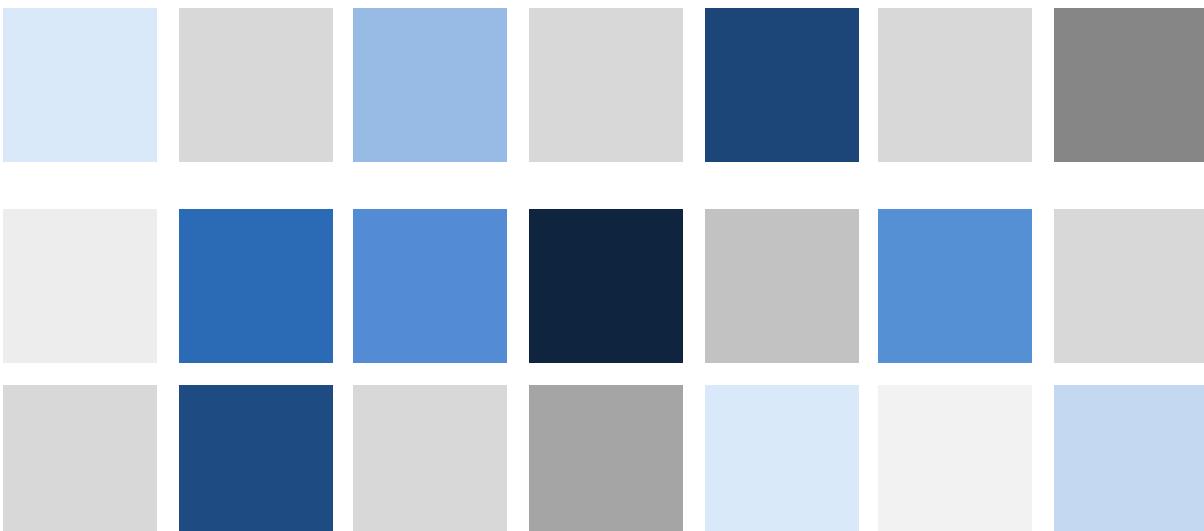


Long-term data for Europe

EURHISFIRM

D1.11: Final 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

<https://eurhisfirm.eu>

Authors:

EURHISFIRM project participants

Approved in 2021 by:

Jan ANNAERT (University of Antwerp)

Wolfgang KÖNIG (Goethe University Frankfurt)

Angelo RIVA (Paris School of Economics)



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

<https://eurhisfirm.eu>

Table of contents

| | |
|--|----|
| I. Executive summary | 6 |
| II. Introduction to the EURHISFIRM project | 7 |
| II.1. The rationale behind EURHISFIRM's creation: to address the lack of a comprehensive European historical financial database and its synchronization with modern financial data | 7 |
| II2. EURHISFIRM and the RI communities | 9 |
| III. The basic principles of the EURHISFIRM design (the work achieved) | 10 |
| III.1. Data extraction technology, transformation, and integration (WP6, WP7) | 10 |
| Further details on the work accomplished in WP7 | 11 |
| Further details on the work accomplished in WP6 | 18 |
| III.2. Meta data, the common data model within a federated system, and technical architecture (WP4, WP5, WP9) | 20 |
| Data inventory and meta data standards selection | 20 |
| The Common Data Model within a federated system | 21 |
| Technical architecture | 27 |
| III.3. Future users, community building, exploitation/dissemination (WP1, WP2, WP8) | 28 |
| Research on future users | 28 |
| Outreach and community building | 29 |
| III.4. Governance and cultural heritage (All Work Packages, and more specifically WP1, WP3, WP9, WP10, WP11) | 33 |
| Adherence to FAIR data principles | 33 |
| Business and governance model/sustainability | 34 |
| Legal governance | 36 |
| Cultural heritage | 39 |
| Memorandum of understanding within the current EURHISFIRM members | 40 |
| IV. Expected impact of EURHISFIRM | 41 |
| IV.1 Contributions to the FAIR data principles | 41 |
| IV.2 Contributions to research, policy making, and socioeconomic impact | 42 |
| V. Conclusions and future directions | 43 |
| VI. REFERENCES | 44 |
| Reports | 44 |
| Websites | 45 |
| VII. APPENDIX | 46 |

| | |
|--|----|
| VII.1. Quick Installation Guide (slide deck) for amending existing files of historical firm data by the EURHISFIRM Legal Entity Identifier (LEI) | 46 |
| VII.2. Memorandum of understanding within EURHISFIRM members | 50 |
| VII.3. Letters of interest in joining future phases of EURHISFIRM from external institutions | 54 |



List of terms and acronyms

| | |
|----------------------|--|
| CDAS | Common Data Access Service |
| CDM | Common Data Model |
| CHIA | Collaborative for Historical Information and Analysis |
| <u>CLARIN</u> | European Research Infrastructure for Language Resources and Technology |
| CRSP | The Center for Research in Security Prices |
| <u>DARIAH</u> | Digital Research Infrastructure for the Arts and Humanities |
| DCU | Data Collection Unit |
| DDI | Data Documentation Initiative |
| DSU | Data Submission Unit |
| EABH | European Association for Bank and Financial History |
| EFII | EURHISFIRM Financial Instrument Identifier |
| ELEI | EURHISFIRM Legal Entity Identifier |
| EOSC | European Open Science Cloud |
| FAIR | Findable, Accessible, Interoperable and Re-usable |
| LEI | Legal Entity Identifier |
| NIC | Network Integration Center |
| RI | Research infrastructure |
| SSHOC/SSH Open Cloud | Social Science and Humanities Open Cloud |
| WGIS | Working Group on Identification and Standardisation |
| WP | Work Package |
| WRDS | The Wharton Research Data Services |

I. Executive summary

EURHISFIRM “Historical high-quality company-level data for Europe” was a design study to build a world-class research infrastructure (RI) compliant to the FAIR (findable, accessible, interoperable, reusable) data principles. The project aimed 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:

- Provided the architecture for FAIR long-run European company-level data enabling the users to connect and combine information from different sources;
- Developed an intelligent and collaborative system for the extraction and enrichment of data, either from historical paper sources or from web-based resources;
- Developed and maintains data quality standards and models for collecting, matching, and connecting data on a European scale.

The focal point of the RI was 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 executed 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, testing the technology for digitising printed data and putting them into a federated database system, connecting the existing databases as well as the potential for linking with future ones, surveying stakeholders on their preferences concerning the RI’s characteristics, assessing the appropriate business model, designing the architecture and security systems, analysing the potential for cultural heritage valorisation, as well as examining the ethical implications of data privacy rights. The project also completed other operational tasks such as outreach and communication, as well as project management and adherence to the FAIR (Findable, Accessible, Interoperable and Re-usable) principles outlined in the data management plans. In the subsequent phases, the project plans to further concretise this design in order to connect, collect, collate, align, and share detailed, reliable, and standardised long-term company-level data for European stakeholders: policy makers, scholars, and private companies.

Central to the EURHISFIRM design was its commitment to fully integrate into the broader European RI ecosystem. It strived towards active involvement in European research infrastructure developments in order to integrate the latest European research technologies and to exploit increasing network effects in Europe. Furthermore, it set-out to build a scientific community based on open science principles to encourage cross-collaboration among researchers in order to enrich scientific contributions and advancements in the social sciences. To this end, the project recognised the importance of collaborating with existing RI projects, policy initiatives (e.g., the European Open Science Cloud (EOSC)), social science

data communities, and other pan-European and national structures in order to ensure the enrichment and long-term sustainability of European RI research.¹

In its subsequent phases, EURHISFIRM plans to initiate or continue further working relationships with other pertinent organisations and developments, in particular with CESSDA ERIC (which is a member of the EURHISFIRM consortium) and the SSH Open Cloud (to which EURHISFIRM contributes). EURHISFIRM will also strive to use these collaborative experiences to establish partnership models of architectural frameworks within RI communities. Such standards would encourage further community-building activities within future RI projects, aligning with the vision of a vibrant and dynamic European research data community. In this regard, it should be noted that a growing number of parties (e.g. NEDHISFIRM and the “Taking Stock. The Amsterdam exchange, investor behaviour, and Dutch economic growth, 1870-1940” projects [see section [II.2](#) for more information] as well as [VII.3 Letters of interest](#)) have expressed concrete interest in collecting historical data and integrating this data in the EURHISFIRM RI.

II. Introduction to the EURHISFIRM project

II.1. The rationale behind EURHISFIRM’s creation: to address the lack of a comprehensive European historical financial database and its synchronization with modern financial data

In light of the major economic recessions in the past decade, as well as current global challenges, the key societal issues facing the European Union are investment, growth, and job creation. 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 meet these urgent social and economic challenges, the European Union needs sound scientific evidence.

Big data offers 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 affect the present and the future. However, because we lack these empirical foundations, this crucial historical understanding of our society remains unfulfilled.

Europe’s huge research potential in the social sciences has not been entirely realized due to a lack of empirical work. The scarcity of long-term (micro)data is particularly notable at the European level. So far, only a very few large stand-alone European long-term (micro)economic databases have been built by both the academic community (e.g. the London Share Price Database of the London Business School) and

¹ These include but are not limited to Huma-Num (the French member of DARIAH ERIC (Digital Research Infrastructure for the Arts and Humanities)), CESSDA ERIC (Consortium of European Social Science Data Archives) (the latter which is a joining member of the EURHISFIRM consortium), CLARIN (European Research Infrastructure for Language Resources and Technology, and the SSHOC (Social Sciences & Humanities Open Cloud).

private companies (e.g. Datastream, commercialised by Refinitiv, based in the UK). Interoperability, if any, remains low among these databases.

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. Digitized historical sources as part of the European cultural heritage represent a shared wealth in terms of citizenship, cultural growth, and economic potential.

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), as 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 US 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 US's dominant position in data production, American companies are frequently and implicitly deemed "representative" or "the norm". Lessons are consequently drawn from their behavior that are supposedly—but are not—applicable everywhere (including Europe), generating issues with research validity and possibly incorrect conclusions.

To summarize, the current lack of high quality long-term empirical European data prevents the usage and testing of models for analyzing 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. There is also a strong need to enable studies on long time series in settings with a low signal-to-noise ratio, such as on financial markets. Creating the data to develop this knowledge requires various interdisciplinary skills, some of which are specific to a country, or even to a region, because of the heterogeneity of historical business rules and practices. These peculiarities call for an ad hoc research infrastructure (RI) that can also connect to other existing systems.

The EURHISFIRM project was conceived to meet the need for such a benchmark RI in Europe. During the course of the project, it designed the framework for a comprehensive long-run economic and financial database to serve as the model for future data infrastructures in the social sciences. The data handled in the design includes information on historical European companies such as accounting, funding and investment, stock exchange data; governance rules; directors; patents; and headquarter locations.

In the subsequent phases, the project plans to concretize this design in order to connect, collect, collate, align, and share detailed, reliable, and standardized long-term company-level data for European stakeholders: policy makers, scholars, and private companies.

II2. EURHISFIRM and the RI communities

Central to the EURHISFIRM design is its commitment to fully integrate into the broader European RI ecosystem. The project recognizes the importance of collaborating with other RI projects, policy initiatives (e.g. the European Open Science Cloud (EOSC)), social science data communities, and other pan-European and national structures in order to ensure the enrichment and long-term sustainability of European RI research.

To that end, EURHISFIRM has engaged with pertinent organizations throughout the project. These include:

- Integration of [CESSDA ERIC](#) in 2020 as a EURHISFIRM consortium member. CESSDA ERIC's expertise in the open data and research infrastructure ecosystem, particularly in the social sciences, enhances the project's technical and scientific competencies and ensures that the project's research infrastructure design incorporates cutting-edge developments from the European data community
- EURHISFIRM's participation in the [EOSC](#) ecosystem via integration into the Social Science and Humanities Open Cloud ([SSHOC](#)) project as a consortium member (formally represented by the three executive committee institutions [Paris School of Economics, University of Antwerp, and Goethe University Frankfurt-SAFE]).
- Collaboration with [CLARIN](#) (European Research Infrastructure for Language Resources and Technology) on analysis of historical resources and the interoperability of heterogeneous data
- Partnership with [Huma-Num](#), the French branch of [DARIAH](#) (Digital Research Infrastructure for the Arts and Humanities), for EURHISFIRM's digital hosting needs (website, data storage, etc.).

It is also worth mentioning that there is a continued interest in collecting national historical financial data in the research community. For example, the [NEDHISFIRM](#) project (collection of Dutch historical financial data by EURHISFIRM researchers and other collaborators) has recently been awarded a grant by the Platform Digital Infrastructure SSH. The team will also receive additional funding from NWO (the Dutch Research Council) to supplement the NEDHISFIRM project with three research projects relating to these topics, collectively under the name of "Taking Stock. The Amsterdam exchange, investor behaviour, and Dutch economic growth, 1870-1940".

In its subsequent phases, EURHISFIRM plans to initiate further working relationships with other pertinent organizations and developments. EURHISFIRM will also strive to use these collaborative experiences to establish partnership models of architectural frameworks within RI communities. Such standards would encourage further community-building activities within future RI projects, aligning with the vision of a vibrant and dynamic European research data community.

III. The basic principles of the EURHISFIRM design (the work achieved)

In practical terms, the project aimed to design the end-to-end process of the RI design: from the digital extraction (i.e. digitization of scanned data from physical paper sources) of financial historical sources until the final step of allowing access for future users to these digitized data.

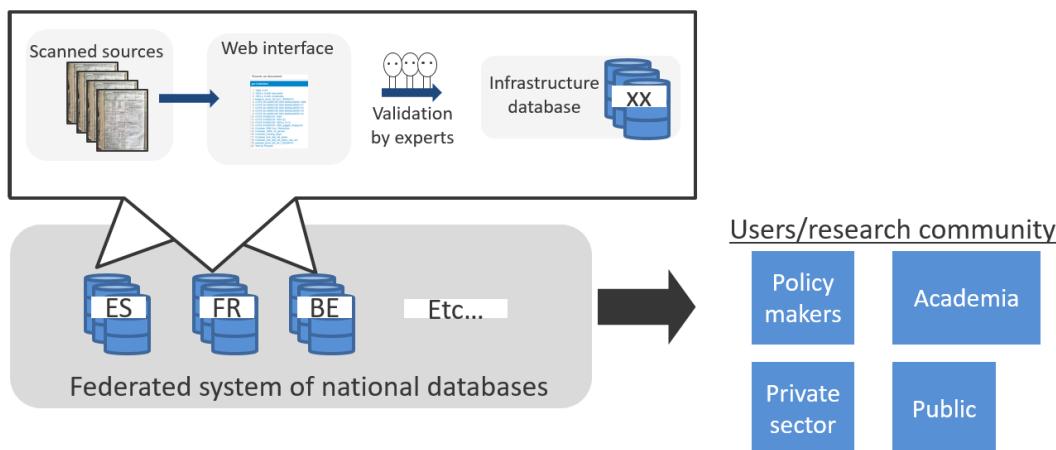


Figure 1: The data extraction, transformation, and integration process

Within this framework, additional work was carried out to design the architectural system and data model, as well as the integration of the extracted data with modern financial data. EURHISFIRM also designed its governance (incl. legal aspects), business model, and cultural heritage analyses for its long-term sustainability.

In this section, we describe this work achieved in the project (i.e. the results of the Work Packages [WPs]). Note that the below sections are general summaries of the work achieved; for further details, refer to the relevant documents of the corresponding WP.

III.1. Data extraction technology, transformation, and integration (WP6, WP7)

As illustrated in Figure 1, the data extraction process begins from the scanned historical paper sources. Through WP7's machine learning and artificial intelligence technology, the data is extracted into digital format and inserted into a web interface that shows the digital text and its paper source; this is validated by experts for cases where there is uncertainty or for training of the machine learning for new types of

data. These data are then inserted into the database for the associated country. Indeed, EURHISFIRM uses a federated system, i.e. each national database retains its specific characteristics and a fair amount of autonomy. Each database then makes up a part of the overall RI system, in which these national data are cross-matched and accessible to users under an open data-compatible platform (i.e. Wikibase). In other words, users will be able to find the information for a company ABC that exists across multiple countries through this single platform, which is made of independent national databases.

Artificial intelligence and machine learning were incorporated into the data extraction phase to ensure continuous addition of new data with reduced manual work. This would enable the RI to continue expanding its data sources from different countries and languages while maintaining data accuracy. For the design study, the data that were covered within the development scope were the following:

| Time period | Yearbooks | Official price lists |
|-------------|-------------------------------|---------------------------------------|
| Before WWI | Germany (Handbuch) 1913-14 | Belgium (Brussels, in French) 1912 |
| Interwar | Spain 1931 | Spain (Madrid) 1931 |
| Post WWII | France (Desfossés) 1962 | France (Paris) 1961-1962 |

Figure 2: Data used in the production of the design study within WP7

The samples were selected according to the available scan quality of the data, differences in language (we prioritized selecting from as many different languages where possible in order to test the system's capacity to handle data heterogeneity), and historical particularities (war periods were avoided due to abnormal data during that time). (NB: WP4 conducted an in-depth study on the types of sources available; see below section [Data inventory and meta data standards selection](#)).

