Long-term data for Europe

EURHISFIRM

D1.2 (version 1, M6): Data Management Plan

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

AUTHORS:

Elisa GRANDI *(École d'Économie de Paris)*

Angelo RIVA *(École d'Économie de Paris)*

Lana YOO *(École d'Économie de Paris)*

APPROVED IN 2018 BY:

Jan ANNAERT *(Universiteit Antwerpen)*

Wolfgang KÖNIG *(Goethe-Universität Frankfurt)*

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

Table of Contents

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

Revision history

Version	*Date*	*Notes*
1.0	12/09/2018	First draft
1.1	21/09/2018	First corrections
1.2	24/09/2018	Second corrections
1.3	28/09/2018	Third corrections (final version of version 1)

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

Executive summary

EURHISFIRM "Historical high-quality company-level data for Europe" is a design study to build a world-class research infrastructure (RI) compliant to the FAIR (findable, accessible, interoperable, reusable) data principles. The project aims to increase the accessibility and usability of historical company-level data (financial, governance, and geographical) and to expand the available pool of this data. At the data and platform levels of the RI, the design study (1) provides the architecture for FAIR long-run European company-level data enabling the users to connect and combine information from different sources; (2) develops an intelligent and collaborative system for the extraction and enrichment of data, either from historical paper sources or from web-based resources; (3) develops and maintains data quality standards and models for collecting, matching, and connecting data on a European scale. The focal point of the RI will be the integration of financial and corporate governance information with data on the location of firms, reflecting, over the long run, the interaction between financial markets and the real economy. To achieve this, the project will execute a number of different data studies, such as selecting the appropriate metadata standards, evaluating possible sources for current and future studies, establishing a common data model based on an in-depth study of data semantics, and testing the technology for digitalising printed data.

Much of European historical company-level data does not yet exist in findable, accessible, interoperable, or reusable data formats. EURHISFIRM envisions to design the RI to make this possible. EURHISFIRM itself will enable the creation of a data management plan on historical company-level data. This document therefore will evolve significantly with each version update. As such, the design study is in itself a data management plan.

The EURHISFIRM project will be run as much as possible in cooperation with existing infrastructures in the field of social sciences and humanities, such as DARIAH and CESSDA.

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

1. Data Summary

1.1 Purpose of the data collection/generation and data utility

The recent economic crisis, usually called the Great Recession, has drawn comparisons with the Great Depression in terms of economic and historical impact. While the causes are complex and spreads across various social, historical, economic factors and beyond, examining the financial markets remains a high priority to fully understand the cause and effects[1]. Growth, investment, and job creation are the key challenges facing the European Union. To take up these challenges, the European Commission is promoting further policy initiatives such as EU-wide capital markets and a Banking Union to improve business access to capital, ensure financial stability, and boost investment and innovation. Economic research, government policy, and society as a whole must possess the data necessary to understand the dynamics of past performance and the way those dynamics structure our present and future. This is why the EU Horizon 2020 Program addresses inclusive long-term growth, as well as reversing social inequality to foster a social and economic framework that promotes sustainability in Europe[2]. Yet, the crucial historical understanding of our society remains totally inadequate, because we lack the requisite empirical basis. The weak empirical foundations of the models used to analyse structural and cyclical changes have become obvious in the recent fierce debates on how to foster economic growth and job creation. One of the main reasons for this uncertainty is the lack of high-quality, long-term and FAIR data on European companies for testing these models. Most of the European scholars rely on the American financial micro-databases. The most widely used database is produced by the CRSP (Center for Research in Security Prices), a production platform managed by Chicago University (http://www.crsp.com/). Its extensive use applied to Europe precludes any understanding of the peculiarities of the European markets and economies, and hinders the development of professional, analytical models and financial products tailored to them.

A few large stand-alone databases have been built by both the academic community and by private companies, but that has been done without any concern for interoperability. Within academia, considerable resources have been devoted to the construction of historical datasets, as often as not 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 lack any systematic comparative or diachronic analytical purpose[3]. The Strategy Report on Research Infrastructure identifies Big Data in the social sciences and the humanities as the first science driver for these fields.[4] FAIR data change the way for carrying out academic research. In spite of the crucial advances “born-digital” big data can bring, they still lack the historical depth that “born-on-paper” data can provide. European cultural heritage represents a shared wealth in terms of citizenship, cultural growth, and economic potential. Hence, the Strategy Report identifies the emerging need and opportunity for research infrastructures (RIs) providing access to this heritage and innovative technologies to analyse and integrate extracted

