EURHISFIRM
Research Infrastructure
Historical high-quality firm level data for Europe

H2020 Infrastructure Development Program

Call:
https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/2052-infradev-01-2017.html

Participants:

Participant organisation	Country	Position	Person in charge	Position
SCOB Research Center - Antwerp University	Belgium	National Coordinator	Jan Annaert	National Coordinator
Paris School of Economics	France	EU coordinator, National coordinator	Angelo Riva	EU coordinator, National coordinator
LITIS - Rouen University	France	Member	Thierry Paquet	Team coordinator
IRISA – Rennes	France	Member	Bertrand Coüasnon	Team coordinator
SAFE Research Center - Goethe University	Germany	National Coordinator	Wolfgang König	National coordinator
Goethe University - Library Information System HeBIS / Hessian Library	Germany	Linked Third Party	Uwe Risch	Team coordinator
Amsterdam International Institute for Social History	Netherlands	National Coordinator	Joost Jonker	National coordinator
Department of Finance - Rotterdam School of Management - University of Rotterdam	Netherlands	Member	Abe de Jong	Team Coordinator
Department of Financial Investments and Risk Management - Wroclaw University	Poland	National Coordinator	Krzysztof Jajuga	National Coordinator
Department of Social Sciences - Carlos III University in Madrid	Spain	National Coordinator	Stefano Battilossi	National Coordinator
GESIS	Germany	Member	Oliver Watteler	Team Coordinator
Center for Economic History - Queen's University in Belfast	United KingdomK	National Coordinator	John Turner	National Coordinator

EURHISFIRM aims at designing a world-class research infrastructure to collect, merge, extract, collate, align and share detailed historical high-quality firm level data for Europe. To achieve this goal, it develops innovative tools and sparks the "Big data" revolution in historical social sciences.

The recent financial crisis led the world into what is now called the Great Recession, in reference to the Great Depression of the 1930s. Economic recovery is still nascent in some parts of Europe. Shortcomings in the working of capital markets undermined corporate investment and spread unemployment. Hence, they rose inequalities, harmed well-being of citizens, and built mistrust vis-à-vis decision makers and scientists. These fallouts affected society's innovation and openness. **The European Commission has identified investment, growth, and creation of jobs as the key objectives of its agenda**. To reach these objectives, it brings forward further policy initiatives such as the EU capital markets union to improve access to capital for businesses, especially SMEs.

The European Commission has identified **sound scientific evidence** as a key element of policy-making at all levels of the European process. Yet, the European huge research potential in social sciences has not been entirely realised due to a **lack of empirical works**. The weak empirical foundations of the analytical models used to analyse structural and cyclical changes has become obvious in the fierce debate among scholars following the financial crisis. One of the main reasons for this shortcoming is the **scarcity of detailed historical high-quality firm level data for Europe** available to test these models, which are crucial to understand the interactions between financial, economic, and social evolutions. This scarcity is **particularly glaring at the European level:** policy for the future must be aware of both the dynamics inherited from the past and the directions these dynamics are structuring the present.

The "Big data" revolution in historical social sciences will bring crucial progress in and fundamental revisions of current knowledge of the EU economy. The scaling-up in both variety and size of available historical high-quality economic and financial data for Europe has the potential for a major **epistemological rupture**. On the one hand, datasets are usually build to test models and bring answers to specific questions, but in the **new paradigm** trends and patterns in the data will emerge due to data mining techniques independently from preliminary research questions: **History is a boundless natural laboratory** for a host of economic and financial experiences. On the other hand, the building of historical **"born-on-paper" big data** – as opposite to "born-digital" data – represent a unique occasion to bring down barriers not only between social sciences but also between social and "hard sciences" in the context of **interdisciplinary digital humanities**.

USA has been investing enormous resources to build and link databases suited for research over the long run. The **Collaborative for Historical Information and Analysis** (CHIA) links academic and research institutions to sustain a Human System Data Resource, connecting variables to analyze many areas of human experience. 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 database in finance, contains prices and dividends for shares listed on the New York Stock Exchange since 1926. A Google Scholar search on scientific papers that have used CRSP data returns nearly 46,000 hits including many papers by Nobel prizes. But the use of US data precludes any understanding of the features of the European economy. 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 supposed to be applicable everywhere, generating many biases.

Today, **only a very few large stand-alone databases have been built** so far on Europe by both academic community (*e.g.* the London Share Prices Database of the London Business School) and private companies (*e.g.* the US Datastream), without any concern for interoperability. Within the **academia**, considerable resources have been devoted to data building, very often with the aim to study very specific issues. Such datasets are without any systematic comparative or diachronic analytic purpose and have no concept for cumulativity or sustainability. Hence, it is nearly impossible to compare the existing fragmented datasets. Moreover, a continuing access is not guaranteed, the data dissemination being often left at discretion of these individuals. Consequently, **with no permanent infrastructure that cares about harmonization and access, these data are in most of the cases lost to the community**.