Additionally, benchmarks and result measurements were incorporated in order to understand the performance and accuracy rates.

Further details on the work accomplished in WP7

WP7 (Data extraction and enrichment system) developed an intelligent and collaborative system for the extraction of structured information from images of historical documents related to companies' financial and economic activities. For the design study, we developed, on yearbooks and price lists, a generic system able to be adapted to a new kind of yearbook or price list. The data extraction system uses redundancies in a document collection (like daily price lists) to drastically reduce the need of manual validation by human experts, while generating high-quality data.

Information extraction in yearbooks

Text Blocks Selection by Document Structure Recognition

A- Graphical Components detector

We built a generic information extraction system for yearbooks. It is based on graphical components such as table separator recognition, with or without existing rulings, contextual segmentation of text lines, rubric header detectors, text alignments detectors... These detected elements are then used in a generic way by a grammatical description of the yearbook layout structure to make a structural analysis of the yearbook organization on text paragraphs (title, subtitle, rubrics) and complex table structures.

B- Structural analysis

To have an easier adaptation to new types of documents of the information extraction on yearbooks, we built it in a generic way: the core recognition of the structure is the same from one yearbook to another and we just have to specify the characteristics of each document before starting the generic analysis of the document structure. Examples of these characteristics are: the font size of the titles, the position of tables regarding the rubrics, the type of the tables (with/without rulings), the presence of a separator between issuers, etc.

C- Evaluation

To evaluate detection and classification of issuers, rubrics, table, balance sheets, administrators tables on yearbooks, we use the ZoneMap metric which is based on bounding box similarity. It provides a score between 0 (better) and arbitrary high values. **We tested a set of 294 images from the 3 yearbooks corpus, and we had a global ZoneMap score of 2.17. We also evaluated the quality of table content structure extraction with the Tree-Edit-Distance-based Similarity (TEDS). On a test set of 56 tables in the French Desfossés 1962 Yearbook we had a TEDS of 96.75% (higher is better) and on a test set of 100 tables in the Madrid 1931 Yearbook, we had a TEDS of 93.31%. We applied the data extraction system on the three yearbooks (Handbuch 1913-1914, Madrid 1931 and 1935, Desfossés 1962) for a total of 7,453 pages and it generated 6,009 issuers, 73,853 rubrics, 2,086 tables and 1,665 balance sheets.** More details are given in D7.4 and D7.2. Some examples of Yearbook structure recognition results are presented in Figure 3.

Top Left: Handbuch 1913 (Left)

Top Middle: Madrid 1931 (Middle)

Top Right: Desfossés 1961 (Right)

Bottom Left: French Desfossés 1961 (Left)

Bottom Right: Madrid 1931 (Right)

| NOMS ET PRÉNOMS | ADRESSES | NOMS DES SOCIETES ET POSTES DANS LE CONSEIL |
|-----------------|---|---|
| SENARD Georges. | 164, Rue de Courcelles, Paris (17 ^e). | Adm.: Banque Parisienne pour l'Industrie, pour les Assurances Industrielles, Paris-Ouest, Transways de Lille, Chemins de Fer Economiques du Nord, Société Parisienne d'Etudes et de Participations, Carbonisation et Charbons Actifs. |
| BENARD Jean. | 15, Rue Daubenton, Paris (5 ^e). | Adm.: Banque Fse du Commerce extérieur. |
| BINARD Lucien. | 18, Rue Spontini, Paris (16 ^e). | Adm.: Signaux et d'Enterprises Électriques Application Générale d'Électricité et de Mécanique S.A.G.E.M. |
| BENASSY Robert. | 148, Avenue d'Italie, Paris (13 ^e). | Adm.: Sfe G. |
| BENAZETH Louis. | 6, Boulevard de Tounis, Marseille (B.-du-Rh.) | P. D. G. : Madagascar Automobile, Morsillaise de Mouloussor, Commission, Consignation et Courtage de l'Océan Indien, Société Commerçante d'Afrique, Agro- et Industrielle de Madagascar, Péninsulaire de Navigation. |
| BENDALL Manley | * Le Roc *, Monsegur (Girande). | Adm.: * Savana * (Industrielle, Commerciale et Financière) |

| Capital, reservas y utilidades | | | | | |
|--------------------------------|--------------------|-------------------------|------------------------|----------------------------|-----------|
| EJERCICIOS | CAPITAL | FONDOS DE RESERVA (I) | Utilidades Liquidadas | | |
| | Nominal Pesetas | Desembolsado Pesetas | Estatutario Pesetas | De amortización Pesetas | Pesetas |
| 1924-25..... | 9.098.000 | 203.602 | | | 1.235.763 |
| 1925-26..... | 9.098.000 | 262.326 | | | 1.389.191 |
| 1926-27..... | 9.000.000 | 329.337 | | | 1.376.916 |
| 1927-28..... | 9.000.000 | 395.600 | | | 1.475.096 |
| 1928-29..... | 10.000.000 | 469.385 | | | 2.068.520 |

Figure 3: Yearbook structure extraction and table recognition : Top: Handbuch 1913 (left), Madrid 1931 (middle), Desfossés 1961 (right); Bottom: Administrator table recognition in French Desfossés 1961 (left), Table structure recognition in Madrid 1931 (right)

Information Extraction from Text Blocks

A- Information extraction and linking model

Financial yearbooks contain a lot of textual information organised in rubrics. Each rubric is reporting about specific financial information and requires a dedicated extraction module. Some rubrics are simply lists of items such as lists of persons for which simple rules encoded with regular expressions may be enough to get the information extracted. Some other rubrics are much more difficult to analyse as shown on Figure 4 below which gives three examples of the rubric "Capital" for the French, German and Spanish yearbooks. The three rubrics shown on this figure highlight with the appropriate color code the typical content of financial yearbooks whatever the stock exchange and the language used. However we can see that there are many different ways of reporting the information depending on the yearbook. For example, while both the French an Spanish yearbooks report the total capital amount evolution through the years, the German yearbook reports the amount of capital increase or decrease over the years. Thus the need to define the proper tagging conventions for each and every rubrics of interest, as depicted in Figure 4 below.

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

M. 3000000, eingeteilt in 3332 Nam.-Aktien à M. 300 und 1667 Nam.-Aktien à M. 1200. Die G.-V. v. 1.13. 1905 beschloss Erhöhung des A.-K. um M. 499800 (auf M. 3000 Q00) in Aktien i M. 1200. Die neuen Aktien, div.-ber. ab 1/1. 1912, wurden 1911 zu 114% be- geben. Gelegentl. dieser Erhöhung wurden 2 Aktien à M. 300 eingezogen u. eine weitere Aktie à M. 1200 ausgegeben. Die G.-V. v. 15./3. 1912 beschloss weitere Erhöhung um M. 1500000 (auf M. 4500000) in 1250 Aktien à M. 1200 zum Begebungskurs von mindestens 105%. Die Zeit der Em. bestimmt der Vorst. u. der A.-R.

Capital. 32. 000. 000 de pesos, enteramente desembolsados. Anteriormente fué de 20. 000. 000, elevado a 26 en 1903 y a 32 en 1905.

TAGS: amount ini-date chg-amount chg-date currency last-amount Cap-incr Cap-dec

Figure 4: three examples of entity extraction in the Capital rubrics for the three yearbooks considered

We adopted a statistical model for Named Entity Recognition (NER) which was presented in D.7.2 and published in the International Conference on Document Analysis Systems, 2020. This system relies on pre-trained word descriptors known as word embeddings that feed a hybrid architecture made of a recurrent neural network and Conditional random fields, following the architecture proposed by Akbik *et. al.* in 2018. This architecture is trained specifically on the EurHisFirm yearbook dataset made of some sample pages that were annotated with our tagging conventions in close collaboration with the historian colleagues for our purpose. Tagging was made possible through the use of a web-based collaborative annotation tool. This tool played a key role for the annotators to visualize the original scanned pages, the transcription made by the OCR, the tagging of this text which the user can modify, and finally the result of the extraction process visualized in a table that highlights the relation between the tags. It is to be noticed that the tagging process operates on the textual transcriptions of document images. For these experimentations the transcriptions have been generated using a professional OCR without any additional processing on our side. This means that the transcriptions contain some errors but they did not impact the extraction process itself, but of course the information extracted are subject to character errors. We did not investigate training a specific OCR for each of these corpora, we estimated this would have required much more time and efforts for a limited impact of this design phase of the project. Of course this can be an option for future works, considering that the OCR technology that was developed for the price lists could apply here as well.

B- Evaluation

More detailed performances are presented in D.7.4. The performance may vary depending on the regularity of the phrasing of the rubric considered, and depending on the amount of manual annotations that were produced to train the system. It is to be noticed that overall ***we analyzed a total of 19 types of***

rubrics in three different languages and distinguished 126 different types of Named Entities. The system was able to tag and extract automatically more than 40 000 named entities exploiting 4844 hand labelled data with a precision higher than 90%.

Information extraction in price lists

The information extraction in price lists system consists of two main tasks: the document structure recognition using a cross validation module with transversal analysis, and the definition of a general-purpose text recognizer (OCR). As the information extraction system in yearbooks this system is built in a generic way to be able to be applied on different kinds of price lists.

Structure Recognition and Transversal Analysis

A- Pricelist Structure Recognition

The price lists structure recognition system has been designed to first extract the global meta table structure where each cell is itself a price list table. The system can detect the reading order between each cell of the meta table structure. It is then able to recover, for example, two price lists which are side by side. The system uses the consistency in price list table headers to validate the reading order (see Figure 5 - top left). As this meta table structure recognition is defined in a generic way, it is the same for all different kinds of price lists.

Then each price lists table is described in a similar generic way to the yearbook data extraction system: the core recognition of the price lists structure is the same from one price list to another and we just have to specify the characteristics of each document before starting the generic analysis of the price lists structure. Examples of these characteristics are: the section title structure, securities on unique or multiple or lines, etc. We can also specify how the data should be structured to produce at the end of the process, an XML file with all the extracted information. An example on the Paris 1961 official list is presented in Figure 5: global meta table recognition with the reading order (top left); column and header recognition (top right); section recognition (top right) with a section starting in a cell and ending in the next cell (in green); and stock recognition (bottom right) on multiple lines.



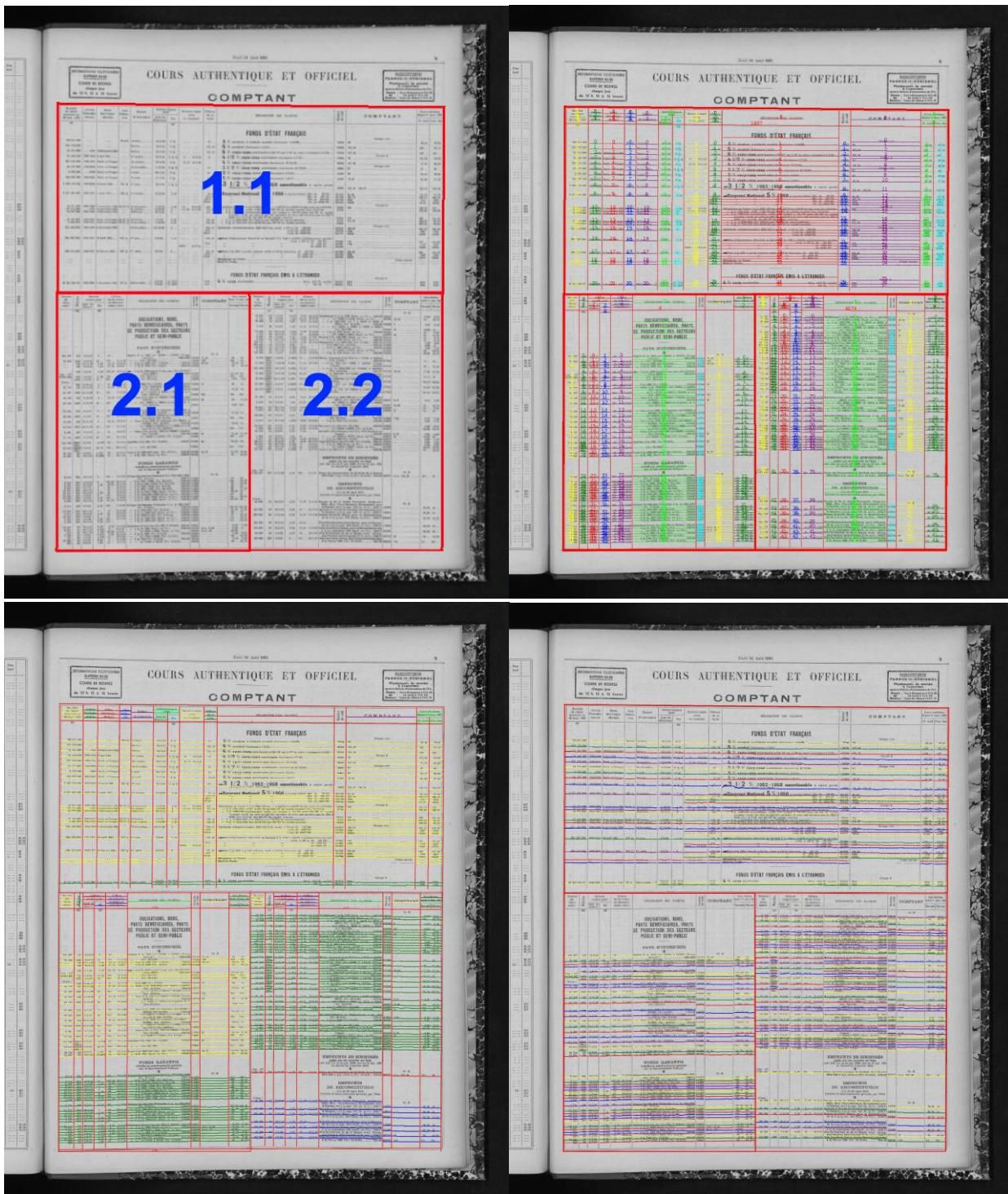


Figure 5: Paris 1961 Price Lists: Top: global meta table recognition with reading order recognition (left); Price list column and header recognition in each cell (right). Bottom: Price list section recognition with reading order (left), each color represents one section (the green section starts in cell 2.1 and ends in cell 2.2); stock lines recognition (right) with stocks on multiple lines, each color represents one stock)

B- Strategy and transversal analysis

A global strategy has been designed to take advantage of the sequential nature of the collection to automatically correct some detection errors of the elements that are known to appear in every publication. Our global strategy is based on an iterative process which allows a cross validation of various information in the document collection through a transversal analysis of documents. The aim of each iteration is to recognize and validate a structural or textual element of the documents: columns, sections, stock names (table entry), and other fields, by using redundancies found in the sequence of daily quotations. At each iteration, the system can produce questions if needed, which are asked to expert users in an asynchronous way. It is mainly for information which is not present in the document, like the unique ID which should be associated with a new security, or some ambiguities on currency abbreviation, for example. This global strategy with cross validation, allows to improve the quality of the data extraction, while reducing drastically the number of expert user interactions.

C- Evaluation

Complete evaluation is available in D7.4. Meta table structure recognition has been evaluated with a **ZoneMap score of 0.76 on a test set of 243 tables from the three price lists (Brussels 1912, Madrid 1931, Paris 1961-1962)**. The transversal analysis applied on column header recognition, on a **test set of 4,055 pages from La Coulisse between 1899 and 1915 showed a reduction of errors from 320 to 18 errors (with global strategy on the document collection)**. An evaluation of the whole data extraction with stock identification on 1696 stock lines from La Coulisse 1899 has been done: the F-measure is improved from **0.914 without collection context to 0.988 with the collection context and expert user interaction**. This quality of data extraction is done while, **on 536 pages and 54,603 stock lines from 6 months of La Coulisse 1899, the number of questions to expert users is reduced from 4,061 to 309 with the collection context**. All the data extracted is then produced in XML (see Figure 6). Results of the process of data extraction on Brussels 1912 (1946 pages), Madrid 1931 (2570 pages) and Paris 1961-1962 (469 pages) Price Lists are presented in D7.4.

General purpose text recognizer

A- System overview

The Optical Character Recognition system consists of a deep neural network made of convolutional layers (CNN) followed by bidirectional recurrent layers (BLSTM), whose main purpose is to predict the characters in an image. The output predictions are then analysed by a Language Model which decodes the predictions to find the best sequence of characters allowed by the language model. The language model into a desired format or item in a list that has been defined. *The neural network* is optimized during a training phase using hand-labelled images so as to adjust its parameters and achieve the best recognition performance. In the end, the trained neural network achieves the image to text transcription. More precisely it provides a lattice of character hypotheses associated with their probability. *The Language Model* consists in the combination of lexical and grammar rules that helps the OCR format the output into a structured text containing the desired elements. For instance, the StockNames column's objective is to find, with

precision, the list of stock names belonging to a list of names. This ensures that the Language Module replaces the variations in the output into the correct entry thus keeping a continuation through the pages.

