Long-term data for Europe

EURHISFIRM

D1.8: Third Data Management Plan

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:

Jérémy DUCROS *(École d'Économie de Paris)*

Johan POUKENS *(Universiteit Antwerpen)*

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

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

APPROVED IN 2020 BY:

Jan ANNAERT *(Universiteit Antwerpen)*

Wolfgang KÖNIG *(Goethe-Universität Frankfurt)*

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

ACKNOWLEDGMENTS:

We thank the following colleagues for their contributions to this document:

Bertrand COÜASNON *(Institut national des sciences appliquées de Rennes)*

Boris CULE *(Universiteit Antwerpen)*

Pantelis KARAPANAGIOTIS *(Goethe-Universität Frankfurt)*

Thierry PAQUET *(Université de Rouen Normandie)*

Lukas Manuel RANFT *(Goethe-Universität Frankfurt)*

Table of Contents

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

Revision history

Version	***Date***	***Notes***
1.0	12/09/2018	First draft (D1.2)
1.1	21/09/2018	First corrections
1.2	24/09/2018	Second corrections
1.3	28/09/2018	Third corrections (final version of version 1)
2.0	04/03/2019	First draft of version 2 (D1.7)
2.1	06/03/2019	First corrections of version 2
2.2	11/03/2019	Second corrections of version 2
2.3	22/03/2019	Third corrections of version 2
3.0	20/04/2020	First draft of version 3 (D1.8)
3.1	23/04/2020	First corrections of version 3
3.2	28/04/2020	Second corrections of version 3

Executive summary

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

The EURHISFIRM project will be executed as much as possible in cooperation with infrastructures in the field of social sciences and humanities, such as 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), and the SSHOC (Social Sciences & Humanities Open Cloud).

Some notable developments since the second version of this Data Management Plan (DMP) (written in March 2019) are 1) the selection metadata standards (the DDI family of standards) and controlled vocabularies (the Thesaurus for Economics (STW) and the Getty Thesaurus of Geographical Names (TGN)) for coding the data's properties and ensuring their adherence to FAIR principles, and 2) the ongoing enhancement of the LEI (Legal Entity Identifier) for adapting to historical nuances and context specific to the EURHISFIRM project (currently named as "ELEI" (EURHISFIRM Legal Entity Identifier) as the potential identification system to be used by the project's common data model.

Much of European historical company-level data does not yet exist in findable, accessible, interoperable, or reusable data formats. EURHISFIRM envisions designing an RI to make this possible.

EURHISFIRM itself will enable the creation of a data management plan on historical company-level data. As such, the design study is in itself a data management plan, and therefore some questions remain naturally unanswered at this stage. This Data Management Plan document will therefore continue evolving in its subsequent version updates.

1. Data Summary

1.1 Purpose of the data collection/generation and data utility

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

A few large stand-alone databases have been built by both the academic community and by private companies, but that has been done without any concern for interoperability. Within academia, considerable resources have been devoted to the construction of historical datasets, as often as not with limited aims, to study specific issues. Moreover, such datasets are scattered and dispersed and do not satisfy the FAIR data principles (Findable, Accessible, Interoperable and Re-usable): they lack any systematic comparative or diachronic analytical purpose[3]. The Strategy Report on Research Infrastructure identifies Big Data in the social sciences and the humanities as the first science driver for these fields.[4] FAIR data change the way for carrying out academic research. In spite of the crucial advances "born-digital" big data can bring, they still lack the historical depth that "born-on-paper" data can provide. European cultural heritage represents a shared wealth in terms of citizenship, cultural growth, and economic potential. Hence, the Strategy Report identifies the emerging need and opportunity for research infrastructures (RIs) providing access to this heritage and innovative technologies to analyse and integrate extracted information to large stakeholders' communities. The EURHISFIRM "Historical high-quality company-level

[1] Directorate-General for Economic and Financial Affairs of the European Commission, 2009

[2] European Commission, n.d. *Europe in a changing world - Inclusive, innovative and reflective societies*

[3] Wilkinson, et al., 2016

[4] Juncker, Tusk, Dijsselbloem, Draghi, & Schulz, 2015

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

data for Europe" project addresses this need with a comprehensive study of: investigating the historical sources available, designing the correct standardisation methods, as well as creating the optimal infrastructures and technology. The design study will be compliant to the FAIR (findable, accessible, interoperable, reusable) data principles.

