE to deal with the remaining 56.98% paragraphs.

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

Verified and correct Keywords	Occurrence	Verified and correct Keywords	Occurrence
CONSTITUTION	1458	CONSEIL D'ADMINISTRATION	62
OBJET	1427	PRÉSIDENT-DIRECTEUR GENERAL	49
SIEGE SOCIAL	1320	DIRECTEUR GENERAL ADJOINT	35
CONSEIL	1308	SECRÉTAIRE DU CONSEIL	26
COMMISSAIRES AUX COMPTES	1228	CENSEURS	19
DIRECTEUR GENERAL	349	PRESIDENT D'HONNEUR	13
SERVICE FINANCIER ET TRANSFERTS	289	CONSEILLER TECHNIQUE	4
CAPITAL	135	FONDE DE POUVOIRS	1
DIRECTION	111	COMITE TECHNIQUE	1

Table 2: 19 keywords automatically extracted at level 2 and their occurrence in company lists

The 19 keyword's specifications of paragraphs identified 14 named entity classes to be extracted in level 3. The following table 3 illustrates the identified named entity classes for each keyword. The keywords that describe the same paragraph's named entities are grouped together and use the same named entities extraction rules. For example, the keywords: CONSEIL, COMMISSAIRES AUX COMPTES, DIRECTEUR GENERAL ADJOINT, SECRÉTAIRE DU CONSEIL, CENSEURS, PRESIDENT D'HONNEUR, CONSEILLER TECHNIQUE, all are followed by sequences of person names and their function title. In other words, once these keywords are detected, we apply the extraction rules of the Name and Function named entities.

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

Keyword	Named entity classes					
	1	2	3	4	5	6
CONSTITUTION	Initial legal statut	Start date of initial legal status	End date of initial legal status	Transformation(s) of legal status	Transformation start date	Transformation end date
OBJET	Whole paragraph text excluding the keyword					
SIEGE SOCIAL	City	Departement (region)	Country			
CONSEIL (alternatives)	Name	Function				
CAPITAL	Changes of capital amount	Date of changes				

Table 3: 14 named entity classes extracted at level 3

The 14 named entity classes were extracted using specific regular expression rules. Currently we are working on the evaluation of the extracted named entities at level 3 using the standard evaluation metrics of precision, recall and F1 score.

III.5. Specification and Design of Man Machine Interface to Browse Correct and Validate extraction results – Task 7.3 (LITIS)

We designed a first prototype of a web visualizer devoted to browse the OCR results and the information extracted on a collection of documents. The OCR outputs stored in ALTO files, associated with their structured description in terms of designation (type of information such as stock name, price, date, person name, capital amount …) stored in METS files, are indexed in a database devoted to visualization purposes. Through its web browser the user will access to the information extracted for a particular collection of documents and browse the collection to view the original digitized documents (right part of the figure 7) and the extracted information (left part of figure below). The user is able to edit the result for correction purposes. Figure 7 below shows an example of the stock names extracted.