B- Evaluation

Complete evaluation results are reported in D7.4. As for a general evaluation, **we were able to learn and analyse 4 different price lists, including 2 languages (French, Spanish). A total of 104,000 images were hand labelled and used for training and testing purposes.** Throughout the training process of the neural network, **we applied a data augmentation technique which reduces the Character Error Rate (CER) by 23%. Overall, the system has achieved an average raw CER of 1.8%.** Moreover, a Language Model was designed specifically for each information considered (column of the price lists) to decode the lattice of character hypotheses and encode the final recognition result into the specified formatting rules of the information. **In some cases the language model also allows for some corrections in the recognition results thus reducing the final CER from few % to more than 50% depending on the column considered.**

FONDS D'ÉTATS, DE DÉPARTEMENTS, DE VILLES

| NOMBRE de titres | VALEUR nominale | DÉSIGNATION DES VALEURS | COURS COTÉS plus bas, plus haut, dernier | COURS précédents | COUPONS DATE du dernier payé | Nos | Montant BRUT | Dernier REVENU | OBSERVATIONS |
|--|----------------------|--|---|---|--|--------------------|---|----------------|--------------|
| 6.000.000 l.st. 34.000.000 lir. 70.966 130.000 4.492.875 lir. 4.500.000 dol | 60 500 500 | FOND D'ÉTATS, DE DÉPARTEMENTS, DE VILLES BRÉSIL 5 o/o (1895) FLORENCE 3% HAITI 5/o Bons de Coupons MINAS GERAES 5 o/o NAPLES 5% SAINT-DOMINGUE Réclamation franco-américaine 4 o/o SAINT-LOUIS 6 o/o oblig. | 40 .. 41 .. 365 .. 366 230 .. 233 | 68 1/2, 57 .. 40 .. 368 .. 84 1/4, 110 .. 235 .. 33 3/4, | 4 février 99 Octobre 98 Janvier 99 15 Janvier 99 Janvier 99 Septembre 98 Novembre 90 6 Mai 98 | 38 3 10 4 | 2 fr. 17 1 Fr. 50 12 Fr. 50 2 Fr. 50 15 Fr. 10 Fr. 0,50 o/o | | |

<<< Previous day 18990227 Next day >>>

FONDS D'ÉTATS, DE DÉPARTEMENTS, DE VILLES [678]

| NOMBRE de titres | VALEUR nominale | DÉSIGNATION DES VALEURS | COURS COTÉS plus bas | plus haut | dernier | Derniers COURS | COUPONS Date du dernier payé | Nº | Montant Brut | Dernier REVENU | Observations |
|--|-----------------|--|-------------------------|-----------|---------|----------------|---------------------------------|----|----------------|----------------|--------------|
| 6 000 000 Livre Sterling (Pound) Titres admis / Total Amount (of capital) | | BRÉSIL 5 % (1895) [106559 - 207602] | | | | 68,5 | 1899-02-04 (DAY) | | 2,17 undefined | | |
| 54 000 000 Lira Titres admis / Total Amount (of capital) | | FLORENCE 3 % [104857 - 207681] | | | | 57 | 1898-10-01 (MONTH) | | | | |
| 70 966 Titres admis / Number of Securities | 60 undefined | HAITI 5 % Bons de Coupons [104916 - 207682] | 40 | 41 | | 40 | 1899-01-01 (MONTH) | 38 | 1,5 Franc | | |
| 130 000 Titres admis / Number of Securities | 500 undefined | MINAS GERAES 5 % [4551 - 207683] | 365 | 366 | | 368 | 1899-01-15 (DAY) | 3 | 12,5 Franc | | |
| 4 492 875 Lira Titres admis / Total Amount (of capital) | | NAPLES 5 % [104858 - 207684] | | | | 84,25 | 1899-01-01 (MONTH) | | 2,5 Franc | | |
| 4 500 000 Dollar US Titres admis / Total Amount (of capital) | | SAINT-DOMINGUE. Réclamation franco-américaine 4% [105466 - 207685] | | | | 110 | 1898-09-01 (MONTH) | 10 | 10 Franc | | |
| 25 000 Titres admis / Number of Securities | 500 undefined | SAINT-LOUIS 6 %. oblig. [106560 - 207603] | 250 | 253 | | 255 | 1890-11-01 (MONTH) | 4 | 15 Franc | | |

Figure 6: Example of the WP7 Price lists data extraction system: 27 February 1899 - La Coulisse, Data extracted in XML (bottom) after the process of the page (top) in the context of a collection of 6 months of quotations. Each extracted data is linked to its location in the image (see in red "SAINT-DOMINGUE")

Further details on the work accomplished in WP6

WP6 (Data connecting and matching) developed and tested technologies to make databases interoperable by developing and defining the conceptual framework and technologies to match existing datasets and databases within the EURHISFIRM RI, as well as with those outside of the EURHISFIRM RI (e.g. web-based information; modern financial data). This work consisted of testing technologies to match national and cross-countries' data.

Deliverable 6.1 formally describes the process of data matching, and tackles the issues that arise from attempting to match data from various sources within the EURHISFIRM project. An overview of existing data matching methodologies is provided, and the extent to which they are, either directly or after suitable adaptations, applicable within our framework is discussed. Furthermore, pre-processing procedures, that, due to the heterogeneity of the data, may need to be applied before the data matching itself can take place, are identified and explained. Such procedures include data formatting, data harmonization and data synchronization.

In this deliverable, a variety of data matching techniques specifically designed for particular data matching problems is discussed. The process starts with schema matching, where, given two (or more) databases, the goal is to identify which tables in different databases may contain the same (or similar) data; and, in a second step, which attributes (or columns) in these tables should be matched to each other (and crucially, how). Once the schemas have been matched, the next step is record matching, where the goal is to identify which records (or rows) in one table can be matched to other records in another table (in the other database).

While the main goal of data matching is to correctly identify the data that should be matched, in the presence of large quantities of data, efficiency is another factor that has to be taken into account. Therefore, the usage of automated techniques whenever possible has been investigated, particularly with an eye on the trade-off between efficiency and accuracy. In some cases, human intervention is necessary to verify (or reject) the possible matches discovered by data matching algorithms. This perspective also leads to the rationale for creating a collaborative environment where human efforts and automatic techniques can register matches between entities in separate databases to improve the outputs of the tools that find those matches.

Deliverable 6.2 presents a similar analysis of issues arising when trying to connect existing external data sources to individual databases belonging to the EURHISFIRM partners and, ultimately, to the central EURHISFIRM database, containing either the integrated data itself or links to other repositories where data is stored. A wide variety of external data sources was studied, with particular attention to the possibility of data coming in different formats, ranging from structured databases, Excel files or text files all the way to publicly accessible web pages. The challenges of processing the data from such a variety of sources were identified, and state-of-the-art techniques for overcoming these challenges were presented.

For this task, the data matching methods introduced in Deliverable 6.1 are central. However, due to the additional complexity of the task, further, more robust, methods have also been developed. Finally, the possible options for storing and maintaining the linking data, once the connections between the datasets have been established, are discussed and evaluated.

Finally, WP6 also resulted in several case studies, putting into practice the data matching and connecting techniques in both relational databases and a wikibase collaborative environment, as reported in Milestones 6.1 and 6.2.

III.2. Meta data, the common data model within a federated system, and technical architecture (WP4, WP5, WP9)

Data inventory and meta data standards selection

WP4 (Data and sources inventory and documentation) established a common standard of documentation for EURHISFIRM's data and sources and inventoried the existing long-term datasets and sources on European companies according to this common documentation standard. It used this inventory to analyze the data semantics and quality.

Overall, WP4 had to tackle the methodological challenge of ensuring standardised approaches whilst allowing for idiosyncrasies of the diverse data types from various countries across time. The work of WP4 therefore consisted of five interrelated tasks. Tasks 4.1 and 4.5 were of a more technical nature and involved the selection of a metadata standard and software for data and sources documentation (i.e. the description of the provenance, characteristics, structure, and contents of datasets and document sources). A preliminary choice of standards and software was made during task 4.1. After a comparison of several metadata standards for the social sciences, it was decided that the Data Documentation Initiative (DDI) Lifecycle standard and the Colectica software were best tailored for EURHISFIRM's needs. These preliminary choices were confirmed by task 4.5, at least as far as the back-office process of harmonisation of data across datasets is concerned. For the front-office process of data documentation of individual datasets, the DDI Codebook standard and Dataverse software are recommended. The final decision on the data documentation standard and software were informed by the experience of tasks 4.2, 4.3 and 4.4 which were of a more heuristic and hermeneutic nature and involved the identification, categorization and contextualization of existing datasets and sources (task 4.2), the analysis of their semantics (task 4.3) and the production of 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 therefore served as a test-case for the suitability of the selected metadata standard.

In order to render the work manageable, the range of potential sources for task 4.2 was limited to printed serial sources of financial, governance and geographical information on publicly traded companies from 1815. The inventory of the data and sources nevertheless contained more than 250 sources from all countries in the consortium (Belgium, Germany, France, the Netherlands, Poland, Spain, and the United Kingdom). An in-depth analysis of existing company-level data and historical serial sources was 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 c) geographical data (e.g. location of headquarters, subsidiaries, and production units). This inventory delivers in-depth knowledge on the type, quality, accessibility, and other key characteristics of yearbooks, stock exchange price lists, and other primary and secondary sources. This inventory is complemented by a report presenting an in-depth analysis of the semantics of the types of data which are commonly found in printed serial sources and datasets with governance, financial and geographical information on publicly traded companies (Deliverable D4.3). Formal documentation of data and sources according to the DDI Lifecycle standard was produced for a selection of sources, consisting of the official price list of the

principal stock exchange and the most important yearbook in each consortium country and the existing long-term databases (SCOBI, D-FIH, EUROFIDAI and London Share Price Database) and datasets collected by consortium members (QUB, SAFE and U3CM). In conjunction with the semantical analysis, this data and sources documentation uncovers the relationship between the terminology used as section or column headings in sources and datasets covering different time-periods and countries on the one hand and their historical denotation (i.e. their meaning in a certain time and place) on the other hand. WP4 also tackled the methodological challenge of ensuring standardised approaches whilst allowing for idiosyncrasies of the diverse data types from various countries across time.²

The Common Data Model within a federated system

Due to the heterogeneity of national historical data sources and the heterogeneity of respective "contents", a Common Data Model (CDM) is necessary in order to - in a federated system - enable a sound integration of newly added data with the existing data of the EURHISFIRM Research Infrastructure (RI).³ ⁴ An important goal is to provide a Common Data Access Service (CDAS) for different classes of users that - without being overly restricted in expressing her/his requests - bridges the substantial semantic difference between queries being expressed in a user-friendly language and the heterogeneously available original data which has in important parts to be pre-harmonised. Towards that end, (aside of singular exemptions) original data - primary data (object identification data), secondary data (complementary object data), and respective metadata - cannot easily (in short time spans, with small investments) "in one step" be ameliorated into a CDM-compliant version. Rather, we define a staging concept - a sequence of amelioration steps of data (some components may be partially automated, others still need to be developed) - that, when executed on original data, - looking bottom-up - increasingly fulfills the CDM-compliance. In fact, the original scans of historical sources and respective data extractions must be retained to enable the user the option to evaluate the results of the various melioration steps.

To that end, **WP5** (Common Data Model) provides the core of an extensible standard CDM to access, navigate, select, and retrieve data from a variety of sources and datasets. Of specific interest is the definition of common semantics and quality standards evaluated in WP4 as a basis for the CDM. Most importantly, a common "biunique"⁵ identification regime ("primary data") of fundamental objects - here: in the realm of finance and firms - has to be adopted by each data source and each relevant data set. A conceptually easy solution⁶ is to amend all relevant original firm data sets by the EURHISFIRM Legal Entity

² D1.8

³ D5.1

⁴ D5.4

⁵ This "biunique" identification regime ensures that each single firm - in Europe - obtains only one single identifier and simultaneously ensures that each identifier is assigned to only max. one firm - and this is ensured also in a federated system with distributed data centers. We name this feature further-on "unique". Actually, in our times of world-wide interconnectedness a world-wide uniqueness of the Legal Entity Identifier (LEI) promises superiority over an identifier that guarantees the uniqueness only when looking at Europe.

⁶ Of course, the implementation of such a concept is expensive and time-consuming - but it is somehow "accountable" because the number of national incorporated companies is limited. But this concept is easy. If we never start with such first steps to enable such a sound basis for a harmonisation process, we will not see sustained

Identifier (ELEI)⁷ which we (for historical Legal Entities) derived from the ISO⁸ standardised LEI⁹. Additional emphasis lays on the elaboration of harmonised metadata (e.g. with respect to accounting regimes or naming conventions) and on ensuring that the metadata standards in the melioration process sequence are consistent with ESFRI Landmark CESSDA ERIC standards.

The formation of the **WGIS** (Working Group on Identification and Standardisation), an inter-WP group consisting of both the technical members of the project as well as - with respect to connected strategic decisions - the WP members (WPs 4, 5, 6, 7 and 9) broadly tied the work of WP5 into the project and acted as a 2+ years self-trial with respect to both the data-structural as well as the procedural parts of the CDM. In fact, we propose a likewise federation-oriented course of action for further development steps. We applied a state-of-the-art software and standards development methodology (Enterprise Architecture) to - in a revolving sequence of information, design, reconciliation, and specification - reach agreements on commonly accepted standards - ideally, when possible, using available industry standards. In the end, the WGIS accepted a substantial set of standards - in this final report we concentrate on sketching the two main design decisions:

- a) Content-wise, the core of the extensible CDM is comprised of three fundamental objects and their relationships (denoted by double-headed arrows, see Figure 7):
 - Legal Entities (which may be generalised to “organisations”) - using the ELEI as the unique identifier
 - Financial Instruments - likewise using the EURHISFIRM Financial Instruments Identifier (EFII)
 - and Markets (for instance stock exchange markets with potentially multiple trade currencies for a given stock as attributes) - likewise using a market identifier

One can think of viewing Figure 7 in three vertical areas. The central vertical area in Figure 7 (“Markets”) contains the object “Currencies”, for example. The left and right vertical areas - Legal Entity (LEs) and Financial Instrument (FIs) - are complemented by firms’ financial statements data and time series objects respectively - both also being related to their respective normalised¹⁰ objects. Below the Legal Entity and Financial Instrument objects, you see “artefact” objects which is a modern means of -

results. And our design of the ELEI also allows for an incremental implementation in a federated (distributed) data environment.

⁷ We provide a quick first step installation guide for amending existing files of historical firm data by the ELEI in the appendix VII.1. to this report. This also acts as an example of how to use this technology for the EURHISFIRM Financial Instruments Identifier (EFII) and other common identifiers (e.g. “sources”).

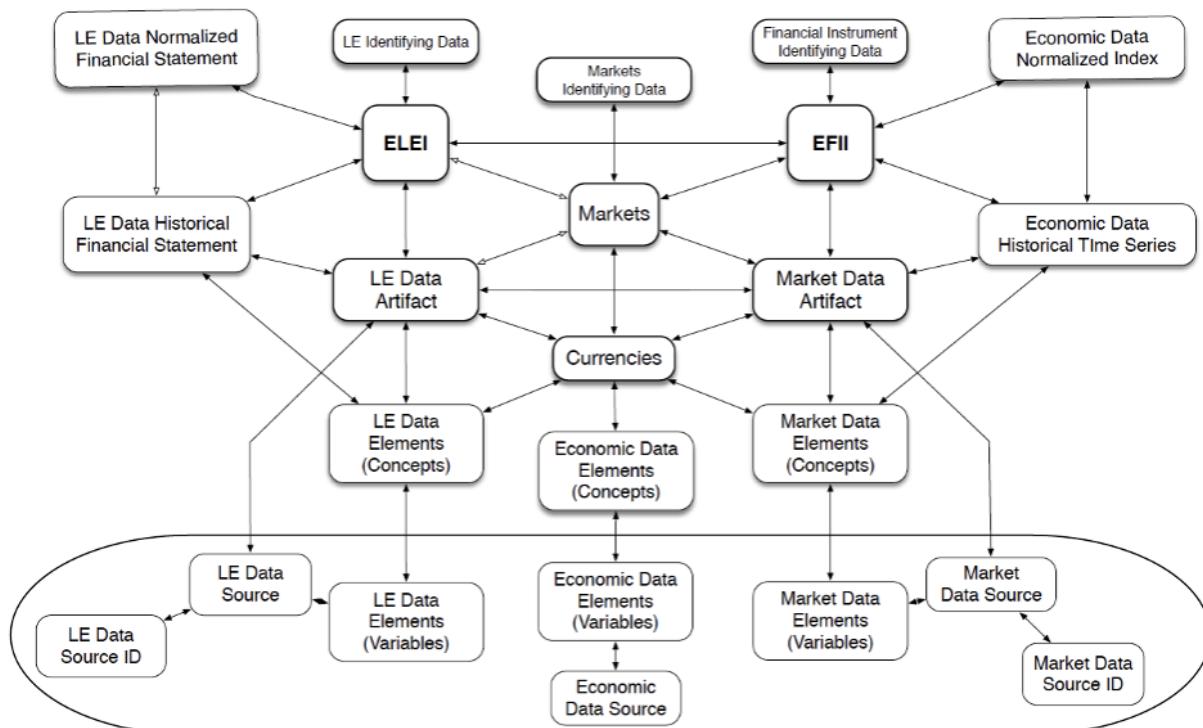
⁸ International Organisation for Standardisation

⁹ The CEO of the Global Legal Entity Identifier (GLEIF) has agreed to support the EURHISFIRM ELEI, so that, for instance, EURHISFIRM has a standardised access to contemporary firm data (in the finance realm) and can also use GLEIF’s sophisticated deduplication routines.