[1] http://ec.europa.eu/economy_finance/publications/pages/publication15887_en.pdf
[2] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/europe-changing-world-inclusive-innovative-and-reflective-societies
[3] Wilkinson, M. D., Dumontier, M., Aalbergsberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 3(160018). doi:10.1038/sdata.2016.18
[4] https://ec.europa.eu/commission/sites/beta-political/files/5-presidents-report_en.pdf

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

information to large stakeholders' communities. The EURHISFIRM "Historical high-quality company-level data for Europe" project addresses this need with a comprehensive study of: investigating the historical sources available, designing the correct standardisation methods, as well as creating the optimal infrastructures and technology. The design study will be compliant to the FAIR (findable, accessible, interoperable, reusable) data principles.

The design study, and the data subsequently generated with the resulting infrastructure, will be useful to various types of organisations with vested interests in the European economy: governmental, academic/scientific, public and other non-profit, as well as private entities.

1.2 Origin, types, formats, and size of data generated/collected

As an infrastructure, EURHISFIRM designs a world-class RI to connect, collect, collate, align, and share detailed, reliable, and standardised long-run company-level data for Europe. The goal is then to provide an infrastructure to extract and enrich data as well as to connect, collate, and align existing and new data.

To make the work manageable, the EURHISFIRM design study first focuses on stock exchange-listed companies because they are larger and better documented. An in-depth analysis of existing company-level data and historical serial sources is carried out for three main types of information related to firm characteristics: a) financial data (stock market data such as securities issued, prices, dividends and coupons, number of traded securities, corporate events such as (reverse) splits, mergers, balance sheets and income statements); b) information on the companies' governance (e.g. evolution of the juridical status, directors, voting and governance rules), and geographical data (e.g. location of headquarters, subsidiaries, and production units).

The origin, types, and size of the existing sources and data will become clearer as the project progresses. WP4 prepares an inventory of existing data and sources. It delivers in-depth knowledge on the type, quality, and other key characteristics of the available data and sources such as yearbooks of companies and stock exchange lists. It produces accurate data and sources documentation according to the chosen common documentation standard (DDI Lifecycle [https://www.ddialliance.org/], see Deliverable D4.1). Specific attention will be paid to data semantics, a scientific challenge particularly relevant to historical data. WP4 explicitly recognizes the methodological challenge of ensuring standardised approaches whilst allowing for idiosyncrasies of the diverse data types from various countries across time.

The project concerns two main data formats: digitised (stored in databases) and raw (not yet digitised and not yet transformed into databases). A few large stand-alone long-term databases have been built by both the academic community and by private companies (e.g. the London Share Prices Database, the Global Financial Data database)[5], but interoperability remains low. Within academia, considerable resources have been devoted to the construction of historical datasets, as often as not 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). Accordingly, national data standards and semantics

[5] It is worthwhile to note the exceptions of the SCOB database at the University of Antwerp and the Data for Financial History Database at the Paris School of Economics which have been built in a coordinated way (both institutions belong to the EURHISFIRM consortium).

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

are developed and harmonised in the process towards a common European data format within WP5. Technologies to merge historical high-quality data and to link them to other historical and contemporary databases are developed by WP6. A platform based on OCR and AI to extract and enrich the data is designed within WP7. The web is a mine of scattered and dispersed information on European companies over the long run: within this framework an ad hoc system will extract and collate this information.

1.3 Re-usage of existing data

A few large stand-alone long-term databases have been built by both the academic community and by private companies (e.g. the London Share Prices Database, the Global Financial Data database)[6], but interoperability remains low. Within academia, considerable resources have been devoted to the construction of historical datasets, as often as not 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). The main goal of EURHISFIRM is twofold: 1) designing the infrastructure to be used by academics and other stakeholders to deposit and connect their data, as well as to 2) inspire new projects of data collection with the next-generation data extraction and enrichment platform developed from EURHISFIRM.

Many members of the consortium have already run extensive data collections. All of them are committed to reverse and integrate their data into the infrastructure to create a first pool of data big enough to raise interest within the "data collectors" community. This gravitational pull will attract already existing data to make them re-usable within the infrastructure once they are documented according to the established data format.

2. FAIR data

To ensure that EURHISFIRM's final output will be a solid, federated RI design consistent to FAIR principles, a working group (Work Group on Identification and Standardisation) has been formed by all interested representative members from all of the WPs. The group has adopted The Open Group Architecture Framework (TOGAF) (http://theopengroup.org/), an enterprise architecture framework that provides a set of standardized guidelines that serves this purpose.