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

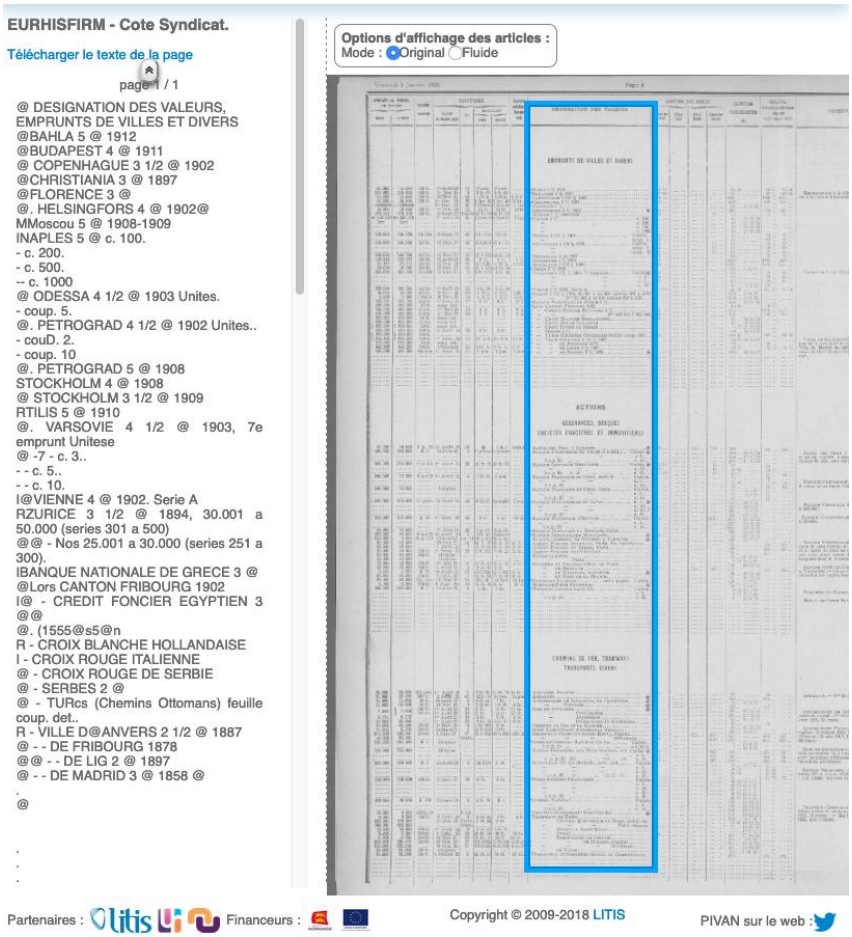


Figure 7: Prototype of man machine interface to browse, correct and validate extraction results

III.6. Automated Linking of Named Entities – Task 7.4 (HeBIS)

Task 7.4 designs a knowledge discovery and linking service to enrich extracted named entities and link them to trusted external knowledge sources like GND, ISNI, VIAF or GLEI. Task 7.4 started with already existing software tools SILK and LIMES. LIMES turns out to be the most suitable choice and will be used for future developments.

As part of the evaluation of linking tools, task 7.4 has already worked on and with existing datasets from EURHISFIRM project partners, external data sources and authority files. Linking of company data between data sets from Germany (SAFE), Belgium (SCOB) and France (DFIH) and with authority files GND and DBPedia showed that manual work is still needed to gain a high quality of links between data sets.

Future work in Task 7.4 needs to address an efficient workflow of linking and discovery in a mixture of automated and manual/intellectual processes.

The work on this task began in October 2018 to ensure a solid technical base for entity linking to be used at the scheduled start on January 2019 (Table 4).

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

Task	10/18	11/18	12/18	01/19	02/19	03/19	04/19
Evaluation of linking tools							
Conception and implementation if automated linking							
Documentation and Report							

Table 4: Schedule of Automated Linking of Named Entities

III.6.1. Evaluation of Linking Tools

As part of this project LIMES (Link discovery Framework for Metric Spaces) was tested as a tool to compute links between entities in different data sources and the first prototypic implementation of an automated method for the linking was conceptualized. LIMES offers a way to link entities using a linked data format and calculating similarity measures between compared data records. The following document describes the goals and the chosen approach for the task.

As part of the evaluation LIMES was tested, as well as SILK, another linking framework. By establishing a local triple store as a test infrastructure, different linking tasks with both frameworks were executed and evaluated. The test data was taken from databases which are publicly available online, such as the company database of BaFin and EBA. The data sources were then linked internally and with other data, such as GND and DBpedia. This step was carried out, to test the tools and the infrastructure for real linking cases. During the evaluation LIMES delivered positive results concerning the performance and possible linking configurations and was chosen as the tool to be used for the linking task.

III.6.3. Finding possible data sources

As test data different sources containing company data were analysed. Criteria for usage were the free access to the data set, export options, csv as file type, size of the data set and its content orientation. After finding a suitable data set, it was exported and its data scheme was analysed. The attributes of the csv file were then transferred to a new table and documented to achieve an overview over the acquired data sources. Additionally, historical company data from Antwerp University was transformed and imported into the local test infrastructure. This data will be linked to future possible data sources (see 5. Next Steps).

III.6.4. Converting the data

To achieve a common format an individual mapping for every data source was created. It transfers the sources attributes into the internal csv format. The csv format (see Table 5) contains following attributes, but is expandable, if needed.

The internal ID consists of a country code (taken from the country attribute) and a five-digit number. Additionally, an identifier of the source is added. This structure enables us to still see the source of an entity, when linking it via LIMES. Attributes as “legalName” and the different address attributes are used to identify the sameness of two entities. More attributes for unique identification of an entity lead to more possible linking queries via a linking framework.

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

The created RDF schema (see Figure 8) uses popular vocabularies to achieve interoperability to other data sources. For attributes about an organization, “schema.org” is used. More common concepts, such as a predecessor and/or successor, “Dublin Core” and “Friend of a Friend” is chosen.

attribute
internalId
externalId
leiCode
legalName
alternateName
country
countryCode (ISO 3166)
state
city
postalCode
street
periodActive
legalForm
predecessor
successor
homepage
sector
source