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

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

As an infrastructure, EURHISFIRM designs a world-class RI to connect, collect, collate, align, and share detailed, reliable, and standardised long-run company-level data for Europe. The goal is then to provide an infrastructure to extract and enrich data as well as to connect, collate, and align existing and new data. The project concerns two main data formats: digitised (stored in databases) and raw (not yet digitised and not yet transformed into databases).

To make the work manageable, the EURHISFIRM design study first focuses on stock exchange-listed companies because they are larger and better documented. Concerning the inventory of the data and sources from all countries in the consortium (Belgium, Germany, France, the Netherlands, Poland, Spain, and the United Kingdom), Work Package (WP) 4 has prepared a detailed inventory, containing more than 250 sources (see Deliverable D4.2[5]). 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 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[6]). WP4 also produced accurate data and sources documentation according to the chosen common documentation standard (DDI Lifecycle (https://www.ddialliance.org/), see Deliverable D4.1[7]) 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 (Deliverable D4.4[8]). 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

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

[6] Poukens, EURHISFIRM D4.3: Report on the semantics of data and sources, 2019

[7] Poukens, EURHISFIRM D4.1: Information system and documentation standards, 2018

[8] Poukens, EURHISFIRM D4.4: Report on data and sources documentation and quality assessment, 2019

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

challenge of ensuring standardised approaches whilst allowing for idiosyncrasies of the diverse data types from various countries across time. To do so, Deliverable 4.5[9] gives a detailed report of the EURHISFIRM standards for the documentation of data, that is description of the provenance, characteristics, structure, and contents of datasets and printed sources.

Regarding the origin, types, and size of the existing data, the size of the SCOB (**S**tudie**c**entrum voor **O**nderneming en **B**eurs)[10] database (based on Oracle) of the University of Antwerp is 29 gigabytes. The DFIH (**D**onnées **FI**nancières **H**istoriques)[11] database (on Oracle) of the Paris School of Economics is 60 gigabytes, with the total size of the image scans at 10,5 terabytes. The SCOB database is a digitised collection of historical data of the Brussels Stock Exchange from 1832; the DFIH database provides this for the Paris Stock Exchange (1795-1976).

Regarding these new sources from other countries in the consortium, the estimated size will become more defined as the project progresses.

1.3 Re-usage of existing data

A few large stand-alone long-term databases have been built by both the academic community (e.g. the London Share Price Database) and by private companies (e.g. the Global Financial Data database), but interoperability remains low. (It is worthwhile to note the exceptions of the SCOB database at the University of Antwerp and the Data for Financial History Database at the Paris School of Economics, which have been built in a coordinated way (both institutions belong to the EURHISFIRM consortium).) Within academia, considerable resources have been devoted to the construction of historical datasets, as often as not with limited aims, to study specific issues. Moreover, such datasets are scattered and dispersed and do not satisfy the FAIR data principles (Findable, Accessible, Interoperable and Re-usable). The main goal of EURHISFIRM is twofold: 1) designing the infrastructure to be used by academics and other stakeholders to deposit and connect their data, as well as to 2) inspire new projects of data collection with the next-generation data extraction and enrichment platform developed from EURHISFIRM. Accordingly, in combination with the study of the existing data models (SCOB and DFIH), national data standards and semantics are developed and harmonised in the process towards a common European data format within WP5. Technologies to match historical high-quality data and to merge them with or link them to data stored in other historical and contemporary databases are developed by WP6. A platform based on OCR and AI to extract and enrich the data is designed within WP7 by using these existing sources, as well as those from other countries in the consortium whose data have not yet been processed.