¹⁰ Some types of transformation of original historical data are typically performed that produce datasets with normalised financial data numerical values (e.g. monetary amounts, market prices). These transformations, for instance with respect to the time frame, are performed to enable a consistent interpretation and analysis.

aside of known data structures - integrating flexible data structures in data containers¹¹. This concept adds substantially to the extensibility of the data structure without the need to alter the up-to-now given fundamental objects. All three vertical areas are complemented by “concepts” data structures¹² that are further down out-spelled in “variables” objects¹³. The bottom part of Figure 7 - depicted in the oval - integrates the data of data sources.

Secondary data can be incrementally added to each object without alteration necessities of the existing data structures. Moreover, complementary objects can be added in later development phases of EURHISFIRM (without alteration necessities of existing objects), for instance (natural) “persons” (e.g. with respect to their role in Boards) and “survey” (if one wants to record for instance historical survey data on markets).



| <--> | Relationship |
|-----------|---|
| ELEI (LE) | EURHISFIRM Legal Entity Identifier (Legal Entity) |
| EFII | EURHISFIRM Financial Instrument Identifier |

¹¹ Think for example of a set of JSON documents comprising key-value pairs of extracted raw data that belong to a newly “found” historical document set being labelled by a timestamp. In D5.4, we have out-spelled the specification.

¹² We could also name these data structures “concept names”. If one gets raw data from different nationalities - for instance “immobile assets” - , we need to harmonize these by means of an overarching European concept name.

¹³ Here, the actual value is assigned (to the, for instance: national, variable).

Figure 7: Central Entities (D5.4)

- b) In light of the principle of preserving historical data in its originality (which has to be stepwise ameliorated to - in the end - be fully compliant with the CDM standard) we developed an - also extensible - sequence of melioration processes which also soundly works in a federated system environment. Figure 8 shows an exemplary four steps sequence of melioration from original ("raw") data - depicted on the bottom layer¹⁴ which are collected by "Data Collection Units (DCUs)" up to the fully-CDM-compliant "Common Data Access Service (CDAS)":
- Data artifact variables tagging and formatting
 - Data artifact concept mapping and local identification
 - Consolidating artifacts / data merging across federatively distributed DSUs
 - Identification, matching, linking and reconciliation across Network Integration Centers (NICs)

With respect to the federated system design, some activities are described in WP9. The organisational units - depicted in Figure 8 on the left hand side - are (from top to bottom):

- a European EURHISFIRM discussion and decision body governing the CDAS and underlying standards
- a Network Integration Center (NIC) can be a national data consolidator
- a Data Submission Unit (DSU) can be an able research institute (say, in Wroclaw)
- Data Collection Unit (DCU) can be a specialised data extraction unit for historical documents (say, in Rennes/Rouen)

¹⁴ For example, WP7 has experimented with "company yearbooks" which have first to be scanned and then further data extraction and interpretation processes take place, and "price lists" which may already come into EURHISFIRM as digital data ("other data sets"). Moreover, respective data come from "existing databases" or might be entered "manually".

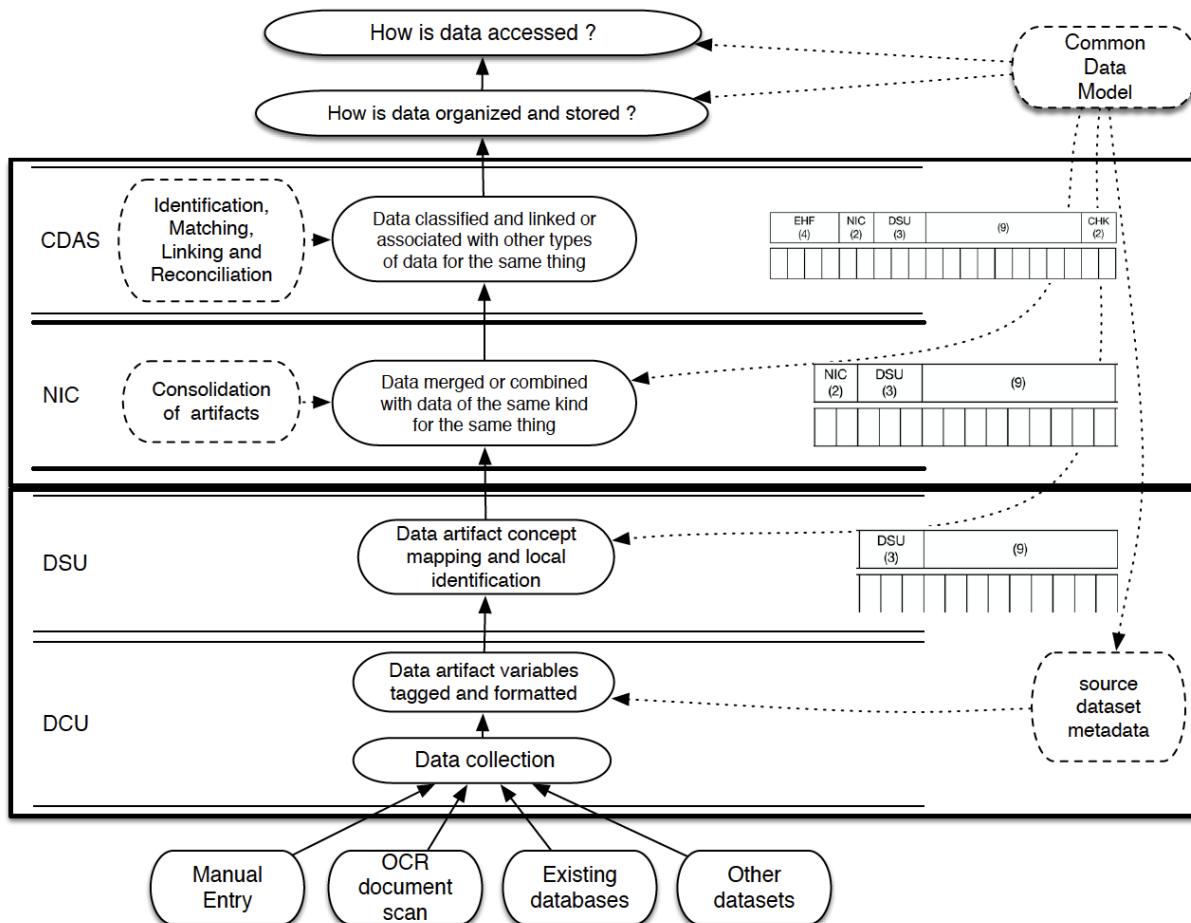


Figure 8: Data Staging

With respect to the unique identification of firms in a federated system environment, in this distributed amelioration hierarchy of firm data from bottom to top, the DSUs have the task to assign the 9-digit identification part of the firm, complemented by the 3digit identification of the respective DCU. This compound assignment is announced to the NIC and the CDAS levels in the sense: The time wise first assignment in the federation will stay the permanent assignment - other assignment candidates which come later into the system have to be consolidated to the first occurrence (WP 9 uses “stamps” to denote this sequence). Likewise, NICs qualify their (network¹⁵) identification consolidation with the assignment of their 2-digit NIC identifier. And, finally, on the CDAS level, the 4-digit European Historical Firm (EHF) identifier¹⁶ qualifies its (European) identification, matching, linking, and reconciliation results. The EHF number is assigned from the Global Legal Identifier Foundation (GLEIF) as an additional specific prefix in their “network” of world-wide Local Operating Units¹⁷ identification denoting “European historical firms”.

¹⁵ For instance: national

¹⁶ which is granted and honored by GLEIF

¹⁷ The label Local Operating Units (LOUs) which indicates for example that there are four German LOUs of GLEIF working in parallel was used in the very early times of GLEIF, but has been dropped meanwhile, because the locality of a unit isn't the only fundamental differentiator. This also allows GLEIF to provide an “European historical firm” operator - EURHISFIRM - an own (unalterable) identification.

One sees also in Figure 8 in the top-right corner, that the accumulated identification compound is amended by a 2-digit security code - the (on the CDAS level) overall identification compound is then fully compliant with the ISO GLEIF standard¹⁸.

The coordination between the different process stages is essential to provide high-quality, consistent data to end-users. The identification, harmonisation, and consolidation of data are crucial to achieving the RI's objectives.

This process sequence of producing CDM-compliant data does not exclude users from examining intermediate, staged data. Each transformation step is transparent, and, similarly to source data and CDM data, the staged data are FAIR. The harmonisation process between DCU and DSU is consistent with the Data Documentation Initiative (DDI) metadata definitions of variables for local data and conceptual variables for harmonised data. Local language labels for financial terms complement common terms for the same data elements during the transformation steps, but links to the original labels are preserved.

Beyond these two main standardisation decisions taken with respect to the core CDM structure and the amelioration staging concept (and the respective CDM extensibilities in a federated system environment) there are more harmonisation necessities of heterogeneous historical source data - part of these debates and decisions have already been performed - other parts are still to be worked-out. Here two examples:

- a) Pre-digitised source data come in different database methodology frameworks (for instance relational vs. graph-oriented databases). In WGIS we have already agreed on an overarching semantic equivalence model among machine-readable data persistence implementations¹⁹ that allows for instance a transformation of data between different logical database structures (e. g. relational to graph-based triplestore data structures).
- b) The semantic harmonisation of diverse attributes calls for an ontology to, for instance, harmonise national idiosyncrasies (e.g. because of different national legal regimes).

It is clear that all harmonisation efforts that are requested by users of EURHISFIRM data but *not* performed by such a RI system prior to a request have to be performed by each end-user individually - and moreover individual solutions most likely are not systematically ameliorated and stored for easy use of follow-up fellows - they thus often must start harmonisation efforts from scratch again.

Moreover, it is clear that the concepts of the core data structure and the amelioration staging may also be applicable to general historical data in the Humanities and thus can pay into the claim to spark the Big Data revolution in the digital Humanities.

¹⁸ The world-wide GLEIF database of contemporary firms and the EURHISFIRM database of European historical firms are not integrated, but loosely coupled. This allows EURHISFIRM for instance, to use the highly sophisticated deduplication procedures of GLEIF. The interconnection between historical and contemporary firm data is possible by adding the GLEIF identifier to the EURHISFIRM data structure of historical firms.

¹⁹ D5.4

Technical architecture

WP9 (Infrastructure policy and architecture) designed the architecture and the operation of the RI, with regard to usage, access control, security, support and maintenance. Users' preferences on data and services design guided the platform architecture and operating. Accordingly, the security system, the maintenance/administration and the desk management of the platform were designed and estimated, with a focus on flexibility in order to accommodate possible unknown requests in the future.

In its reports and milestones, WP9 has put attention on three aspects of the architecture of the EURHISFIRM RI:

1. What policies are available and applicable to the RI, with the end goal in mind. This has resulted in indirect functional specifications covering software quality, security, access control (D9.1);
2. Business and application architecture elements, covering functional requirements with the end goal in mind (D9.4);
3. Technical architecture elements, taking the current national implementations as a starting point and suggesting how these may move forward (D9.3).

Results from other work packages (WP3, WP5, WP7, WP8, WP10) have been used to guide the WP9 documents.

The architecture design allows for scalability and loosely coupled components interacting in designed ways. This is helped by implementing parts of the functionality at different autonomous locations and using standard interfaces to exchange data and metadata. This implies software development at the national existing systems, if they want to be part of the federated solution. D9.1 proposes to adopt development standards in line with what CESSDA, EOSC, SSHOC and FAIR envision.

EURHISFIRM is designed to have access control through user roles, allowing flexibility on the one hand and compliance to regulations on the other hand. The idea is to implement a system that is resilient to future decisions. D9.1 addresses the current ideas on access roles and usage auditing.

D9.4 and M9.2 foresee the need for at least one syndicate competence centre (NIC) within a federation of autonomous implementations, where the NIC offers common or generic functionality and communicates with the autonomous locations via agreed interfacing. The way in which each existing national system may comply hasn't been decided yet. Neither do we know for sure what contributions EU-wide RIs may bring to implement parts of the EURHISFIRM functionality (specifically Dariah and the SSHOC SSH Open Marketplace in the near future). This makes assessing the exact future technical infrastructure and costs involved at each element challenging for D9.3. As this is highly ICT related as well, any written estimate will be outdated within a calendar year. For this reason the WP9 reports chose to be more functionally oriented in design and to advise further development to be based on prioritization by interacting members of the EURHISFIRM community. This should be aligned with the business model proposed by WP10, as described [in the Business model/sustainability section](#). Another factor is costs involved, since some of the prioritized backlog items defined in the EURHISFIRM community may have to be developed and maintained locally, centrally or licensed from available RI's (essentially, a make or buy decision for each backlog item).

In an implementation project of the common NIC functionality, compliance to privacy regulations and security considerations will be essential. D9.1 describes what to consider.

The WP9 documents propose to stimulate the EURHISFIRM community to communicate on progress and requests as well as contribute towards EU level data availability in projects. The characteristics of such a communication platform have been described in D9.2. This will benefit both inspiration as well as support within the EURHISFIRM community, in line with the requirements stated for the SSHOC SSH Open Marketplace.

Contentwise, by having data producers state a “completeness” level of data in their projects, EURHISFIRM helps stakeholders to see where contributions are needed the most and to what extent data is usable for research already. For this some sort of model of data stages is needed, as DDI, D9.1 paragraph 3.1 and the developed common data model in WP5 describe. Purpose of such a model is to express to what extent historical financial data has been sourced and processed into comparable EU level data. The stages imply a process of getting from source to target, without requiring that all steps are taken at all times. The stages could be used by the NIC to make transparent to stakeholders what the community of data producers is up to in getting towards the end goal of the RI. A visible prioritized backlog for data production may stimulate others to join.

III.3. Future users, community building, exploitation/dissemination (WP1, WP2, WP8)

Research on future users

WP8 (Interaction with users) completed research on EURHISFIRMs future users to provide specific recommendations for the optimal design of the data and services that the RI should provide. Recommendations were developed by gathering and analyzing the preferences of potential end-users and key stakeholders in both academia and industry. This information was gathered via an online questionnaire, which was carried out between October 2018 and January 2019 (D8.2), and also through semi-structured interviews with potential users, which were conducted between June and August 2019 (D.8.3). All steps adhered to the GDPR guidelines. Following on from these stages, and the completion of the Synthesis Report (D8.4), WP8 concluded the following specific recommendations for the design of the EURHISFIRM infrastructure:

a) Content

- Priority should be given to data relating to the twentieth century, as this was the most popular time period among our respondents.
- Regarding forms of company data, priority should be given to ordinary equity market data, which was by far the most popular with respondents.
- The EURHISFIRM platform should also aim to provide access to accounting data. Specifically, total assets, total debt, revenues, and profits.
- Data related to government and corporate bonds should be less of a priority.
- In terms of frequency of prices, daily and monthly data are recommended, where possible.

- As regards geography, the United Kingdom was the most popular country, followed by Germany and France.

b) Usability

- The EURHISFIRM platform should enable users to download the data in bulk, in MS Excel format, and with minimum restrictions on downloads per day. The EURHISFIRM platform should use Wharton Research Data Services as an example of best practice in this area.
- For reassurance as to the accuracy of data, users should be able to ‘click through’ to a scan of the original document and EURHISFIRM should provide a full citation of any source material.
- EURHISFIRM should provide an explanation of the methodology and rationale for any interpretation or manipulation of data carried out by EURHISFIRM researchers.
- For less popular data, EURHISFIRM should provide simple, non-tabulated PDF scans of the original source document, where possible.
- It is important that a feedback channel is available for users. An email address would be sufficient.²⁰

Outreach and community building

Within community building efforts, **WP2** (Dissemination and communication) was responsible for external communications (incl. website, social media), project identity/branding, outreach, and public event organizations (e.g. general assembly). Due to the COVID-19 global pandemic, the previously planned outreach events (e.g. an informational panel in Brussels, research dissemination and discussion forums, etc.) were not able to take place.

The strategic goal of WP2 was active communication and dissemination of knowledge about the project among the widest possible audience.

The first steps focused on creating the identity of the entire project as a basis for further identification during all promotional activities.