Table 5: Internal csv-format

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

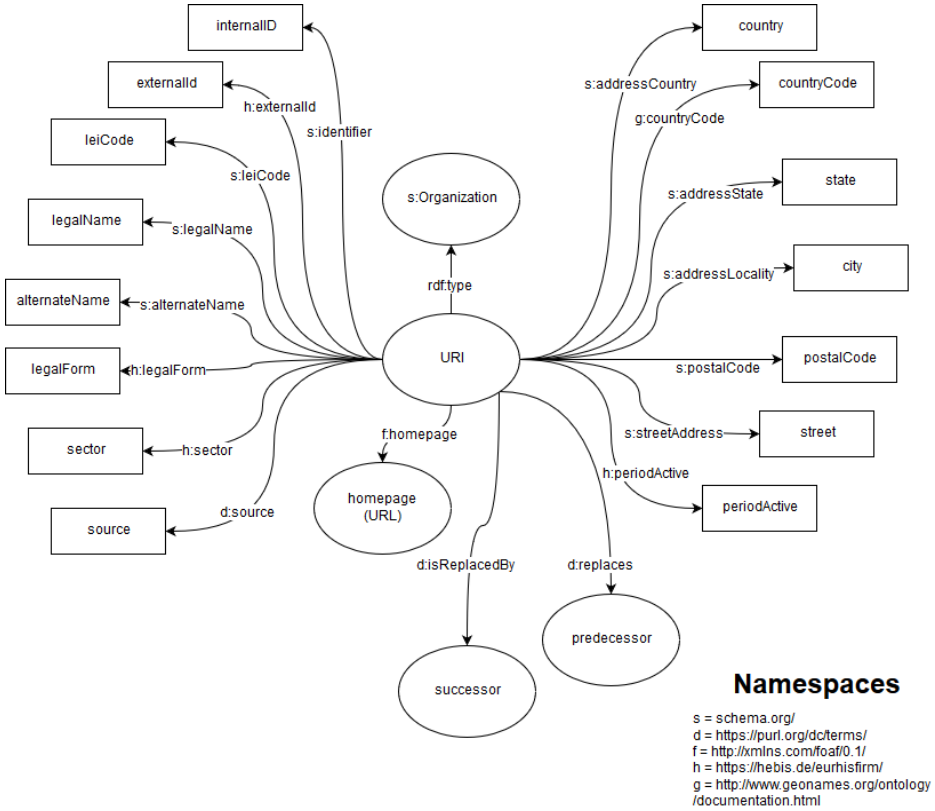


Figure 8: Internal RDF-Format

III.6.5. Infrastructure

After the conversions the RDF data sets were uploaded to a Fuseki triple store via an upload server. A RDF4J-API is used to execute lifecycle maintenance operations, such as create, updates and deletions. To differentiate between the ingested resources, the data was stored in different graphs in the triple store. For the evaluation of the frameworks the available data sources were converted, ingested into the database and then used for test linking tasks (Figure 9).

III.6.6. Configuring LIMES & SILK

When the data is stored, the frameworks were used to link the contents of the triple store to different sources. DBPedia can be queried by using a SPARQL-Endpoint. An excerpt containing company data was taken from the GND and also transformed and added to the local triple store. By using graphs for structuring, linking between sources on the same triple store was possible. In LIMES and SILK different configurations were used. Depending on the outcome the queries were analysed, changed and optimized to achieve the best possible results. After the evaluation a prototypical automated service for linking of entities in different sources will be conceptualized and implemented.