2.1 Making data findable

Data are only useful if they are discoverable and useful to the relevant users. These can depend on two factors: good organisation via metadata usage and the location (and availability) of the data storage.

In order to render the data findable for future users, the EURHISFIRM design study must select the appropriate metadata format. A number of standards have been under study by WP4 and WP5, and the optimal method for the type of data EURHISFIRM envisions has been chosen as the DDI Lifecycle (https://www.ddialliance.org/) due to its compatibility with historical datasets, especially in dealing with

[6] It is worthwhile to note the exceptions of the SCOB database at the University of Antwerp and the Data for Financial History Database at the Paris School of Economics which have been built in a coordinated way (both institutions belong to the EURHISFIRM consortium).

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

various elements of the data that may change in format and content over time[7]. Unique identifiers will be assigned to datasets stored within the infrastructure and updated in case of several versions of the concerned dataset. Search by k-words will be provided.

2.2 Making data openly accessible

EURHISFIRM aims to design an RI of open-access data under the constraint of a sustainable business model developed by WP10.

As a research infrastructure, EURHISFIRM aims to become the reference *repository location for historical company-level data*. The *software, method(s)* and possibly *license(s)* required will be decided with the project progression, based on their abilities to provide open access to potential users under the constraint of a sustainable business model. In-depth documentation about the software used to access the data will be included. As much as possible, the relevant software produced within the EURHISFIRM project will be open source, under the constraint of the consortium agreement clauses concerning software produced by members before the EURHISFIRM project. The data, associated metadata, documentation and code will be deposited within the EURHISFIRM infrastructure.

The very nature of the concerned data does not seem to require a data access committee. However, to ensure ethical use of data, as well as to comply to the General Data Protection Regulation (GDPR)[8], the design will enforce that any data that may reveal personally identifiable data will be properly handled. The final design could, therefore, mandate that such data would be anonymised or even be made inaccessible. In particular, the governance data may be particularly sensitive to this issue, as they may contain personally identifiable information of company individuals.

As the data sources become more detailed throughout the project progression, further details on the plans for data accessibility will be elaborated especially with WP3's work on data privacy and information protection, while a system to record data downloads will be elaborated within WP9.

2.3 Making data interoperable

As the interoperability of historical European company-level is currently low, EURHISFIRM aims to create an RI design to specifically overcome this obstacle.

The reasons for the low interoperability are:

- For both digitised and non-digitised data: different languages, types of markets, formats, normalisation of changes over time e.g. company name evolutions and/or M&As
- Digitised data: stored in multiple databases in various data formats, which increases the incongruency of the data and therefore makes analysis and cross-comparison difficult

[7] See deliverable *D4.1: Information system and documentation standards* (author: J Poukens)
[8] https://ec.europa.eu/info/law/law-topic/data-protection_en

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

- Non-digitised data: in printed format. As they are not digitised, searching, analysis and cross-comparisons are extremely cumbersome

In the EURHISFIRM project, WP5 will create a common data model to overcome these interoperability challenges in historical European company-level data. The current model developed by WP5 so far describes a system in which the local data sources (pink layer) will be treated through data integration gateways (yellow layer) and then integrated within a common access system through which data users can consume the data (green layer). The details of these processes are currently under discussion within the WP5 tasks.

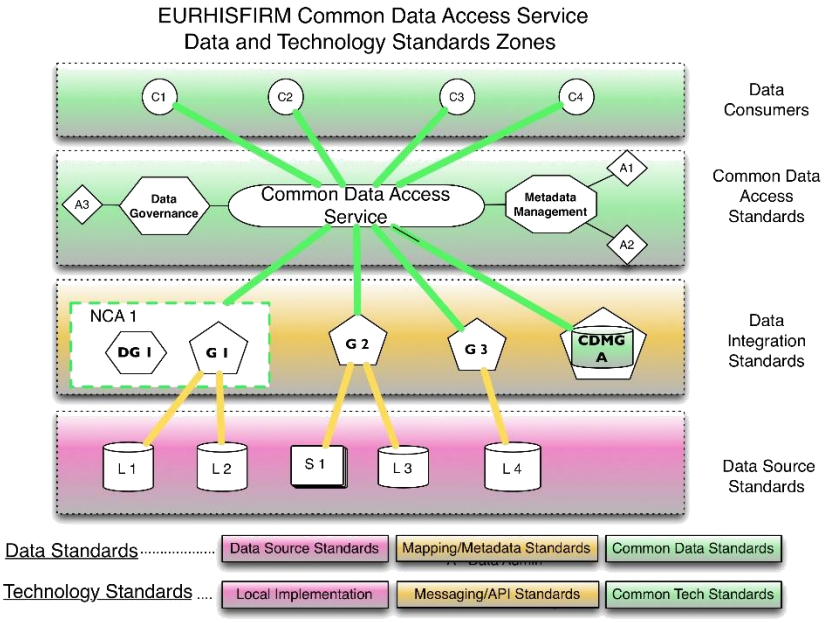

Figure 1: proposed Common Data Access Service (provided by W König (GUF, WP5))