The identity should meet the criteria of professionalism and user-friendliness at the same time. It manifested itself in both typography and the adopted colors. Therefore, a professional visual identification system was developed for the project, which included a package of templates for word processor documents, presentation slides, typography, colors, fonts, etc., allowing for the unification of the visual side of the EURHISFIRM project each time.

The visual identification system functioned in both internal and external communication during the entire project, which allowed for the creation of a recognizable brand among stakeholder groups.

The main tool of external communication was the website of the project <https://eurhisfirm.eu>.

²⁰ M8.1

The website devoted to the EURHISFIRM project mainly includes the work of individual WP teams. Regularly published information allows stakeholders to observe the activities and progress of the project. It is a key aspect in terms of achieving the set goals, but also a tool stimulating interest in the EURHISFIRM project. In addition, the website contains descriptions and contacts to people directly related to the project, which makes it easier for interested parties to reach decision-makers in a specific area.

Also, connecting to a social networking site increases its effectiveness and penetration range of the target group.

Most of the promotional activities were based on social media linked on the website. The active use of the potential of Twitter or ResearchGate allowed for a regular increase in the group of recipients. The selection of the social media most appropriate to the information provided allowed to strengthen the overall philosophy of the EURHISFIRM project.

The key value of the project was also reaching a specific group of selected stakeholders with information about EURHISFIRM.

Due to the importance and invaluable nature of the project, it was important for various stakeholder groups from various environments to provide key information and interest in this unique, innovative and invaluable project.

The stakeholder group consisted of key decision makers, academia, business, society, financial institutions and regulatory institutions.

Information about the project has been successfully delivered to over 190 recipients all over Europe in 11 countries.

Such important direct communication allowed for the expansion of brand and project awareness on a significant scale.

Activities carried out in co-promotion by placing information about EURHISFIRM on websites and in social media of partner companies constituted another promotional effect for the project. Cooperation with partner projects and institutions, such as SSHOC and CESSDA ERIC, made it possible to reach an even wider audience. At the same time, the rank of these institutions and their recognition contributed to the positive strengthening of the EURHISFIRM brand.

Due to the COVID-19 global pandemic, the previously planned outreach events (e.g. an informational panel in Brussels, research dissemination and discussion forums, etc.) were not able to take place.

As a consequence of special restrictions resulting from direct contacts during the global pandemic, the EURHISFIRM promotion has been moved to online channels.

EURHISFIRM Twitter was also used to increase the users subscribed to the project's Twitter account by retweeting and replying to relevant material.

In addition, the international value of the project was actively promoted at the following conferences:

1. The Session at the World Economic History Conference planned in 2021 is postponed to 2022.
<https://wehc2021.sciencesconf.org/>
2. A workshop on “The Price of Everything, But the Value of Nothing” under the auspices of EURHISFIRM, Amsterdam, 27 November 2020
3. 4th Consortium meeting of the SSHOC,
8th and 9th September 2020, EURHISFIRM was presented after joining the SSHOC project
4. Presentation at the 2020 Business History Conference. Charlotte, North Carolina
12-14 of March 2020
Christopher Coyle, Robin Adams "The Wee Divergence: Entrepreneurship and Political Turmoil in Ireland Before 1900"
5. Presentations made by Angelo Riva who presented the Project:
 - a) Hommage à Georges Gallais-Hamonno,
Online, 5 th of October 2020.
Riva A., From GGH to Eurhisfirm. The ‘Big Data Revolution’ in financial history. Some experiments
 - b) Conseil Supérieur des Archives, Ministère de la Culture,
Paris, 11 th of December 2019.
Laperdrix M., Riva A., L’Exploitation “Big data” des Archives de la Compagnie des Agents de change de Paris.
 - c) Data for Financial History Workshop
Paris, 13th of December 2018.
Paris School of Economics, Riva A., Eurhisfim.
 - d) Les archives, patrimoine et richesse de l’action publique, 2ème Rencontre du réseau archives des ministères économiques et financiers,
Paris, Bercy, 10 th of April 2018.
Riva A., The “Big data” Revolution in Financial History. Some experiments”.
 - e) EABH (European Association of Banking and Financial History) Workshop “The data dilemma: a risk or an asset?” (joint event with INFUTURE 2017 International Conference “Integrating ICT in Society”),
Zagreb (Croatia), 8-10 of November 2017.
Riva A., “The ‘Big Data Revolution’ in banking and financial history? The French Experiments: Successes, Failures and Developments
7. On the 29th of June 2021 will be held the conference, where the milestone of WP5 at ICTeSSH2021 will be presented.
Title of presentation : An Extensible Model for Historical Financial Data with an Application to German Company and Stock Market Data.
Dennis Gram, Pantelis Karapanagiotis, Jan Krzyzanowski, Marius Liebald and Uwe Walz
8. Prepared a poster to promote the EURHISFIRM project at the Conference of the Verein für Sozialpolitik in Regensburg. Hanna Floto-Degener was presented with the poster.
Mobility and Migration in Historical Perspective III. Congress for Economic and Social History
20 - 22 of March 2019.

9. EURHISFIRM project information during the onsite evaluations at SAFE held in the past years (for entering the Leibniz Association). Since 2020 SAFE has been a member institute of the Leibniz Association, which is a union of 96 German non-university research institutes. The Leibniz Institutes cooperate closely with universities and are funded publicly by the federal government and the federal states (total budget of €1.9 billion).
10. Dissemination information by CESSDA ERIC. *Prepared by:* Martina Drasci, Ivana Ilijasic Versic, Ron Dekker
 - a) Twitter: Tweet on CESSDA presentation at project Kick-off (https://twitter.com/CESSDA_Data/status/1106593700128403458) (Impressions 514; engagements 20); Mar 15, 2019
 - b) Website:
 - Presenting EURHISFIRM project on CESSDA webpage ([https://www.cessda.eu/About/Projects/\(offset\)/5](https://www.cessda.eu/About/Projects/(offset)/5) and <https://www.cessda.eu/About/Projects/Current-projects/EURHISFIRM>) January 2021
 - News piece on EURHISFIRM joining SSHOC on SSHOC project website: (<https://www.sshopencloud.eu/news/eurhisfirm-joins-sshoc-partner-consortium>), 18 Sept 2020
 - Presenting EURHISFIRM project on SSHOC project website: <https://www.sshopencloud.eu/partners/eurhisfirm>; 18 Sept 2020
 - Article on EURHISFIRM on CESSDA website: <https://www.cessda.eu/News-Events/News/CESSDA/CESSDA-ERIC-joins-the-EURHISFIRM-project>, 11 Dec 2020
 - c) Annual report. CESSDA Annual Report 2020 - Item on EURHISFIRM in Section "EC Projects" (in publication) - will be available here: <https://www.cessda.eu/News-Events/Annual-Reports>
 - d) Presentations:
 - 13th CESSDA Service Providers' Forum, Bergen, 10 October 2019 - agenda item 13.08 Reports and Information from Main Office announcing CESSDA joining EURHISFIRM project (internal use only, not public)
 - 6th CESSDA General assembly, Copenhagen, 21 November 2019 - agenda item 6.06d Report from Main Office - EC projects on amendment aiming to add CESSDA to EURHISFIRM consortium
 - 14th CESSDA Service Providers' Forum, virtual, 7 April 2020 - agenda item 14.08 European issues, including EC projects overview (internal use only, not public)
 - Announcing EURHISFIRM project and members joining SSHOC project (SSHOC 2nd Consortium meeting, Florence, 14 Oct 2020, WP1 presentation, Martina Drascic - internal project use only, not public);
 - Introducing EURHISFIRM project and members to join to the SSHOC Strategic Board (SSHOC 2nd Consortium meeting - Strategic Board pre-meeting, Florence, 14 Oct 2020, Ivana Ilijasic Versic, Ron Dekker - internal use only, not public);
 - 8th CESSDA General Assembly, virtual, 24 November 2020 - agenda item 08.07 Report from Director/MO, EC projects overview (internal use only, not public)

- 16th CESSDA Service Providers' Forum, virtual, 15 April 2021 - agenda item 16.09
EC projects and CESSDA (overview, internal use only, not public)

WP1 (Project Management) assisted in the above tasks where needed for communication and dissemination in addition to its normal project management and report drafting duties to assure that the project's operations are fulfilled in the most efficient way possible and to ensure that the project progresses according to the timelines established. WP1 also targeted community building/outreach with other organizations (see [EURHISFIRM and the RI communities](#) in section II), disseminated the project's public documents on the [OpenAIRE](#) platform, and also coordinated and/or initiated communications with relevant organisations for realised and potential collaborations and for participation in pertinent EURHISFIRM activities, such as the general assembly. WP1 also represented EURHISFIRM in its participation in the SSHOC activities.

WP1, with the Executive Committee and Steering Committee, also worked on ensuring the project's subsequent steps after the design phase. This included confirming interest from other organizations outside of the current EURHISFIRM consortium for their participation in the project's future developments. Four institutions have thus responded in their support and interest in joining the consortium for phase II (see [letters of interest](#) in appendix).

III.4. Governance and cultural heritage (All Work Packages, and more specifically WP1, WP3, WP9, WP10, WP11)

Adherence to FAIR data principles

EURHISFIRM is committed to adhering to the FAIR data principles and being FAIR by design in its subsequent phases, as it has been during the course of the current phase by ensuring that the project is conceived right from the beginning as FAIR by design. These details are elaborated through the several versions of the data management plan created by WP1 with assistance from relevant WPs throughout the document's evolution. Indeed, as EURHISFIRM in its current phase was a design study, it could be viewed itself as a data management plan that evolved as the characteristics of the design were fleshed out by the various WPs throughout its course.

WP9 has also conducted a self-assessment on how the datasets and the software development of the RI would comply to the FAIR principles by using a tool (<https://www.andis-nectar-rds.org.au/fair-tool>) and found that they would be highly compliant to the principles, except for one condition within the "Accessible" terms concerning the requirement for specialised protocols or tools.²¹

In all other Work Packages, EURHISFIRM endeavored to being FAIR by design by incorporating these principles as much as possible, where possible. For example, WP6 uses Wikibase as part of its work, which

²¹ D9.1

is an open-source platform. Indeed, the data connection task of this WP tested the interoperability of two independent data sets, as would be the case when new types of data are added to the RI. The work from WP5 and WP9 support a federated data system designed to handle the interoperability of new data content (as well as with new languages and historical nuances, etc.) with the existing ones. Ultimately, the goal is to enable researchers to access a rich archive of stand-alone data sets that remain fully interoperable with one another. This would allow the possibility of not only accessing the data sets themselves, but to also studying the relationships between them and possibly uncover much more hidden research potential that have not been available to this day due to their inaccessibility.

Additionally, an external organization had expressed interest in EURHISFIRM's data, but ultimately this cooperation did not proceed due to the other party's preference for privatising the data, which EURHISFIRM believed would be contrary to the FAIR principles.

Additional details on the work completed by EURHISFIRM to ensure adherence to FAIR data principles are also described in section [IV.I.](#)

Business and governance model/sustainability

In order to ensure a long-term impact and a good return on investment of the project funding, EURHISFIRM is well aware of the importance of a sustainable business model. **WP10** (Business model and governance) worked on researching these models and selecting the one most appropriate for EURHISFIRM, based on documentations from relevant international and national structures, as well as on a survey administered to a large and diverse group of stakeholders in the eight countries participating in the network.

The proposed business model represents an approximation toward sustainability for EURHISFIRM in its transition from the implementation to the operation stage. The described scenarios will be continuously refined in light of the progress made with our user community. Therefore, WP10's outputs should not be considered as a commitment to a final plan to be executed exactly as it is currently formulated, but rather as an essential guidance in the process of developing our final business plan.

The business model must ensure the sustainability of EURHISFIRM as a data-oriented International Distributed Research Infrastructure specialized in the federation, management, storage and curation of large company-level historical datasets. EURHISFIRM will operate on the base of a central shared coordination model, with distributed and operational nodes, and a coordinating mechanism based on a central node.

During the design stage, we have identified CESSDA ERIC as a key strategic partner, inviting them to join the project. CESSDA ERIC is now a full member of the EURHISFIRM consortium. The participation of GESIS-Leibniz Institute for the Social Sciences (the largest German research infrastructure for social sciences) in EURHISFIRM has effectively ensured from the outset our co-ordination with CESSDA ERIC's standards and practices and in-depth knowledge on RI design. In the implementation stage, EURHISFIRM will deepen its coordination with CESSDA ERIC's standards and practices with the aim of integrating within its network of Service Providers in the operation stage. EURHISFIRM also aims to integrate into the European open cloud ecosystem for social sciences and humanities promoted by the SSHOC (Social Science and Humanities

Open Cloud) project, adopting the interoperability principles of FAIR data and developing cloud-ready tools. EURHISFIRM is also open to collaborative partnerships with other RIs, digital libraries, archives and repositories, such as DARIAH-EU, Europeana and Clarin.

EURHISFIRM will supply three types of services: 1) Data services based on a Wide Access mode to raw data to guarantee the broadest possible access; 2) Support, community building and training services to expand its user community; and 3) Infrastructure services for the production, standardization and enrichment of data from digitized primary sources. EURHISFIRM also contemplates the possibility to offer additional value-added services requested by specific users for academic or commercial purposes.

EURHISFIRM's key resources include: 1) Data Extraction using an Artificial Intelligence-based platform; 2) Data Merging in order to connect existing historical and contemporary company-level data; 3) a European Common Data Model with interfaces to the process architecture; 4) a Common Data Access Service to ensure the interoperability of different datasets; and 5) Data Visualization to give users maximum contextual information.

EURHISFIRM will generate both economic and socio-economic benefits. In terms of economic benefits, the availability of FAIR research data through EURHISFIRM will translate into a significant reduction in time costs, storage costs, license costs and research retraction. Jointly, time spent and storage costs account for 95% of the total cost of not having FAIR data, according to the European Commission. Their total estimated cost is €4,476 per researcher/year, equivalent to 8.2% of the average annual salary of an academic researcher. Their total estimated cost for the European research community in economic, financial and business history is estimated to €2 238 000 per year. This is a lower bound, as the community of EURHISFIRM users will include a large number of researchers who only occasionally use historical data, both in the academic sector and in research centers of public and private institutions, such as regulators or central and commercial banks.

Its socio-economic benefits include: 1) research-enhancing effects (through publications, citations, research grants, scientific users, research collaborations); 2) research training and other educational and outreach activities; 3) the production of expert advice and resources in support of public policies; 4) the development of collaborative projects with non-academic partners. The main beneficiaries of EURHISFIRM will be researchers, universities, libraries, archives, research funding organization, other Research Infrastructures and e-infrastructures (digital libraries, archives and repositories). The channels through which they can benefit from EURHISFIRM range from the use of its services to the establishment of long-term partnerships, up to full membership. However, non-academic research institutions, both public and private, will benefit from the existence of EURHISFIRM.

WP10's outputs also provide a first approximation to the human capital requirements of the coordinating central node and to its cost structure at the outset of the operation stage, including personnel costs and costs related to the use of infrastructures and licenses. This are estimated to €394,000 per year approximately.

As to funding, the costs of EURHISFIRM, including both the central node and the distributed nodes, will have to be covered by a core stream of long-term structural funding from public agencies. This is justified

by the fact that EURHISFIRM's main beneficiaries are the international research community and the general public. Structural funding will cover EURHISFIRM's basic functionality, usability and support services. The funding of the central node will be shared by the consortium members, with contributions based on country GDP and the possibility to contribute in-kind. Individual members will be responsible for funding the operation of distributed nodes. As a complementary revenue stream, EURHISFIRM will pursue project funding as well as contracts or fees for additional value-added services requested by specific users for academic or commercial purposes.

During the stage of implementation of the research infrastructure, the EURHISFIRM consortium will operate on the base of a Memorandum of Cooperation Agreement signed by the participating institutions. As it enters the operation stage, EURHISFIRM will become a separate and independent legal entity with the ability to hire employees and carry out commercial activities. The precise legal form of this independent entity will be decided at the end of the implementation stage.

At the operation stage, the governance model of this legal entity will be aligned to that of other RIs in the social sciences. It will be based on three main bodies: 1) General Assembly; 2) Executive Committee; 3) Scientific Advisory Board/Steering Committee. The General Assembly is the highest authority, composed by representatives of EURHISFIRM members and responsible for decisions about budget and long-term strategies. The Executive Committee reports to the Assembly and is composed of the Director of the central node, the Head of the Scientific Advisory Board/Steering Committee, and a Board of Directors (one for each participating country). The Steering Committee will be composed of experts from the EURHISFIRM network and from international independent experts. In addition, we contemplate the possibility of a permanent Stakeholder Forum as an interface with stakeholders who are not members (hence, are not represented in the General Assembly). Since EURHISFIRM will be a distributed RI, we are also considering the creation of a Committee of National Coordinators, to integrate and coordinate national activities at consortium level. Finally, for day-to-day management the Director will be supported by a Head Office located at the central node.