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

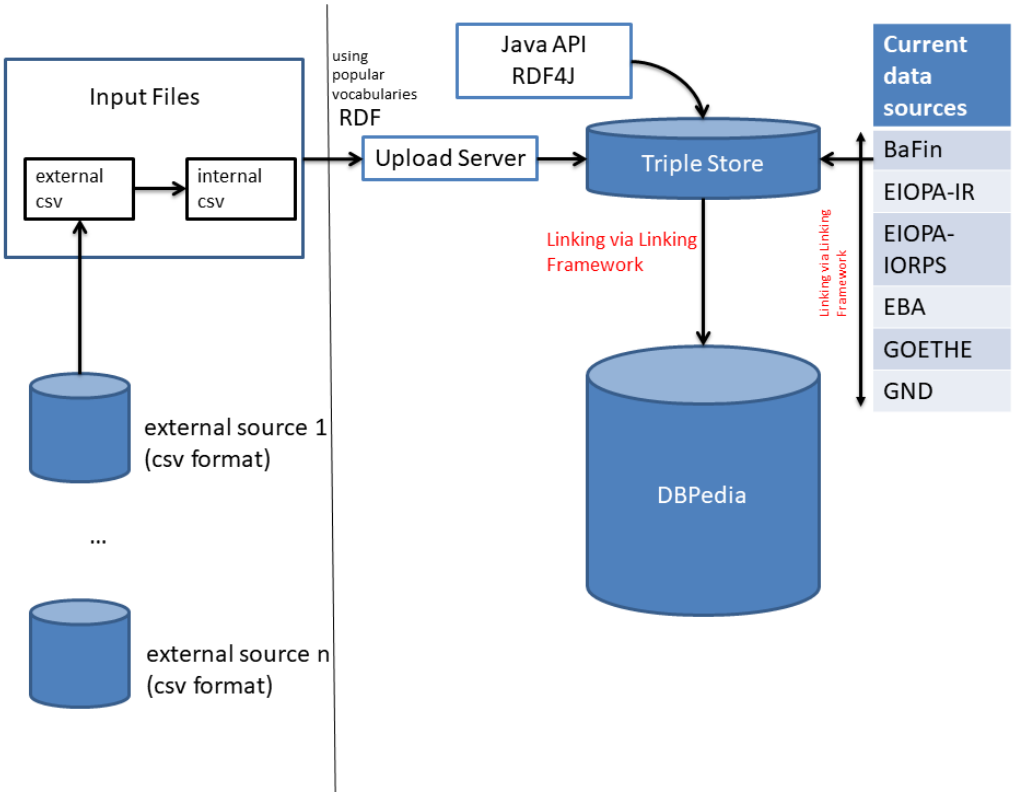


Figure 9: Infrastructure concept

III.6.7. Next Steps

The next step is to configure linking tasks between the data from the selected sample containing company names in French. The internally used data format has to be aligned with the work on the Common Data Format (Work Package 5) to ensure common ground for further developments.

A concept to safe the links between companies persistently is being worked on. The goal is to enrich the entities, that are identified as the same, with a common persistent identifier. This can be achieved by importing the results from a LIMES query into a triple store. Via the use of an “owl:sameAs” property between the entities, sameness is expressed. Following the creation of corresponding triples, a process will check, if the entities are already connected to a common identifier. If not, a new one will be created and both entities will be linked to it. If one entity already has such an identifier, it is added to the other entity.

For links established on the basis of a critical significance level a manual review can be done. The links can be deleted from a review file and after its adjustment, the remaining triples can be ingested into the workflow described in Figure 10.

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

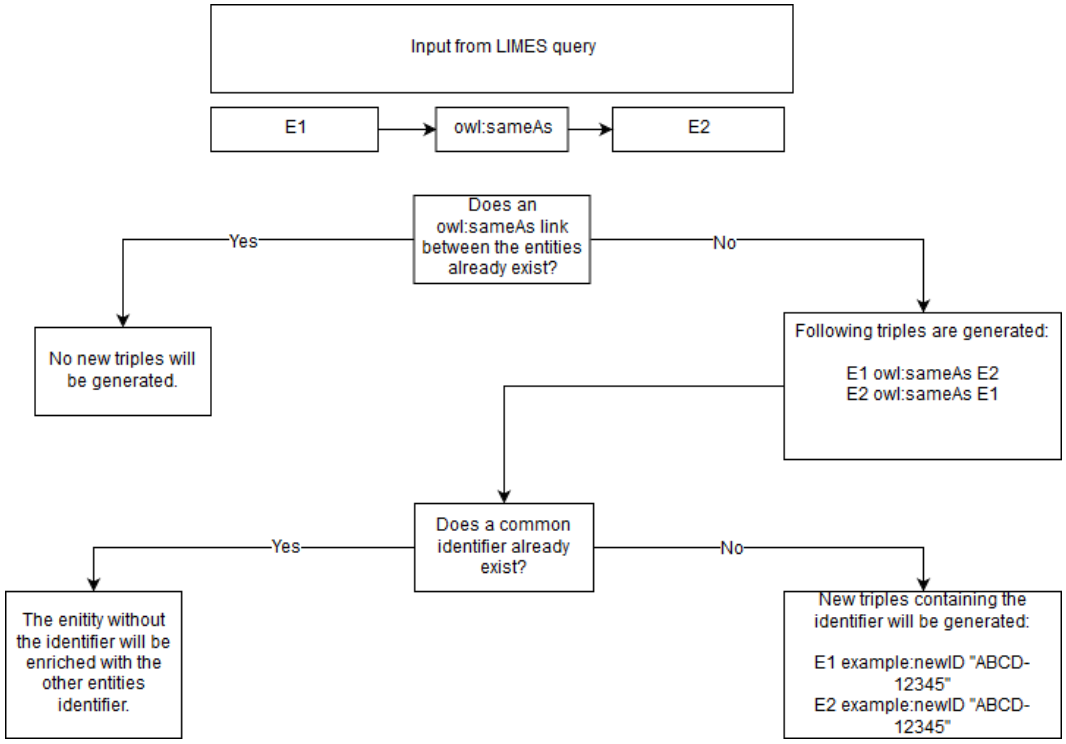


Figure 10: Ingestion workflow

III.7. Deliverables and Milestones

The work presented in section III.1. on the library of document components detectors for structure recognition, and in section III.3. on the general-purpose text recognizer (OCR) will constitute the deliverable D7.1 (Software libraries developed) at M12. The next deliverable is D7.2 (First version of the data extraction system), which corresponds to the milestone M7.1, both at M24.