On the other hand, the very few historical series that are contained in some **commercial databases** are sometimes unsuitable for research, but daily used by business and academia. They can give rise to serious errors, because they are poorly documented and flawed because based on easy-to-find but inappropriate sources. The building of such data 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 (RI) able to connect to other already existing ones.**

The **EURHISFIRM project** meets the need for such a **benchmark research infrastructure** in Europe. It will operate **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, location of headquarters. The creation of a **vibrant European community** will support the development of **revolutionary RI technology**, which in turn will enable a scientifically reproducible, technically sound and socio-legally robust **evidence-base for the stakeholders**. **Not only policy makers and scholars will benefit from, but notably private companies**: on the one hand, companies are major data users, the global spend for market data, an industry where US holds a quasi-monopoly, amounting to nearly 30 billion of dollars in 2015 (Burton and Taylor, 2016); on the other hand, the disruptive technologies developed within the RI will push further the technological frontier and bring major spin-offs to the European IT industry.

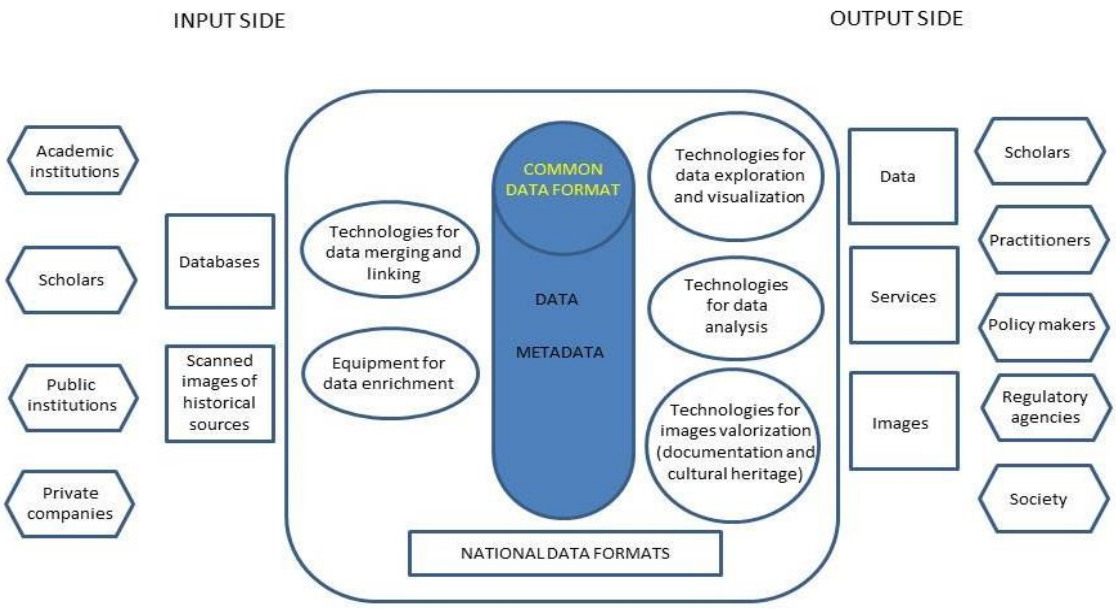
This project stems from **the experience of the research group Eurhistock** that brings together specialists in economic and financial history on a yearly basis since **2009**. This group acknowledged both the incompleteness of the existing datasets, the fragmentation of the initiatives, and the heterogeneity of the data collection practices in Europe. This observation led some countries like France and Belgium to put in place coordinated initiatives to build long-run **structured data with digital techniques at the technological frontier**[1]. Other countries in the consortium have started to collect data or are reflecting on the comparative issues of their datasets.

The **concept of a RI** developed in this project relies on innovative technologies to collect, merge, extract, collate, align and share **detailed historical high-quality firm level data for Europe** (Figure 1). Concerning the **input side**, the RI will develop innovative data connecting and extracting tools. The project develops technologies to **merge** existing historical high-quality data and **link** them to other

[1] for Belgium, the SCOB database of the University of Antwerp (http://www.scob.be/what.html); for France, the Equipex DFIH "Data for Financial History" of the Paris School of Economics (http://www.parisschoolofeconomics.eu/en/grand-emprunt/equipex-d-fih-donnees-financieres-historiques); for Germany, the Goethe project (http://safe-frankfurt.de/research/research-projects/research-projects-details/controller/project/projectname/historical-german-stock-market-database-goethe.html)

historical and contemporary databases. To **enrich** existing data**,** the project develops a **data enrichment-equipment** that sparks the **"Big Data Revolution" in the field of historical social sciences**. In spite of technological advances, data from historical sources are still manually collected, an expensive and erratic procedure. This path-breaking equipment will collate data from **the web** and extract data from **scanned images of historical printed serial sources**. It will bring down the cost of data collection and improve data quality.