As mentioned above, WP6 will also study the interoperability between the existing data, as well as their compatibility with the to-be digitised data from the other partner countries in the project. Eventually, based on the models resulting from this design study, the ideal goal would be to increase the interoperability of historical financial data with as many other European countries as possible.

2.4 Increase data re-use (through clarifying licences)

Reusability remains a priority for historical European long-term company-level data. If data (particularly historical/long-term data) are not re-usable, then reiterative comparisons as well as innovative data use would not be possible. EURHISFIRM's design study aims to permit the creation of data that could be re-used by other institutions and individuals. To make this possible, the first 3 criteria mentioned above (findable, openly accessible, and interoperable) must be fulfilled. As mentioned before, WP5 and WP6 will be the main actors to complete these tasks.

WP8 will also survey the interested stakeholders in academia, business, and policy in order to understand the users' needs and interests concerning long-term financial company-level data. As these survey

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

respondents come from diverse fields, understanding their data needs will ensure that the RI design will consider its usage across domains in the long term, ensuring that the data could be re-used in an openly accessible manner to serve various purposes in the future. WP9 will also study and recommend sound infrastructure policy and architecture that would most optimally support the sustainability, and therefore the reusability, of the data. Additionally, WP11 will also explore on the ways that digitised historical financial data can be used to promote and deepen European research, culture, and heritage.

The legal details concerning the data license will become more known with the progression of WP3's work.

3. Allocation of resources

Regarding financial resources, the costs for conforming to the FAIR principles established within the EURHISFIRM project are under calculation. In case of data that do not conform to the FAIR principles, the costs will be paid by institutions willing to deposit data. EURHISFIRM's goal is to minimise these costs.

Regarding the data management plan, the responsibility is held by all of the consortium members, but the final decisions are approved by the Executive Committee with input from the Steering Committee, the General Assembly, and the Project Advisory Board.

Data preservation policy will be decided by the governance structure of the EURHISFIRM infrastructure as designed within the design study (WP10). Working hypothesis on this policy will be laid down by WP9.

4. Data security

Although EURHISFIRM designs the RI by envisioning an open-access data system within a sustainable business model, proper data security measures are high priorities to ensure that the data are used and maintained with proper handling. These issues are to be examined by the WP9 in particular to be agreed upon and to establish the proper technology infrastructure. WP9 is responsible for the data access, security, and maintenance. WP9 will also design the infrastructure policy and architecture, with anticipated contributions from the IT collaborators from WP5 and WP6.

These policies are to be studied and elaborated in future versions of this document.

Note on this data management plan

This data management plan is based on the European Commission's Horizon 2020 Data Management Plan template (http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm).

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

References

DDI Alliance. (2014). *DDI Lifecycle 3.2*. Retrieved from http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/

Directorate-General for Economic and Financial Affairs of the European Commission. (2009). *Economic Crisis in Europe: Causes, Consequences and Responses.* Retrieved from http://ec.europa.eu/economy_finance/publications/pages/publication15887_en.pdf

European Commission. (2018). *Data protection*. Retrieved from https://ec.europa.eu/info/law/law-topic/data-protection_en

European Commission. (n.d.). *Data management*. Retrieved from Research & Innovation Participant Portal H2020 Online Manual: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

European Commission. (n.d.). *Europe in a changing world - Inclusive, innovative and reflective societies*. Retrieved from Horizon 2020: https://ec.europa.eu/programmes/horizon2020/en/h2020-section/europe-changing-world-inclusive-innovative-and-reflective-societies

Juncker, J.-C., Tusk, D., Dijsselbloem, J., Draghi, M., & Schulz, M. (2015). *Completing Europe's Economic and Monetary Union.* Retrieved from https://ec.europa.eu/commission/sites/beta-political/files/5-presidents-report_en.pdf

Poukens, J. (2018). *EURHISFIRM D4.1: Information system and documentation standards.* University of Antwerp, Antwerp.

Wilkinson, M. D., Dumontier, M., Aalbergsberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data, 3*(160018). doi:10.1038/sdata.2016.18

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