IV. Conclusion

We have defined the document dataset on which the prototype of the document recognition system will be evaluated. The French documents from 1962 have been selected to start on building a prototype on yearbook. To introduce knowledge in the recognition system, specifications on those documents have been defined. The library of document components detectors for structure recognition and the general-purpose text recognizer (OCR) have been developed. A grammatical description of tables of administrators and list of companies found in the yearbook Desfossés 1962, have been defined, leading to a first structure recognizer prototype. A named entity extraction module on the same yearbook has been developed, as well as a first prototype of man machine interface for browsing, correcting and validating extraction results. A first prototype of automated linking of named entities has been defined to enrich extracted named entities.

Next steps will be on working on the strategy at the collection level to enable the system to cross validate information between pages to improve the recognition quality, and to drive the complete system. We will need to combine the different modules on document structure recognition with the named entity

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

extraction module and the man machine interface. We will need to design new descriptions like balance tables in yearbooks, tables of stock price lists. Name entities extraction will have to be adapted to the possibility of detecting stable names across documents. The introduction of financial knowledge in number recognizer in yearbooks and stock price lists will have to be done.

On the French documents we will also combine the automated linking of named entities with the document recognition system. We will then be able to generate a first prototype for yearbooks and stock price lists with web-linking.

To introduce financial knowledge in this prototype to improve the recognition quality, we will need to have the specifications on all the documents from the sample dataset (Belgium, German and Spanish) and try to define the common knowledge between all those documents to limit the adaptation of the system to a new kind of document. This common knowledge is also related to the work done in the Work Package 5 on the common data format. This work is also needed to start to adapt the system to the different kind of document of the samples dataset.

V. References

[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.

[Graves 2006] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks,” in Proceedings of the 23rd international conference on Machine Learning. ACM Press, 2006, pp. 369–376.

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

Work Package 8: Interaction with users

I. Introduction

The aim of Work Package 8 is 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 are:

- 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 guide the design and data policy of the EURHISFIRM RI.

We have completed the first two objectives and we are currently working on the third.

II. Progress: Logistics and Resources

Robin Adams has joined the team as a Research Fellow. Robin completed his PhD in Economic History at Oxford.

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 involves 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. We are currently in the process of identifying interviewees.

IV. Conclusion

The findings of the online questionnaire (D8.2) have provided a firm basis from which to proceed with the qualitative interview stage (D8.3) of research. Examined in detail in the D8.2 report, these findings have yielded valuable insights relating to the content and usability desired by potential end-users and stakeholders of the EURHISFIRM RI.

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

In terms of content, the twentieth century was the most popular period among the survey respondents, and the United Kingdom was the most popular country, followed by Germany and France. Ordinary equity market data was considered the most useful form of company data, followed by accounting data, data relating to government and corporate bonds, macroeconomic data and governance data. Regarding frequency, the respondents expressed a preference for daily data and monthly data, but not for weekly or annual data.

As regards usability, Bloomberg was rated the best data service provider for ease of use, while CRSP/Compustat scored best for data availability and for bulk downloads, and Yale Investors' Monthly Manual was considered best value for money. Respondents were invited to share some examples of good practice that, in their experience, make data services easier to use. A recurring theme in their answers was the importance of bulk downloads, with the speed and quantity of downloads given particular emphasis. The ability to manipulate data was also important to the respondents, as was the clarity of the user interface.

The qualitative interview stage (D8.3) of this project will build on the findings of D8.2, allowing a deeper, more nuanced analysis of the preferences of the EURHISFIRM RI's potential end-users and stakeholders.

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

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

As per EURHISFIRM agenda, the Work Package is to officially commence late 2019, so it is in a very preliminary stage at this point. Steps taken so far are all in preparation for the official start of the project sections. 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 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 meeting with all group members: To assure transparency and an awareness of EURHISFIRM goals, aims and deadlines a catch up meeting had taken place on January 9, 2019.
    - Attendees: Joost Jonker, Abe de Jong, Coen Fierst van Wijnandsbergen, Joanna Kinga Sławatyniec, Marlies Vreeswijk, and Juan de Weger.
    - The meeting was followed up by a Skype session with Lana Yoo
    - The overarching aim of the meeting was to ensure that all the project requirements are being met and EURHISFIRM deadlines will be fulfilled.
- In addition to the January 9, 2019 meeting, a couple other meetings and Skype sessions were held between key project group members, including:

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

- A general introduction session for Joanna Kinga Sławatyniec by Joost Jonker held in Amsterdam on December 10, 2018.
- An introduction session for Joanna Kinga Sławatyniec by Coen Fierst van Wijnandsbergen covering steps taken so far in terms of Work Package 9. This took place on December 11, 2018 in Rotterdam.
- 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.
- Other Skype sessions

Early work has been carried out for Work Package 9 (a work in progress document has been uploaded by Coen Fierst van Wijnandsbergen and is accessible via Seafile). Preliminary architecture design was developed and ideas regarding the frontend model were proposed. Other issues addressed include, among other, maintenance, functionality and database administration design.

