

EURHISFIRM: Historical high-quality company-level data for Europe

Project description

Authors:

Angelo Riva (Professor, European Business School and Paris School of Economics)

Jan Annaert (Professor, University of Antwerp and Antwerp Management School)

Wolfgang König (Professor, Goethe University Frankfurt/Main — House of Finance)

Lana Yoo (EURHISFIRM Project Manager, Paris School of Economics)

Contact: eurhisfirm@eurhisfirm.eu

Complete list of project consortium members: see [page 6](#)

Document last updated on January 2021

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 analyze, develop, and evaluate effective strategies to promote investment and economic growth. To achieve this goal, EURHISFIRM develops innovative tools to spark a “big data revolution” in the historical social sciences and to open access to cultural heritage in close cooperation with existing RIs.

A. Background and rationale

A.1 The need for scientific evidence

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

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



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

A.2 The European empirical shortage

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

So far, only a very few large stand-alone European long-term databases have been built by both the **academic community** (e.g. the London Share Prices Database of the London Business School) and **private companies** (e.g. the US Datastream). Interoperability, if any, remains low among these databases.

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

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

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

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

To summarise, **the current lack of high quality long-term empirical European data prevents the usage and testing of models for analysing structural and cyclical changes, which are crucial for understanding the interactions between financial, economic, and social evolutions.** Creating sound future policy requires the understanding of both past and current dynamics. Creating the data to develop this



knowledge requires sharp interdisciplinary skills, some of which are specific to a country, or even to a region, because of the heterogeneity of historical business rules and practises. These peculiarities call for an ad hoc **Research Infrastructure** that can also connect to other existing systems.

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

B. The project

B.1 The foundations

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

B.2 The concept

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

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



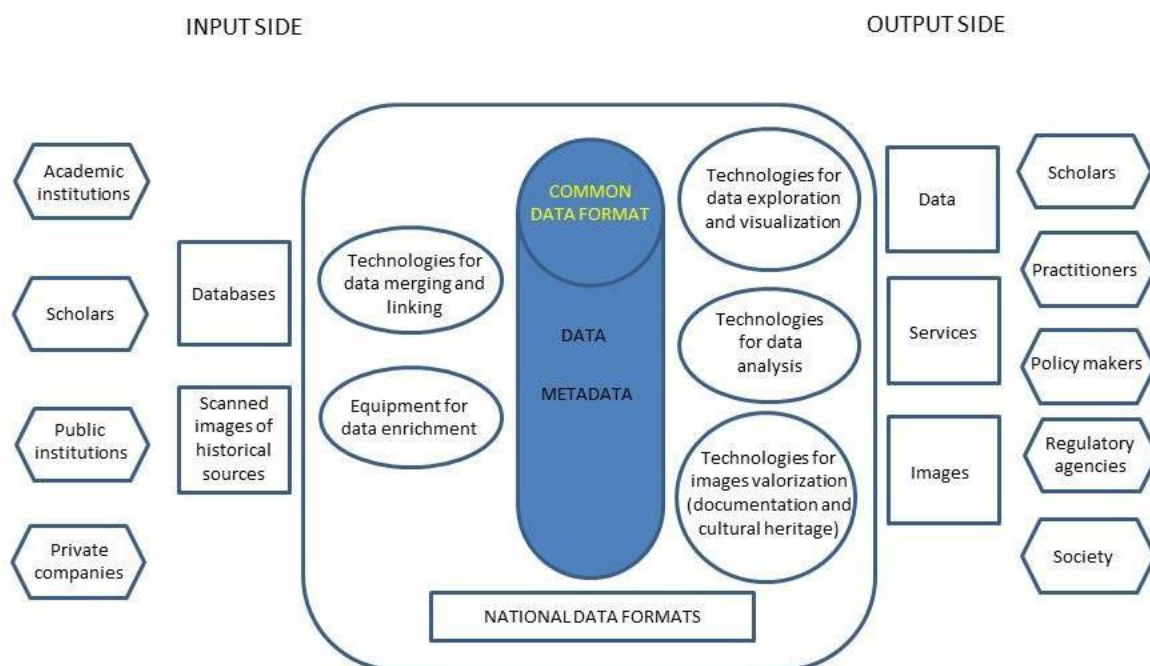


Figure 1 The EURHISFIRM Research Infrastructure Concept

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

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

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

B.3 Methodological approach

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

The **data design** is based on an in-depth **survey and assessment of both the available data and the companies' historical sources (WP4)**. To make the work manageable, the survey will be limited to 19th-

and 20th-century historical printed serial sources on publicly traded companies. Accordingly, **WP5** 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 **WP6**.

European archives and libraries have preserved a wealth of serial printed sources on companies. WP7 will design a set of tools to extract high-quality data from these sources at low costs. Additionally, the web is a mine of scattered and dispersed information on European companies over the long run, and an algorithm will extract and collate this information.

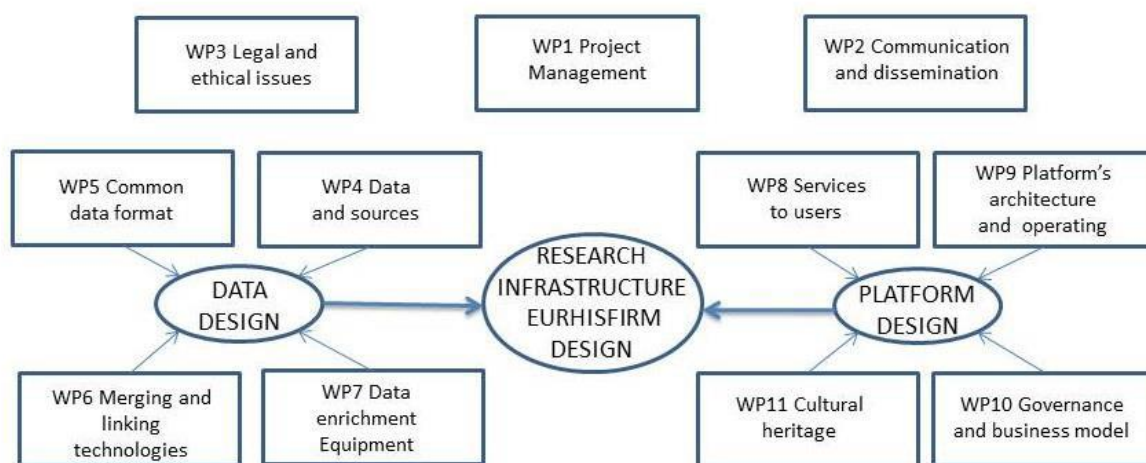


Figure 2 : Methodological Approach to the Concept

The **platform design** focuses on EURHISFIRM’s future **services** to the community: the services are conceived and designed in close cooperation with the stakeholders (**WP8**). The service design will guide the **platform’s architecture and operations** (**WP9**). Tight interconnections with the community and the analysis of related initiatives will drive both the **governance and business model** designs of EURHISFIRM (**WP10**). 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** (**WP 11**).

This approach is supported by the **project management** (**WP1**) in charge of both the overall coordination of the project and the **final design study**; the **communication and dissemination** Unit (**WP2**) for establishing and expanding a **vibrant stakeholder’s community**; and the **legal and ethical unit** (**WP3**) for exploring issues related to the dissemination and use of data and images, partnerships, contracts and the consortium agreement.

EURHISFIRM consortium members