Legal governance

WP3 (Legal and ethical design) worked on the legal issues stemming from the protection of intellectual property (WP3.1) and the protection of privacy and personal data (WP3.2) within current and future data sources. Since there is no EU copyright code or an unified unfair competition regulation, WP3.1 referred to applicable EU regulations or directives to the extent possible. Otherwise the applicable law in Germany and the UK (which has recently left the European Union – the implications are being developed at the moment) was used to exemplify legal issues. Data privacy laws appear to have less impact on EURHISFIRM's current and potential data, since, in principle, only information on living natural persons are covered. Intellectual property laws vary more by the data source's characteristics.

(1) Protection of intellectual property (WP3.1)

Only some categories of source materials will be subject to intellectual property rights (IPRs); those are:

- Handbooks, newspapers and similar content displaying a minimum of creative effort published anonymously since 1. January 1951 or by named authors who died on or after 1. January 1951.
- Source databases with creatively selected or arranged content published anonymously since 1. January 1951 or by named editors who died on or after 1. January 1951.
- Other source databases published since 1. January 2006.

Other source materials are in the public domain, thus intellectual property law has no impact on their use. Source materials that are subject to IPRs may only be reproduced (copied) or made available to the public either with the authorization from the respective rightholders or on the basis of statutory limitations and exceptions to those rights. The most relevant exception is Art. 3 Digital Single Market Directive. This provision allows for text and data mining of source materials for the purposes of scientific research. This purpose requires the database to be operated on a non-commercial basis, i.e. to be publicly financed or to reinvest all profits into the database. Additionally only materials to which the EURHISFIRM members have lawful access can be text and data mined on the basis of Art. 3 Digital Single Market Directive. Thus, as long as the source materials are available in public libraries or access is lawfully acquired, they may be digitized and integrated into the EURHISFIRM database for the purposes of scientific research. Depending on the existing national laws and on the transposition of Art. 3 Digital Single Market Directive into national laws, the resulting corpus of the database may be made available to a limited circle of researchers for their joint research or to individual third persons for monitoring the quality of scientific research (see for example Sect. 60d para. 1 German Copyright Code). It is however not permissible to make the digitized source materials included in the database corpus available to the general public via the database. Instead the raw data needs to be extracted and may then be displayed in the database, since the IPR protection does not extend to raw data.

(2) Protection of privacy and personal data (WP3.2)

On the European level several legal rules or codifications of rules exist protecting privacy and personal data. In addition, the case law of the (European) courts plays a significant role interpreting and applying the quite often vague and ambiguous rules. This holds foremost for rules enacted on the highest levels of the norm hierarchy: Article 8(1) of the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR), Articles 7 and 8 of the Charter of Fundamental Rights of the European Union (CFR) which has been integrated into the primary law of the EU by the Treaty of Lisbon, Article 6(1) Treaty of the European Union (TEU). Another principal legal source of the primary law of the EU for assessing the protection of personal data is Article 16 of the Treaty on the Functioning of the European Union (TFEU).

For the practical work of EURHISFIRM, the secondary law of the EU, in specific the General Data Protection Regulation (GDPR), plays a predominant role, in general also for the parts of EURHISFIRM residing in the United Kingdom. The UK has retained so far the majority of the rules as national law. Other codifications of the secondary law are either not (yet) relevant for EURHISFIRM or are in the process of a general revision, like the Regulation on Electronic Communication which should have entered into force together with the GDPR. Even if an agreement seems to be in sight now this is a different field of law despite its

interconnection with the work of EURHISFIRM. In the course of the work on WP3.2, a (new) Regulation on Data Processing by EU Institutions (IDPR) was created, two years after the GDPR, laying down specific data protection rules which apply (only) to EU institutions, bodies, offices and agencies. So far the rules of this Regulation are not applicable for EURHISFIRM. In case, however, it will be decided to set-up the RI in the future as such a body, this Regulation will have to be obeyed by EURHISFIRM.

Aside from this, national law of the Member States is in principle precluded since the GDPR has to be considered as a conclusive and exhaustive codification of the field. In case of conflict, the national law must not be applied. The rules of the EU law prevail. National law is, however, insofar relevant as the EU law needs transformation or concretization or is explicitly allowed by opening clauses. Such limitations are rare as it was a crucial goal of the legislation to create a uniform set of rules in this field for the whole EU. Unfortunately, at least one of these derogations or limitations is relevant in view of EURHISFIRM's work: the exemptions for research or archives.

On the other side, it is *most important* to notice that the relevant legal rules cover only personal data, i.e. information about a *natural person*. The vast bulk of the information processed for and in the data-base will not fulfil this criterion with the effect that the GDPR is insofar *not applicable*. But exceptions exist. The name of natural persons might be included or attached to the information on e.g. a stock corporation even if it is of no significance for the research purposes.

Another very *crucial* limitation is the confinement to *living* persons. It is almost unanimous opinion backed by the legislative motives that only information on *living natural persons* is protected by the GDPR. Since the protection of the deceased does not belong to the domain of the Regulation, national law may contain *specific* rules to protect the memory of the deceased. The German law, for example, does not contain a general rule of this kind. Only protection against defamation is granted by the courts. This can easily be avoided by EURHISFIRM. Apart from this, the question has to be judged on a case by case basis in the affected Member State. Details are elaborated in the Report. To create a comprehensive synopsis of the national rules would be a task for future comparative legal research.

Unfortunately the GDPR does not contain a general exemption for research or archives. This was prevented by influential forces not considering sufficiently that free access to information and scientific research is also protected by the primary law and that the real dangers for privacy and personality rights lie somewhere else where the Regulation in effect shows little effectiveness.

The privileged treatment of data processing in the public interest or in the exercise of official duties might be helpful but in general the prerequisites will only be fulfilled by EURHISFIRM under certain circumstances. Details are left to the national law. The relaxations from the requirement of purpose limitation for archives and scientific research also refer to national law. Derogations by Union or Member State law from certain rights are allowed without prejudice for scientific or historical research as well as for statistical purposes. Processing for archival purposes has to be in the public interest. From this differentiation can be gathered that the law considers the research purposes always to be in the public interest. For details see the Report.

Under the influence of U.S. law firms and interested groups in Europe the talk on “ethics” and “ethical rules” have high season; probably disproportionate to the actual behavior. Often it is a linguistic misunderstanding since the terms “ethics” and “ethical” have in the English language a considerably wider meaning than categories of moral philosophy. Usually they also meant (company) policy or at most “code of conduct” with limited legal significance. In Europe and in specific in the context of the EU they are increasingly used as an instrument to enhance morality and to intrude into the private lives of the people, although not seldom they lack a sound and undisputed legal basis not to speak of an enactment by a democratically elected representation of the people. It is often executive bodies who impose burdens on the weaker parts this way, like their employees or applicants for subsidies or grants.

The most which is acceptable might be codes of conducts or guidelines (with no moral philosophical elevation) which constitutionally may not bind the courts. The primary law of the EU only knows regulations and directives as binding and orders in Article 283 (3) TFEU explicitly that guidelines are non-binding.

(3) Reports and Clearance Protocols

The clearance protocols for both intellectual property rights and personal data and privacy are detailed in the milestone produced by WP3 (Data Ownership and Ethics Protocol <https://eurhisfirm.eu/wp-content/uploads/2021/04/MS6-Data-Ownership-and-Ethics-Protocol-final.pdf>). In addition you can find more detailed information on IPR issues in the D3.1: Report on Intellectual Property Rights Design, available at: https://eurhisfirm.eu/wp-content/uploads/2020/05/EURHISFIRM_D.3.1.pdf. Detailed information on privacy and personal data protection can be found at: <https://eurhisfirm.eu/wp-content/uploads/2021/03/D3.2-Report-on-the-Legal-Issues-Related-to-the-Protection-of-Privacy-and-Personal-Data.pdf>

Cultural heritage

Though quantitative data form EURHISFIRM’s main concern, the project consortium recognizes the imperative of adding digitized material documenting the cultural dimension of the European corporate and financial experience to the RI. That dimension shows to best effect Europe’s defining characteristic as a region, as a continent: unity in diversity. **WP 11** (Cultural Heritage) identified a two-fold fundamental problem concerning the cultural heritage material that interests us. First, the material is highly diverse; second, unlike the quantitative material, it is randomly preserved, scattered over numerous collections, and poorly catalogued, so difficult to find. At the same time the material is of great importance to economic historians and it has great outreach potential to other disciplines and even the wider public.

It is in EURHISFIRM’s interest to solve this problem. This might be done, for instance, through launching initiatives, on the one hand, to develop a standard classification of relevant material, so national projects can start identifying and cataloguing it; and on the other, to bring institutions with big collections together for collaborating on themed exhibitions. Both might best be achieved through raising interest for

EURHISFIRM's goals with the European Association for Bank and Financial History (EABH, Frankfurt), an association of the historians and archivists of major European banks and other financial institutions.²²

Cultural heritage material also poses additional, serious IT complications if they are to be accessible for quantitative databases and other disciplines at the same time. Consequently EURHISFIRM needs to find a partner with which to develop ways of presenting the cultural heritage materials. A survey of 48 websites identified Europeana and CLARIN as desirable partners.

From the EURHISFIRM perspective, a collaboration with either would also serve very different objectives. Europeana is a vast and varied virtual museum, library, archive, and service provider rolled into one. As portal of and collaboration between Europe's main museums, libraries, and archives, Europeana. CLARIN is something entirely different, an RI for Language Resources and Technology which offers sophisticated tools for text analysis. Both would reinforce EURHISFIRM's outreach, but do so in different ways and for different audiences. Whereas Europeana is the ideal showcase for cultural heritage material in the widest sense of the word to a wide public, CLARIN would open up underused text resources for both the economic or economic history disciplines and for research by linguists and other humanities researchers.

EURHISFIRM should preferably team up with both. Europeana offers a wide field of opportunity since social, economic, and business history have, at the moment, surprisingly little presence in its many and varied exhibitions. While it is of course not up to EURHISFIRM to fill that lacuna, EURHISFIRM might provide an impetus for establishing such a presence. However, that appears neither easy nor obvious save for its most basic level, proposing themed exhibitions. All other options either require a formal, durable organization or would be best undertaken by EURHISFIRM as part of a bigger team of institutions aiming to establish and maintain a proper presence of social, economic, and business history on Europeana.

By contrast, linking up with CLARIN requires no formal organization or agreement. CLARIN is a federation of language data repositories, service centres, and knowledge centres which brings together digital language resources in any form with tools to discover, explore, exploit, annotate, analyse or combine them for researchers in the humanities and the social sciences. That is to say, all EURHISFIRM and/or its consortium members need to do is provide access to the textual databases built in the framework of stock exchange and corporation data collection projects for CLARIN members to work with. Links with CLARIN were quickly established following initial talks, leading to a trial on textual resources supplied by EURHISFIRM consortium members that started in March 2021.

Memorandum of understanding within the current EURHISFIRM members

While EURHISFIRM is currently finishing its phase I (design study), it has ambitions to continue onto the subsequent phases if successful in the future calls for projects.

²² D11.1 and D11.2

The [memorandum of understanding](#) (see appendix) is the formal agreement of the current members of EURHISFIRM that they commit to the future phases of the project and will implement the propositions from this report, including the FAIR and open science principles.

IV. Expected impact of EURHISFIRM

IV.1 Contributions to the FAIR data principles

As described in the section under III.4, [Adherence to FAIR data principles](#), EURHISFIRM ensured that the final output will be a solid, federated data format consistent with FAIR principles. As such, a working group (WGIS - 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), an enterprise architecture framework that provides a set of standardised guidelines that serves this purpose.²³

Moreover, EURHISFIRM has been designed to make data (1.) easily findable; (2.) openly accessible; (3.) interoperable; (4.) reusable. Individual WPs have been working to contribute to this goal in various ways. This includes the work done by WP5 and WGIS on the common data model and the EURHISFIRM Legal Entity Identifier (ELEI) standards based on the Legal Entity Identifier (LEI) standards developed by the Global LEI Foundation (GLEIF)²⁴. There is also the work done by WP6, which studied the linking and merging between the existing data, as well as their compatibility with the to-be digitised data from the other partner countries in the project. WP7 worked with Data Source Standards in order to explore the most suitable methods for transforming the sources into exportable outputs. WP8 completed a survey with interested stakeholders in academia, business, and policy in order to understand the users' needs and interests concerning long-term financial company-level data. WP9 studied and recommended infrastructure policy and architecture that would optimally support the sustainability, and therefore the reusability, of the data. WP10 analysed the most viable methods for operational sustainability through business and governance models suitable for the RI. WP1 and other contributing WPs also handled the Data Management Plan (which in itself could be seen as the EURHISFIRM project itself, as it is a design project) in order to ensure that the design incorporates the FAIR principles outlined in the Data Management Plan templates. Finally, WP11 explored the ways that digitised historical financial data can be used to promote and deepen European research, culture, and heritage. The main findings of the respective WPs have been summarised in this final report; for more information, see their respective deliverables, which are all referenced in this final report as well as available on <https://eurhisfirm.eu/index.php/publications/> (for public deliverables).

The expected impact of the project's incorporation of the FAIR principles is that 1) more historical financial data will be made available as well as accessible, and 2) researchers will be able to reuse and recombine these heterogeneous data, which would otherwise not be possible, in order to generate new research

²³ TOGAF: <http://theopengroup.org/>.

²⁴ Global Legal Entity Identifier Foundation, 2019

insights and discoveries. Indeed, EURHISFIRM's vision is to unleash the potential of this data that currently remains inaccessible, in order to strengthen research possibilities in Europe.

EURHISFIRM's involvement within the RI community also reinforces its commitment to open science and FAIR principles. As part of the [SSHOC project, which is an ESFRI Science Cluster and a proponent of open science](#)²⁵, EURHISFIRM will also be integrated within this effort and will seek to collaborate with other open science and FAIR-principle guided research organisations and communities in the future.

IV.2 Contributions to research, policy making, and socioeconomic impact

Scientific research, government policy, and society as a whole may benefit from historical data to properly understand the dynamics of the past and how these affect the present and the future. Creating the data to develop this knowledge requires various interdisciplinary skills, some of which are specific to a country, or even to a region, because of the heterogeneity of historical business rules and practices. These peculiarities call for an ad hoc research infrastructure (RI) that can also connect to other existing systems.

Europe in the 21st century faces and has faced a number of unprecedented challenges which are and were inextricably linked to socioeconomic repercussions, such as health, education, and employment. These effects not only affect the immediate aftermath of the situation, but many years as well thereafter. However, it is also important to remember that all events are also linked to those that came before them, i.e. history and the past. In addition, history teaches us how possibly similar challenges have arisen in the past and how these have been dealt in good (or otherwise) ways, and how we could learn from these to avoid repeating the same mistakes and to form better policies that could benefit societies. In order to analyse these types of events, as well as ours, reliable empirical and factual data are needed. However, the lack of this type of data prevents such studies from being realised, presenting a significant obstacle to uncovering a rich mine of information that could be used to inform policies for the good of society.

The EURHISFIRM project was conceived to meet the need for such a benchmark RI in Europe that would be the gateway to providing quality empirical data for policy. We believe the designed infrastructure to have a positive impact on research and economic policy making, which in the long run will inevitably cause socio-economic advances. EURHISFIRM sets out to obtain this goal in a number of ways.

Firstly, it challenges the US's dominant position in data production. Before, a lack of data on the European level meant American companies were frequently and implicitly deemed "representative" or "the norm". However, this may have led to findings which are not directly applicable on the European level. EURHISFIRM attempts to address this issue by offering data based directly on European cases thus ensuring a higher validity in European financial research, also resulting in better informed policy decisions. EURHISFIRM's commitment to the FAIR principles ensures that the data will be accessible, findable, interoperable and reusable, which would further enable its potential to impact policy making in positive ways.

²⁵

<https://sshopencloud.eu/news/esfri-science-clusters-position-statement-expectations-and-long-term-commitment-open-science>

Secondly, the creation of EURHISFIRM leads to the enrichment of social science data to be “cross-studied” in concordance with the existing RI ecosystems and partners, such as SSHOC, CLARIN, EOSC, and CESSDA ERIC. Naturally, by creating a pan-European research infrastructure, EURHISFIRM also fosters further collaboration in financial research across European countries which will undoubtedly have a positive effect on research and informed policy making.

Finally, and more generally speaking, by making data on the European level more accessible, EURHISFIRM encourages and improves (young) scholars’ participation and training. Not only will EURHISFIRM help researchers save time by presenting relevant research data in a digitised form, but it also serves as a quality check by making such harmonised data accessible via NICs.

V. Conclusions and future directions

As stated in the previous sections, the EURHISFIRM project addresses the crucial need for a reliable and standardised long-term company-level data in Europe. It benefits European citizens and societies, the scientific domain, public policy and public organisations, and the private sector.

The EURHISFIRM project succeeded in designing a research infrastructure compatible with open science and FAIR data principles that will provide a comprehensive platform to access heterogeneous European long-term company level data in a standardised, reliable, scientifically sound, and technologically advanced way.