III. Progress: Project Achievements

As discussed above, currently Work Package 9 is in a very nascent state. Nonetheless, 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 first milestones and deliverables (i.e. D9.1 Report on RI policy as part of the conceptual design report (M27)). To ensure the milestones are reached and deliverables are submitted on time, Joanna Kinga Sławatyniec will work closely predominantly with Coen Fierst van Wijnandsbergen, who has been working on the Work Package for a couple of months already. Furthermore, collaboration will also take place between Joanna Kinga Sławatyniec, Joost Jonker and Abe de Jong.

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

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).

Since the start of the project, we made significant progress in the acquisition of key information related to the governance and business models of existing RIs in Social Sciences, contributing to T1 and T2.

II. Progress: Logistics and Resources

In the next months, we will hire the human resources required to the achievement of the tasks and deliverables assigned to Work Package 10. We expect to complete the process of recruitment during Spring 2019, so that the hired professional(s) will start working full-time at the implementation of Work Package 10 in September 2019.

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[1] [2]. 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

[1] OECD Global Science Forum and CODATA. "Business models for sustainable research data repositories". OECD Science, Technology and Innovation Policy Papers, n. 47, December 2017.

[2] OECD Global Science Forum and CODATA. "Coordination and support of international research data networks". OECD Science, Technology and Innovation Policy Papers, n. 51, December 2017.

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

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 (see Figure 1).

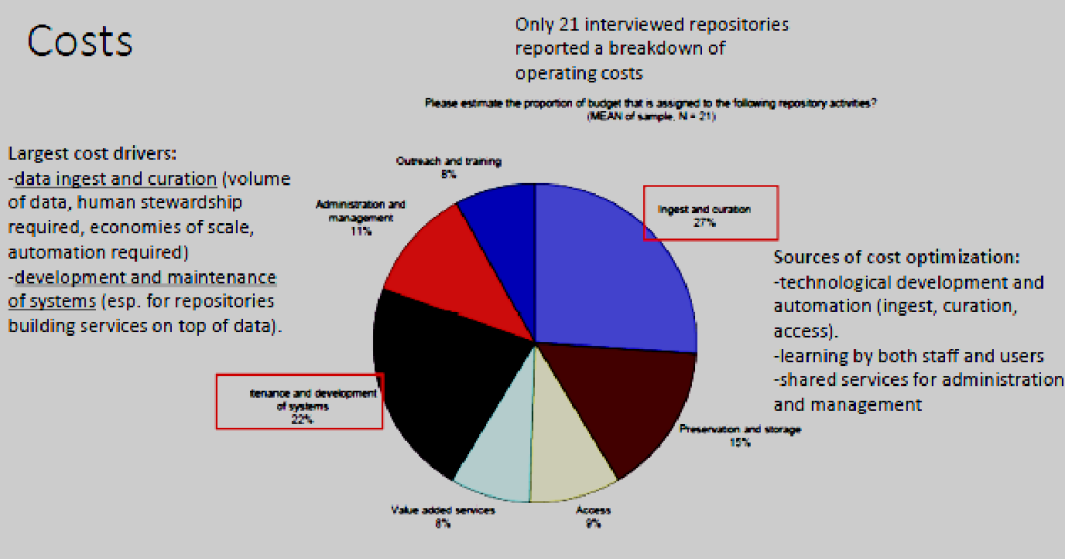


Figure 1. Source: OECD Global Science Forum and CODATA. "Business models for sustainable research data repositories". OECD Science, Technology and Innovation Policy Papers, n. 47, December 2017 (p. 25)

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 in generally limited, as a majority of RIs report funding from one source only (see Figure 2).