Figure 1 The concept of EURHISFIRM Research Infrastructure



Common format and semantics safeguard the coherence of the data. They require a harmonization process that gradually transforms local and national heterogeneities – due to institutional differences or simply to data ownerships - into shared standards. To handle the complexity, data formats and semantics are to be set up first at country levels within the national focus points (consortium's national coordinators) in close cooperation with national communities, and then processed in a recursive way toward European norms.

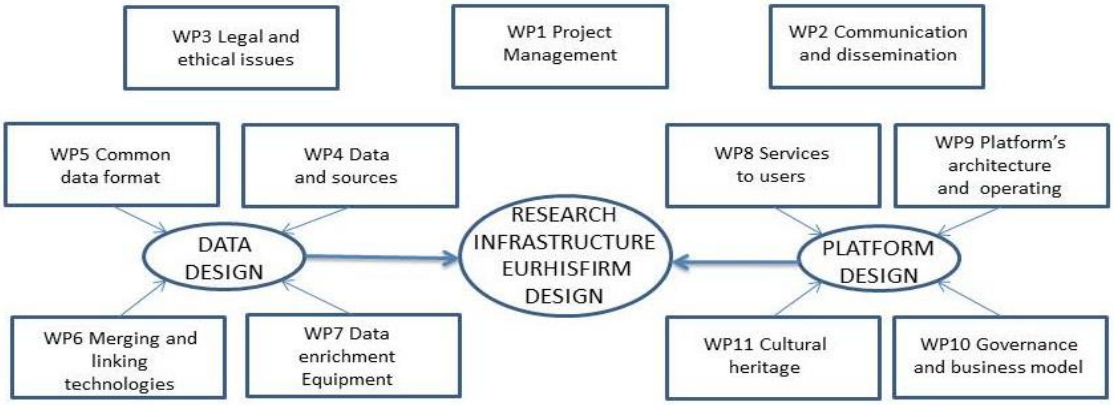
Concerning the output side, the RI offers to the stakeholders' community **data**, **services** and images that contribute to the **European cultural heritage**. It develops differentiated **technologies to explore and visualize** large and complex amounts of financial data in a user-friendly manner, making information easily accessible to both experts and citizens. It develops **technologies for data analysis and mining**. It makes available knowledge, data connecting and extraction technologies in order to inspire new data collections, particularly by **young scholars**, and make EURHISFIRM an expanding community. It provides **images of historical sources** to supply an exceptional historical **documentation** for the data and preserve the **European cultural heritage**.

Data standards, services to users and principles to prioritize data merging, collating and collecting are jointly determined with the **community of stakeholders**, within the **Research Data Alliance** context. Notably, according to our vision, EURHISFIRM is the required and complementary extension of two existing RI in the social sciences area, CESSDA and DARIAH. EURHISFIRM could develop and extend the **historical dimension of ESFRI CESSDA**: the planned RI has the potential to go well beyond its original focus on companies and encompass broadly economic and social quantitative data. The

EURHISFIRM's first choice to value the images of historical sources is a **partnership with the ERIC DARIAH**.

The **methodological approach to the design of the RI** is based on the integrated 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 available data and historical sources on companies** (**WP4**). In the design study, the circumscription of the survey to post-1815[2] historical printed serial sources related to publicly traded companies is required to make the WP manageable. Accordingly, **national data standards and semantics** are developed and harmonized in the process towards a **European common data format** (**WP5**). This convergence enables the development of **technologies to spark "Big Data Revolution" in historical sciences and push further the technological frontier. Technologies to merge** historical high-quality data **and link** them to other historical and contemporary databases are developed (**WP6**). An **equipment to enrich data** is designed (**WP7**). European archives and libraries preserve a wealth of serial printed sources on companies. An equipment to extract high-quality and low-cost data from these sources will be designed. The web is a mine of scattered and dispersed information on European companies over the long run: an algorithm will extract and collate this information.

Figure 2 Methodological Approach to the Concept



The **platform design** is focused on the **services** the RI will supply to the community: the services are conceived and designed in close interactions with the stakeholders (**WP8**). The design of the services will guide the **platform's architecture and operating** (**WP9**). Tight interconnections with the community and the analysis of other experiences will drive the design of both the **governance and business model** of the RI (**WP10**). Images produced to nurture the equipment will constitute at the same time an exceptional source of documentation for the data and a precious contribution to the valorisation of 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.

[2] In 1815, the Congress of Vienna brings together the European powers to create a new order after the defeat of Napoleon.