Consortium members such as Carlos III University and Goethe University, who also have already run extensive data collections, are committed to reverse and integrate their data into the infrastructure design (which, as mentioned in the previous paragraph, will work with the SCOB and DFIH databases in the first phases of development) to create a first pool of data big enough to raise interest within the "data collectors" community. This gravitational pull will attract already existing data to make them re-usable

[9] Poukens, EURHISFIRM D4.5: Report on EURHISFIRM documentation standard, 2019

[10] Annaert & Buelens, 2017

[11] Hautcoeur & Riva, 2018

within the infrastructure once they are documented and structured according to the established (meta)data format.

In future phases of the project, EURHISFIRM envisions building a community-based infrastructure which will be able to integrate existing economic history data from other European countries and institutions outside of the consortium.

1.4 Data utility

A whole Work Package (WP8) is devoted to develop and run a large-scale survey, via an online questionnaire and interviews, in order to ascertain the preferences of potential users and key stakeholders for the design of data and services that EURHISFIRM RI should provide. Based on answers of more than 120 potential users of the EURHISFIRM platform (see Deliverable D8.2[12]), and on qualitative interviews with potential users and stakeholders (see Deliverable D8.3[13]), priority should be given to data relating to the twentieth century and concerning the UK, Germany and France, ordinary equity market data (and if possible daily price data), accounting data (i.e. total assets, total debt, revenues, and profits).

Potential users seem to be willing to manipulate the data themselves directly on the EURHISFIRM platform. Specifically, users will wish to be able to download the data in bulk (e.g. in MS Excel or csv format), and with minimum restrictions on downloads per period. Deliverable D8.4 also recommends allowing users to 'click through' to a scan of the original document for reassurance as to the accuracy of data. It also recommends that EURHISFIRM provide an explanation of the methodology and rationale for any interpretation or manipulation of data carried out by EURHISFIRM researchers.

2. FAIR data

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

2.1 Making data findable

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

In order to render the data findable for future users, the EURHISFIRM design study must select the appropriate metadata format. A number of standards have been under study by WP4 and WP5, and the DDI family of standards (http://www.ddialliance.org) has been chosen as the optimal method for the type

[12] Adams, Campbell, Coyle and Turner, EURHISFIRM D8.2: Report on surveys results, 2019

[13] Adams, Campbell, Coyle and Turner, EURHISFIRM D8.3: Report on interviews results, 2019

of data EURHISFIRM envisions (i.e. microdata), especially in dealing with various elements of the data that may change in format and content over time and over different phases in the research lifecycle[14]. Unique and persistent identifiers will be assigned to datasets stored within the infrastructure and the underlying metadata records will be updated in case of several versions of the concerned dataset. Search by k-words will be provided. EURHISFIRM envisages using existing controlled vocabularies such as the Thesaurus for Economics (STW, https://zbw.eu/stw) for subject terms and the Getty Thesaurus of Geographical Names (TGN) for geographical keywords as much as possible the enhance findability. To expand the RI's compatibility with contemporary financial data, EURHISFIRM is testing a way to facilitate the process of migrating the data from their current relational databases to the open-data Wikibase. The Financial Industry Business Ontology (FIBO) (created by the EDM Council (https://edmcouncil.org/)), to be possibly used as the grammar standard within this process, is also currently under investigation.

In summary, the following metadata standards, controlled vocabularies and "grammars" will be used or are in evaluation to code the data's properties (naming conventions) to adhere to FAIR principles: DDI Lifecycle and DDI Codebook, STW and TGN (metadata compatible with printed serial sources and academic and historical databases), FIBO (under evaluation; compatible with contemporary financial data standards), and Wiki (to promote open and collaborative data usage). Incidentally, Wikibase also allows the creation of version numbers. WP5 and the WGIS (Working Group on Identification and Standardisation, an inter-WP work group dedicated to the standardisation of the common data model) are also studying the Legal Entity Identifier standards (developed by the Global LEI Foundation (GLEIF))[15] as the potential identification system to be used by the common data. The group is studying on adapting the LEI for EURHISFIRM's purposes in order to fit the project's historical contexts and nuances.