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

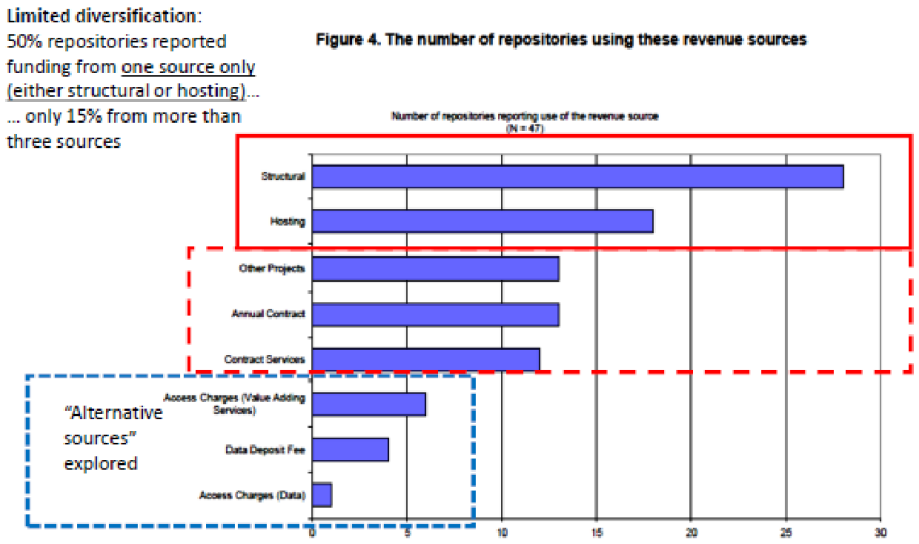


Figure 2. Source: OECD Global Science Forum and CODATA. "Business models for sustainable research data repositories". OECD Science, Technology and Innovation Policy Papers, n. 47, December 2017 (p. 23)

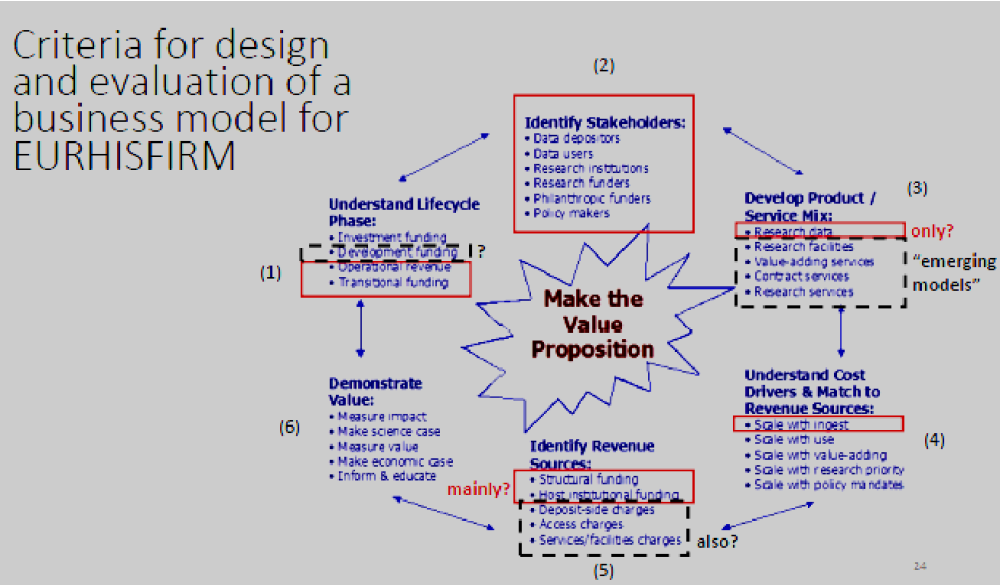
We expect these characteristics will prevail as EURHISFIRM matures, scales up from the design to the operational phase and starts providing an ongoing reliable service to users. However, unlike most existing RIs in social sciences and humanities that work almost entirely with national data, EURHISFIRM will include elements of a federation of national repositories, as research based on its data is inherently international. As a consequence, we expect EURHISFIRM to face challenges similar to those typical of international research data sharing networks, for which, as emphasised in OECD 2017b (p.36)[3], current funding arrangements are inadequate and the multiplicity of funding agencies and schemes required represents "the single biggest barrier to effective data sharing". Our main objective will be to explore the possibility to arrange an international consortium of research infrastructure funding bodies consistent with the characteristics of EURHISFIRM as a domain repository – i.e. a RI that manages data related to a specific field of research in the area of economics and business. We will also explore and discuss the applicability to EURHISFIRM of the ERIC (European Research Infrastructure Consortium) legal framework, as well as the possibility to cooperate with existing RIs operating within the ERIC framework. Since experience suggests that for RIs in social sciences the key challenge is typically fast growth in demand for data curation and sharing, we will pay special attention to develop a business model that is responsive to scaling in the volume of data or number of datasets ingested and hosted, human curation and data preparation. We'll also explore innovative sources of revenues, such as data deposit fees, or charges related to value-adding services provided by EURHISFIRM to other parties or research contracts.

Finally, in the elaboration of EURHISFIRM's business model in this initial lifecycle stage of our RI, we intend to follow the suggested sequence of actions: identify stakeholders, develop the product/service mix, understand cost drivers, identify revenue sources, demonstrate value and make value proposition to

[3] OECD Global Science Forum and CODATA. "Coordination and support of international research data networks". OECD Science, Technology and Innovation Policy Papers, n. 51, December 2017.

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

stakeholders (see Figure 3). The last stage will be especially critical, since, as emphasised in OECD 2017a[4], "engaging and maintaining structural, institutional, philanthropic, or other funders depends on their understanding of the value proposition, and ensuring such engagement may involve repositories undertaking detailed benefit/cost, value, and impact analyses."