To accomplish this goal, EURHISFIRM was comprised of 11 Work Packages that can be roughly grouped into the following categories: economic history foundations, Information technology work (data extraction technologies, common data modelling, infrastructure architecture), practical aspects of the infrastructure operations (business plan [including user target research], legal plan, and cultural heritage of the data produced), project administration (communication, community building, logistics and strategy/vision [including compliance to open science frameworks and FAIR data]). These Work Packages have managed to progress in accordance with the timeline set out in the project proposal. Their main contributions have been summarised in sections II and III.

Due to the COVID-19 pandemic, the project encountered some delays and deviations from original plans, despite our efforts to mitigate these as much as possible. We had already discussed this with the project officer and had anticipated that we will encounter difficulties, and the project officer had let us know that this will be acknowledged and that we are to include notes on these in the reporting tool on the submission platform. However, despite these changes, the project accomplished its intended goals of completing an RI design that is ready for continuing in its subsequent phases of development. Additionally, the original plans for expanding the EURHISFIRM research community (for example, through participation in related in-person conferences and organising events pertinent to the research infrastructure community) were cancelled and/or changed. Despite these setbacks, the project aimed to continue its efforts to extend its research community by focusing on these efforts exclusively online to collaborate with relevant organisations.

Looking towards the future, the EURHISFIRM project will continue to focus on the following key priorities:

- Continuing to engage with the European research infrastructure community and technological advancements.
- Expanding the project's community by increasing citizen science/public engagement in the project (possible ideas include: analysis on stock markets, contribution to new data).
- Sharpening the long-term vision of EURHISFIRM in terms of utility, flexibility and sustainability, considering recent developments in the European research infrastructure, and research communities, as well as possible new sources of data whose characteristics may deviate from the ones existing in the EURHISFIRM RI.
- Thinking of ways to deal with more sources in social science data beyond stock exchange data.

To this end, it should be noted that a growing number of parties have expressed their interest in expanding on the foundations of the EURHISFIRM infrastructure. Apart from the aforementioned NEDHISFIRM project in the Netherlands, these interested parties include the institutions that have signed the letters of commitment in the Appendix.

After the completion of the design phase within this current project, EURHISFIRM looks forward to concretizing the results from this design study in the future phases in order to provide the European research community, policymakers, the public and private sector with sound empirical historical company-level data to serve a better society.

VI. REFERENCES

Reports

Adams et al., EURHISFIRM D8.4 Synthesis Report, 2019.

Cule, EURHISFIRM D6.1: Report on data matching issues and methodologies, 2020.

Cule, EURHISFIRM D6.2: Report on data connecting issues and methodologies, 2020.

Cule et al., EURHISFIRM M6.1: Data Matching Case Study, 2020.

Cule et al., EURHISFIRM M6.2: Data Connecting Case Study.

Ducros et al., EURHISFIRM D1.8: Third Data Management Plan, 2020.

Karapanagiotis, EURHISFIRM D5.1: Technical document on national data models, 2019.

Poukens, EURHISFIRM D4.2: Report on the Inventory of Data and Sources, 2018.

Ranft et al., EURHISFIRM D5.4: Updated technical document on the preliminary data model, 2021.

All other public deliverables and milestones are available at: <https://eurhisfirm.eu/index.php/publications/> and on the EURHISFIRM OpenAIRE page: https://explore.openaire.eu/search/project?projectId=corda_h2020::612830f55f1f92d36a5477538163d4e5.

Websites

CESSDA ERIC: <https://www.cessda.eu/>.

CLARIN: <https://www.clarin.eu/>.

DARIAH: <https://www.dariah.eu/>.

Europeana: <https://www.europeana.eu/en>.

EURHISFIRM OpenAIRE page: https://explore.openaire.eu/search/project?projectId=corda_h2020::612830f55f1f92d36a5477538163d4e5.

EURHISFIRM website: <https://eurhisfirm.eu/>.

EOSC: <https://eosc-portal.eu/> and SSHOC: <https://sshopencloud.eu/>.

ESFRI Landmarks: <http://roadmap2018.esfri.eu/projects-and-landmarks/>.

Huma-Num: <https://www.huma-num.fr/>.

TOGAF: <http://theopengroup.org/>.

VII. APPENDIX

VII.1. Quick Installation Guide (slide deck) for amending existing files of historical firm data by the EURHISFIRM Legal Entity Identifier (ELEI)

(An example on how to use this technology also for unique identification of other objects, e.g. the EURHISFIRM Financial Instruments Identifier (EFII), data sources, etc.)

1.

Long-term data for Europe
EURHISFIRM

ELEI Assignment and Resolution
Data Submission Unit (DSU) and
Network Integration Center (NIC)
"Quick Start" Guide

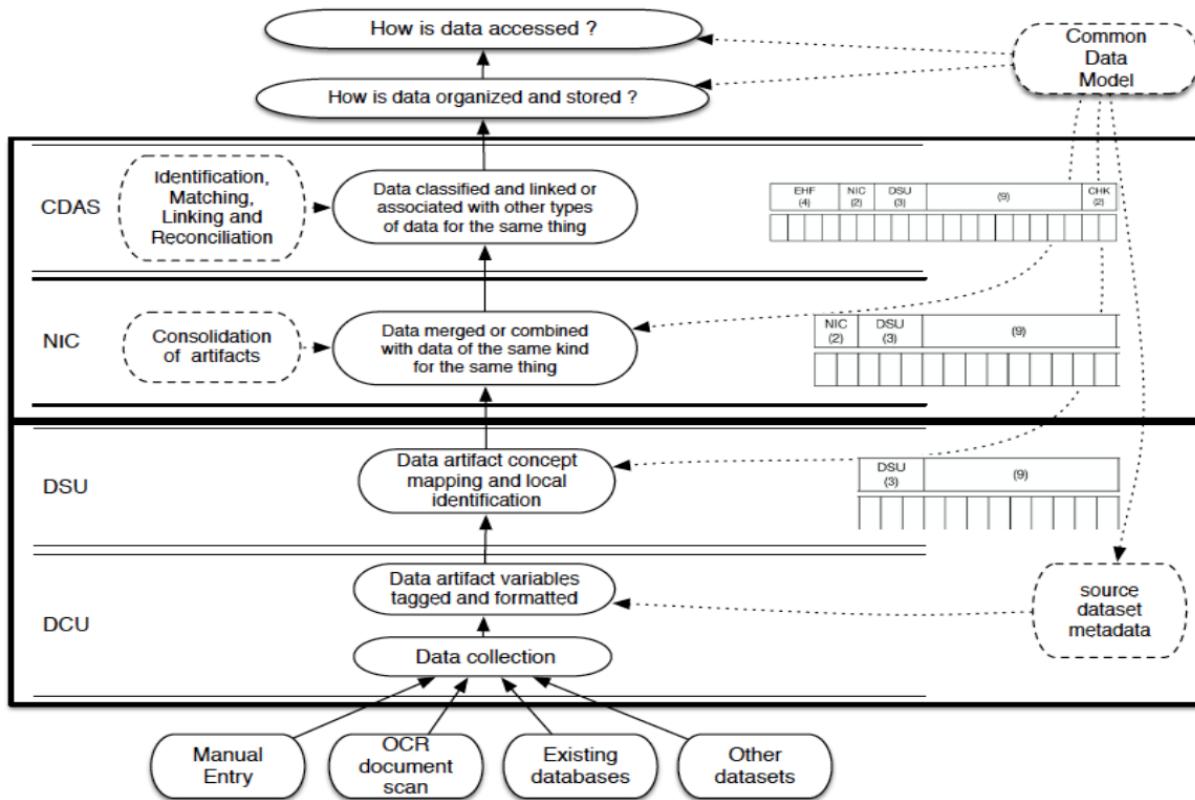
Jefferson Braswell



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

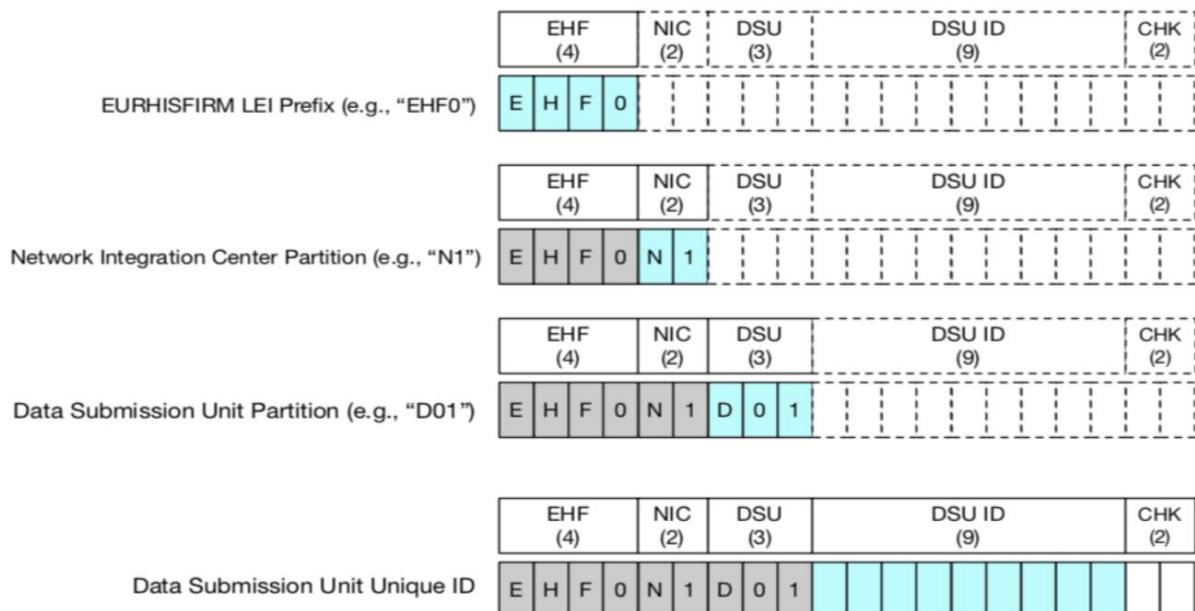
<https://eurhisfirm.eu>

2.



3.

ELEI Code Structure Partitions

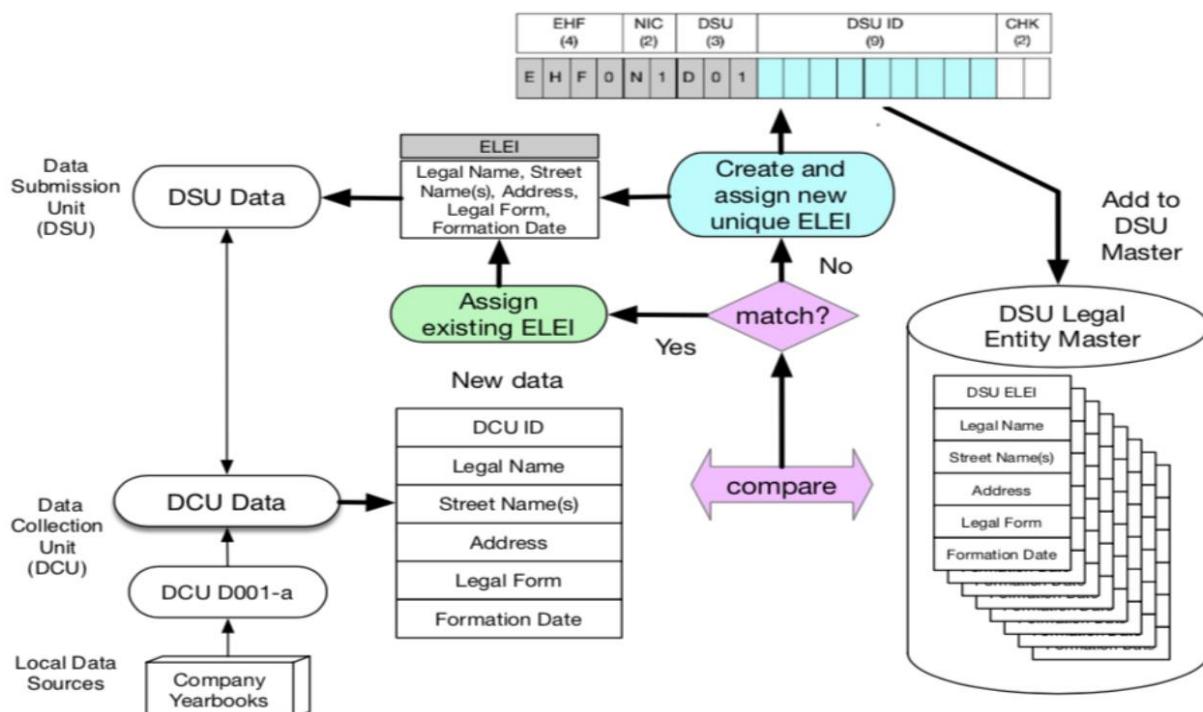


4.

The following slide depicts the process by which historical data that is collected on companies from local sources is compared with previous data that has been collected in order to:

- Determine, for each company for which data has been collected:
 - 1) If the company has already been encountered in the local data collection process -- in which case the ELEI that was previously assigned to the company is applied to the recent data, or
 - 2) The identifying reference data for the company does not match the reference data for any other company in the local company master – in which case a new ELEI is generated and assigned to the new company, and the local company master is updated with a new addition

5.

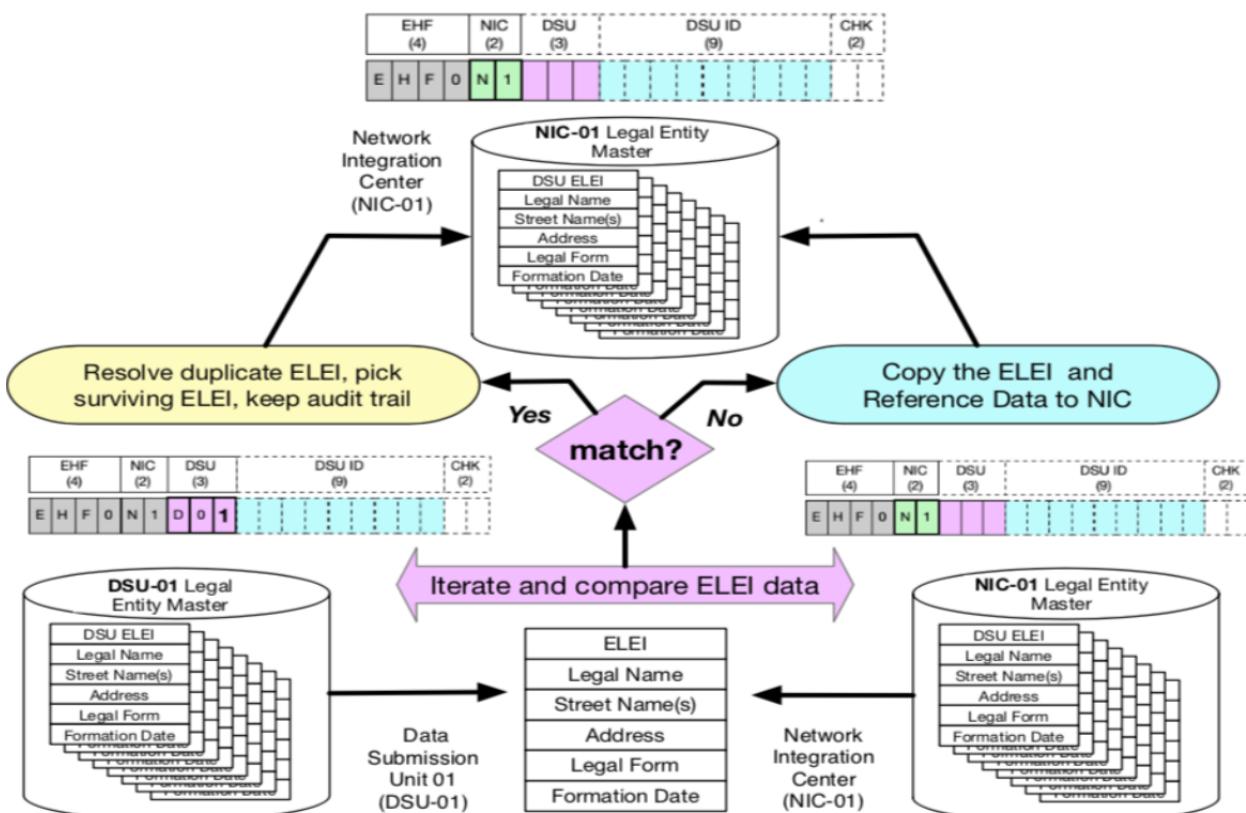


6.

The following slide depicts the process by which ELEI data that is submitted to Network Integration Centers (NICs) by Data Submission Units (DSUs) are analyzed in order to:

- Determine, by comparing the identifying ELEI reference data in DSU submissions, for each instance of ELEI data that is submitted:
 - 1) If a DSU has identified a company from their respective local data collection sources that has already been identified and assigned an ELEI -- in which case the previously assigned ELEI would survive and the ELEI submitted by the DSU would be designated as a *Duplicate* (but retained in the date lineage audit trail for drill-down purposes back to the source);
 - 2) ELEI data whose identifying reference data are determined to not match the reference data of any other ELEI in the NIC ELEI company master would be added to the NIC ELEI company master

7.



VII.2. Memorandum of understanding within EURHISFIRM members

MEMORANDUM OF UNDERSTANDING

THIS MEMORANDUM OF UNDERSTANDING (MOU) is made on 17 June 2021 by and between:

1. Paris School of Economics (EEP-PSE), the Coordinator
2. University of Antwerp (UANTWERP)
3. Goethe University Frankfurt (GUF)
4. Erasmus Rotterdam University (EUR)
5. Wroclaw University of Economics and Business (UE WE WROCLAWIU)
6. The Queen's University of Belfast (QUB)
7. Royal Dutch Academy of Science (International Institute for Social History – IISH) (KNAW)
8. Universidad Carlos III de Madrid (UC3M)
9. Université de Rouen Normandie (URN)

Acting in the name and on behalf of Laboratory LITIS, located at Université de Rouen Normandie, UFR Sciences et Techniques, Avenue de l'Université 76800 Saint-Étienne-du-Rouvray, France, represented by Mr. Laurent Heutte

10. Institut National des Sciences Appliquées de Rennes (INSA RENNES)

Scientific, Cultural and Professional Public Establishment, whose registered office is located 20 avenue des Buttes de Coësmes, CS 70839, 35708 Rennes cedex, represented by its Director, M'hamed DRISSI,

Hereinafter referred to as "INSA Rennes"

Acting on its behalf and by mandate of the Centre National de la Recherche Scientifique (CNRS) for the activities of the Institut de Recherche en Informatique et Systèmes Aléatoires, joint research unit (UMR) 6074, managed by Jean-Marc JEZEQUEL, hereinafter referred to as "IRISA"

11. GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS)
12. Consortium of European Social Science Data Archives European Research Infrastructure Consortium (CESSDA ERIC)

Individually referred to as a “Party” or collectively as the “Parties”.

WHEREAS

The Parties (with the exception of CESSDA ERIC, who joined the project after it has been established), initiated contacts between themselves and submitted a proposal for a collaborative project in response to the H2020-INFRADEV-2017-1 call for projects launched by *the European Commission* (called the Grantor);

The Parties have agreed to name this collaborative project as *EURHISFIRM* (hereinafter referred to as the “Project”);

The Parties, having received a positive evaluation on the proposal for phase I of the Project (herein referred to as Proposal I) by the Grantor, have received the signature of the Grant Agreement and have negotiate between them a Consortium Agreement, in which the last version was signed in December 2020 and is currently in effect;

The Parties acknowledge that a Party may at any time wish to stop its involvement in the Project;

The Parties agree to prepare and complete the call for projects under the Horizon Europe framework for the phase II of the Project (herein referred to as Proposal II) in order to continue the work established under phase I of the Project.

For the purpose of the submission of Proposal II and for the preparation and negotiation of a future grant agreement and a consortium agreement for this phase II of the Project (herein referred to as Grant Agreement II and Consortium Agreement II), the Parties intend to disclose information to each other, which they wish to keep confidential.

THE PARTIES AGREE AS FOLLOWS:

1. Proposal preparation

- 1.1. The Parties wish to prepare and submit together the Proposal II as explained in the preamble.
- 1.2. The Parties agree that each Party shall not prepare or submit any additional proposal with the same research activities for the topic identified in Proposal II;
- 1.3. The Parties agree that the consortium may be expanded with additional members, including those who have explicitly expressed interest to do so (cf. the signatories of the “Letters of interest in joining future phases of EURHISFIRM from external institutions”)
- 1.4. With the purpose to prepare and submit the Proposal II in due time, the Parties agree:
 - i) to designate a Coordinator of the project (hereinafter referred to as the “Coordinator” to represent the Parties towards the Grantor in due time;

- ii) to meet or correspond as necessary to prepare and decide the details of the Proposal II;
- ii) that each Party shall use its best endeavours to prepare all the documents, data and information necessary for the preparation of the Proposal II and to provide them to the Coordinator in due time. In particular, each Party shall provide the Coordinator with its participant identification code (PIC).

1.5. The Coordinator agrees not to modify, without previous consent, any document, data or information supplied by the other Parties.

1.6. The Coordinator shall keep the Parties informed of the progress of the preparation of Proposal II and, at any Party's request, it shall make available a copy of all letters, emails or any other documents concerning the Proposal II that were sent to the Grantor or received from it before the submission of the Proposal II.

1.7. The Parties agree to prepare the Proposal II with a commitment to implement the principles and work completed in the Project's phase I, including FAIR (findable, accessible, interoperable, and reusable) and open science principles.

2. Grant Preparation

2.1. Provided that the Proposal II has a positive evaluation and that the Parties are invited to prepare the Grant Agreement II, the Parties wish to collaborate with the purpose to conclude a Grant Agreement II for the Grantor.

2.2. The Parties agree that the Coordinator shall be responsible for conducting the preparation foreseen under clause 2.1. of this MoU.

2.3. The Coordinator shall keep the Parties informed of the progress of the grant preparation and, at any Party's request, it shall make available a copy of all letters, emails or any other documents that were sent to the Grantor or received from it for this purpose, before the signature of the Grant Agreement II. The Coordinator shall, in any case, send to the other Parties a copy of the invitation to prepare the grant.

2.4. Upon request of the Coordinator, the Parties shall attend the meetings with the Grantor.

2.5. The Parties agree to assist the Coordinator in the preparation and to provide it with the necessary documents, data and information in order to allow the signature of the Grant Agreement II in due time.

2.6. Adjustments shall be negotiated in good faith by the Parties. Any adaptation or modification concerning the work packages shall be accepted by the Coordinator only with the prior written agreement of the Party concerned.

3. Negotiations of the Consortium Agreement

3.1. Provided that the Proposal II has a positive evaluation and that the Parties are invited to sign the Grant Agreement II, the Parties wish to conclude the Consortium Agreement II before the signature of the Grant Agreement II. The conclusion of such a Consortium Agreement II is dependent on mutual consent and must be reduced to written form.

3.2. The Parties agree that the Coordinator shall be responsible for conducting the negotiations foreseen under clause 3.1. of this MoU.

4. **Non-binding effect of this MoU**

The matters set forth in this MoU constitute an expression of the mutual consent of the Parties and do not constitute a binding agreement between the Parties. Any such binding agreement will only arise once the Consortium Agreement II is signed by the Parties.

By attaching this document into the final reports of the EURHISFIRM Project, the Parties hereto have caused this MoU to be executed as of the date stated above.

VII.3. Letters of interest in joining future phases of EURHISFIRM from external institutions

| Institution name | Country | Signatory | Position |
|--|-------------|-------------------|----------------------------------|
| Leibniz Institute for Financial Research SAFE (Sustainable Architecture for Finance in Europe) | Germany | Uwe Walz | Deputy scientific director |
| | | Muriel Büscher | Administrative director |
| Luxembourg Centre for Contemporary and Digital History | Luxembourg | Andreas Fickers | Director |
| | | Benoît Majerus | Professor |
| NOVA School of Business and Economics | Portugal | Pedro Vicente | Head of the research office |
| ISEG Lisbon School of Economics & Management, Universidade de Lisboa | Portugal | João Peixoto | Vice-Dean for Scientific Affairs |
| Fondation pour l'Institut de Hautes Etudes Internationales et du Développement (IHEID) | Switzerland | Nathan Sussman | Director, full professor |
| Binghamton University, on behalf of the Stockholm School of Economics | USA/Sweden | Kristian Rydqvist | Professor of Finance & Economics |

École d'Économie de Paris - Paris School of Economics
Angelo Riva
Boulevard Jourdan 48
75014 Paris
France

Leibniz Institute for
Financial Research SAFE e.V.
Theodor-W.-Adorno-Platz 3
60629 Frankfurt am Main
Phone + 49 (0) 69 798 30067
demoor@safe-frankfurt.de
www.safe-frankfurt.de

Tax No.: 0452558687
VAT Reg. No.: 325945321
Register Court: Frankfurt am Main
Register number: VR16523
Management Board:
J. Krahen, M. Büscher, U. Walz

Frankfurt, 7 May 2021

REF: Letter of Commitment to the phase II of the EURHISFIRM

Dear Mr. Riva,

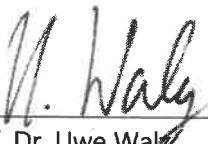
We, herewith undersigned, Prof. Dr. Uwe Walz and Dr. Muriel Büscher, on behalf of the Leibniz Institute for Financial Research SAFE ("Sustainable Architecture for Finance in Europe") (hereinafter SAFE), with VAT Reg. No.: DE325945321 and established in Germany, welcome the opportunity to participate as Partner Organisation in the subsequent phase II of the EURHISFIRM European research infrastructure project for historical financial data, under the following terms and conditions.

SAFE is a Research Institution whose mission is to produce original, high-quality research and research-based policy advice in all areas of finance, with a special focus on Europe. EURHISFIRM, which is currently in progress under phase I, participates in the INFRADEV-01-2017 framework under the H2020-EU.1.4.1.1. - Developing new world-class research infrastructures programme funded by the European Commission's Horizon 2020 programme. The project, consisting of a consortium of 12 European institutions, is 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.

SAFE will contribute to the future objectives of the phase II of the project. In addition, SAFE will participate throughout the entire duration of phase II of the EURHISFIRM project as a member of the Steering Committee, attending the relevant meetings and network events.

Yours sincerely,
Leibniz-Institut für Finanzmarktforschung SAFE e.V.
Goethe-Universität Frankfurt
House of Finance
Theodor-W.-Adorno-Platz 3
60323 Frankfurt am Main

Designation and official stamp of the Organization



Prof. Dr. Uwe Walz
Deputy Scientific Director



Dr. Muriel Büscher
Administrative Director

ESCH SUR ALZETTE, 31st of March 2021

REF: Letter of Commitment to the phase II of the EURHISFIRM project

Dear Madam/Sir,

I, herewith undersigned, Prof. Benoît Majerus, on behalf of the Luxembourg Centre for Contemporary and Digital History (hereinafter C²DH), with VAT number LU 19805732 and established in LUXEMBOURG, welcome the opportunity to participate as Partner Organisation in the subsequent phase II of the EURHISFIRM European research infrastructure project for historical financial data, under the following terms and conditions.

C²DH is the University of Luxembourg's third interdisciplinary research centre, focusing on high-quality research, analysis and public dissemination in the field of contemporary Luxembourgish and European history. It promotes an interdisciplinary approach with a particular focus on new digital methods and tools for historical research and teaching. EURHISFIRM, which is currently in progress under phase I, participates in the [INFRADEV-01-2017](#) framework under the [H2020-EU.1.4.1.1. - Developing new world-class research infrastructures](#) programme funded by the European Commission's Horizon 2020 programme. The project, consisting of a consortium of [12 European institutions](#), is 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.

C²DH will contribute to the future objectives of the phase II of the project. In addition, C²DH will participate throughout the entire duration of phase II of the EURHISFIRM project as a member of the Steering Committee, attending the relevant meetings and network events.

Yours sincerely,



Prof. Dr. Andreas FICKERS
C²DH DIRECTOR

Prof. Benoît Majerus
C²DH

Lisbon, 05 April 2021

REF: Letter of Commitment to the phase II of the EURHISFIRM project

Dear Madam/Sir,

I, herewith undersigned, Prof. MARIA EUGÉNIA MATA, on behalf of the Nova School of Business and Economics (hereinafter Nova SBE), with VAT number 501559094 and established in PORTUGAL, welcome the opportunity to participate as Partner Organisation in the subsequent phase II of the EURHISFIRM European research infrastructure project for historical financial data, under the following terms and conditions.

Nova SBE is an Academic Institution aiming “to be a community dedicated to the development of talent and knowledge that impacts the world through degree and non-degree programs, generating cutting-edge knowledge, both academic and applied. Nova SBE **spreads this footprint** beyond Portugal, reaching Europe and the world to establish a culture of impact, based on our graduates’ achievements, for a more equitable, sustainable and **peaceful world**”.

EURHISFIRM, which is currently in progress under phase I, participates in the INFRADEV-01-2017 framework under the H2020-EU.1.4.1.1. - Developing new world-class research infrastructures programme funded by the European Commission’s Horizon 2020 programme. The project, consisting of a consortium of 12 European institutions, is 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.

Nova SBE will contribute to the future objectives of the phase II of the project. In addition, Nova SBE will participate throughout the entire duration of phase II of the EURHISFIRM project as a member of the Steering Committee, attending the relevant meetings and network events.

Yours sincerely,

Maria Eugénia Mata

NAME OF LEGAL REPRESENTATIVE: Pedro Vicente, Professor

POSITION OF LEGAL REPRESENTATIVE: Head of the Research Office

HAND-WRITTEN SIGNATURE OF LEGAL REPRESENTATIVE:



Lisbon, 7 June 2021

REF: Letter of Commitment to the phase II of the EURHISFIRM project

Dear Madam/Sir,

I, herewith undersigned, Prof. João Peixoto, legal representative, on behalf of the Lisbon School of Economics and Management, University of Lisbon (hereinafter ISEG), with VAT number 502 488 603 and established in Lisbon, welcome the opportunity to participate as Partner Organisation in the subsequent phase II of the EURHISFIRM European research infrastructure project for historical financial data, under the following terms and conditions.

ISEG is an Academic Institution aiming to create, share and enhance the socio-economic value of knowledge and culture in the fields of Economics, Finance and Business, through an approach based on plurality, the guarantee of freedom of intellectual and scientific expression, and respect for ethical principles and social responsibility. EURHISFIRM, which is currently in progress under phase I, participates in the [INFRADEV-01-2017](#) framework under the [H2020-EU.1.4.1.1. - Developing new world-class research infrastructures](#) programme funded by the European Commission's Horizon 2020 programme. The project, consisting of a consortium of [12 European institutions](#), is 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.

ISEG will contribute to the future objectives of the phase II of the project. In addition, ISEG will participate throughout the entire duration of phase II of the EURHISFIRM project as a member of the Steering Committee, attending the relevant meetings and network events.

Yours sincerely,



João Peixoto

Vice-Dean for Scientific Affairs

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

Geneva, 01 April 2021

REF: Letter of Commitment to the phase II of the EURHISFIRM project

Dear Madam/Sir,

I, herewith undersigned, Prof. Nathan Sussman, on behalf of the Fondation Pour L'Institut de Hautes Etudes Internationales et du Développement (hereinafter IHEID), with VAT number CHE-113.890.516 TVA and established in Switzerland, welcome the opportunity to participate as Partner Organisation in the subsequent phase II of the EURHISFIRM European research infrastructure project for historical financial data, under the following terms and conditions.

The IHEID is an Academic Institution aiming at the study of world affairs, with a particular emphasis on the cross-cutting fields of international relations and development issues. The IHEID aims to combine world-level research on social sciences with policy impact and engagement with international organisations, NGOs, governments and multinational companies.

EURHISFIRM, which is currently in progress under phase I, participates in the INFRADEV-01-2017 framework under the H2020-EU.1.4.1.1. - Developing new world-class research infrastructures programme funded by the European Commission's Horizon 2020 programme. The project, consisting of a consortium of 12 European institutions, is 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 IHEID will contribute to the future objectives of the phase II of the project. In addition, the IHEID will participate throughout the entire duration of phase II of the EURHISFIRM project as a member of the Steering Committee, attending the relevant meetings and network events.

Yours sincerely,

Nathan Sussman

Nathan Sussman
Full Professor, International Economics & Pictet Chair in Finance and Development
Director, Centre for Finance and Development

Binghamton University

Kristian Rydqvist
Professor of Finance & Economics
School of Management
Binghamton, New York 13902
Phone: +1 607 777 2673
Fax: +1 607 777 4422
Email: rydqvist@binghamton.edu

4 May 2021

REF: Letter of Commitment to the phase II of the EURHISFIRM project
Dear Madam/Sir,

I, herewith undersigned, Prof. Kristian Rydqvist, on behalf of the Stockholm School of Economics (hereinafter SSE), established in Sweden, welcome the opportunity to participate as Partner Organisation in the subsequent phase II of the EURHISFIRM European research infrastructure project for historical financial data, under the following terms and conditions.

SSE is an academic Institution aiming at research and education. EURHISFIRM, which is currently in progress under phase I, participates in the INFRADEV-01-2017 framework under the H2020-EU.1.4.1.1. - Developing new world-class research infrastructures programme funded by the European Commissions Horizon 2020 programme. The project, consisting of a consortium of 12 European institutions, is 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.

SSE will contribute to the future objectives of the phase II of the project. In addition, SSE will participate throughout the entire duration of phase II of the EURHISFIRM project as a member of the Steering Committee, attending the relevant meetings and network events.

Sincerely,



Kristian Rydqvist