IV. Conclusion

At the present stage, we have been able to use the experience of existing RIs to provide a clear picture of the main challenges that EURHISFIRM will have to deal with for the development of a sustainable business plan as it enters the operational phase and starts providing an ongoing service to users. In the next months, we will hire the human resources required to perform the tasks and deliverables assigned to Work Package 10.

[4] OECD Global Science Forum and CODATA. "Business models for sustainable research data repositories". OECD Science, Technology and Innovation Policy Papers, n. 47, December 2017.

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

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

As per EURHISFIRM agenda, Work Package 11 is to officially commence late 2019, so at this stage the Work Package is in a very preliminary stage. Steps taken so far are all in preparation for the official start of the project sections. 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 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 meeting with all group members: To assure transparency and an awareness of EURHISFIRM goals, aims and deadlines a catch up meeting had taken place on January 9, 2019.
    - Attendees: Joost Jonker, Abe de Jong, Coen Fierst van Wijnandsbergen, Joanna Kinga Sławatyniec, Marlies Vreeswijk, and Juan de Weger.
    - The meeting was followed up by a Skype session with Lana Yoo
    - The overarching aim of the meeting was to ensure that all the project requirements are being met and EURHISFIRM deadlines will be fulfilled.
- In addition to the January 9, 2019 meeting, a couple other meetings and Skype sessions were held between key project group members, including:
    - A general introduction session for Joanna Kinga Sławatyniec by Joost Jonker held in Amsterdam on December 10, 2018.

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

- An introduction session for Joanna Kinga Sławatyniec by Coen Fierst van Wijnandsbergen covering steps taken so far in terms of Work Package 9. This took place on December 11, 2018 in Rotterdam.
- 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.
- Other Skype sessions

Preparatory work and research has been started for Work Package 11. Literature on stock exchange markets is being reviewed for an even better understanding of their intricacies, and to ensure a solid fundament for future milestones and deliverables.

III. Progress: Project Achievements

As discussed above, Work Package 11 is currently in a very nascent state. Nonetheless, 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 first milestones and deliverables (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.

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

Conclusions

The project has progressed satisfactorily, particularly after the planned recruitments had been made. Indeed, the various teams faced greater difficulties than expected in recruiting people with the right profiles for the planned objectives. Once the recruitments were completed, the project has been able to structure its activities and to develop the necessary procedures and the appropriate coordination tools. While the first deliverables were submitted with some delays, the subsequent deliverables have adhered to the deadline schedule.

With regard to risk assessment, the Executive Committee identifies three main challenges for the coming year:

1. **Coordination among Work Packages**, which has been satisfactory this year, must be the subject of renewed attention, particularly regarding activities aimed at developing new technologies. During the first year, the project's work progressed according to the established programme. In the second year of the project, this progress will lead the teams to undertake more complex actions that will require a stricter level of coordination than in the first year. In particular, the development of technologies will require enhanced cooperation not only among the IT teams, but also among the IT teams and the economic historians since they are shared within different Work Packages and are carried out by people based in different institutions.

2. **Community building:** Technological success is necessary but not sufficient to make a Research Infrastructure effective. It needs to have enthusiastic users, and it must become sustainable on a stand-alone basis. It is therefore of utmost importance to be able to build and maintain a vibrant community of users. We should therefore address at least the following challenges:

    - 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?

3. **Degree of (de)centralisation of the common data model:** One major system design variable is the degree of decentralisation – and, closely interconnected to this, the depth of system federalism. We explore a range of alternatives between the extremes of highly centralised and highly decentralised systems. A centralised system promises to be efficient with respect to costs of system operations, and many of our stakeholders have fewer challenges to easily comprehend the centralised approach. However, decentralised systems are less vulnerable to single point of failure of a central node.

    Complementing decentralisation with federalism, we see that federated systems also make it possible for regional and national data centres to incorporate local sources of data and information that extend

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

and go beyond the system-wide data models and standards that are common to all nodes in the federated infrastructure.

This ability to maintain core data using common standards in tandem with local sources of data that are specific to a regional node does make the design, operation and maintenance of such a federated system more complex than centralised systems. (This is for instance due to the fact that data is distributed and stored on multiple nodes in a federated system.)

Furthermore, this is by far not just a technical or data management optimisation question. The political processes – and in particular the many 'stumble stones' in the processes of European unification – teach us that only a decentralised and federated system architecture will enable a unified, EU research infrastructure with common access to all data while avoiding the (potentially controversial) need to establish a central data centre in a single national jurisdiction.

Hence, our challenge and our desire is to employ technical and organisational advancements (and accepting higher overall operative costs) to render a substantially higher degree of local (national or regional) freedom while – at the same time, via a common data access service – to use European data in an integrated way.

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