2.2 Making data openly accessible

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

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

The nature of the data does not require a data access committee. However, to ensure ethical use of the data, as well as to comply with the General Data Protection Regulation (GDPR)[16], the design will enforce that any data that may reveal personally identifiable data (i.e. any data that would allow for the

[14] Poukens, EURHISFIRM D4.5: Report on EURHISFIRM documentation standard, 2019

[15] Global Legal Entity Identifier Foundation, 2019

[16] European Commission, 2018

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

identification of living persons) will be properly handled. WP3 will verify the compliance of EURHSFIRM policy with these regulations. The final design could, therefore, mandate that such data would be anonymised or even be made inaccessible.

In the final infrastructure design, the current plan will include a tool that will allow users to customise their data exports from the infrastructure. This tool will include the relevant documentation to facilitate user access.

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

Another type of data generated by EURHISFIRM are survey data (WP8) to collect information on target users' historical financial data needs. This information helps us to design the RI that would well serve the interests of the European scientific, public sector, and private sector communities.

WP6 matches data in historical databases to data in other historical or contemporary databases for the purpose of merging (in case of partner databases) or linking (in case of external databases) data on the same entities (companies, securities or persons). Its output will consists of matching algorithms and linking tables. The matching algorithms automatically match entities based on similarity measures of certain characteristics (e.g. name, date of incorporation, price, dividend). The matches proposed by the algorithms are then checked by a human expert and the confirmed matches are stored in a linking table. These linking tables contain the respective identifiers of entities that are common to two or more datasets. For the design study, the matching algorithms were implemented and performed in Python, an open-source programming language. Once complete, the matching and linking software will be made available through an appropriate code repository (e.g. GitHub). Future updates of the code, also during the implementation and operational phases of EURHISFIRM, will be made available in the same manner. In the EURHISFIRM RI, linking tables will be stored as part of the identification system (see next section) that is being developed by WP5 and will be made available as such. Different scenarios for matching and connecting data from consortium members' databases (i.e. SCOB and DFIH) to external databases (i.e. Eurofidai and LSPD), from a basic exchange of identifiers to a complete merger, are being discussed. Access to the data on linked entities stored in external databases will be contingent on the result of these ongoing discussions and on the licence and access modalities of these databases. Furthermore, efforts are being made to develop a collaborative environment based on the Wikibase format, in which the data will not only be open, but the environment would also allow a much wider field of researchers to contribute to the task of data matching, as well as to the improvement in data quality. Additionally, algorithms (bots) could also be used to discover potential matches in this data format.

Additionally, the EURHISFIRM project's public outputs (including deliverables, reports, milestones) are openly available through the CC-BY licence and accessible via the project website (https://eurhisfirm.eu/) and the OpenAIRE platform (https://explore.openaire.eu/search/project?projectId=corda__h2020::612830f55f1f92d36a5477538163d4e5) via the Zenodo depository (https://zenodo.org/). The deliverables on Zenodo also include metadata and persistent identifiers (DOI).

2.3 Making data interoperable

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

The reasons for the low interoperability are:

- For both digitised and non-digitised data: different languages, types of markets, formats, normalisation of changes over time e.g. company name evolutions and/or M&As
- Digitised data: stored in multiple databases in various data formats, which increases the incongruence of the data and therefore makes analysis and cross-comparison difficult
- Non-digitised data: in printed format. As they are not digitised, searching, analysis and cross-comparisons are extremely cumbersome

In the EURHISFIRM project, WP5 is working on a common data model (CDM) to overcome these interoperability challenges in historical European company-level data. The current model developed by WP5 so far describes a system in which the local data sources (pink layer) will be treated through data integration gateways (yellow layer) and then integrated within a common access system through which data users can consume the data (green layer). The details of these processes are currently under discussion within the WP5 tasks. WP5 plans to advance in the CDM's interoperability by continuing its investigation in the best practice commonalities of the existing data models of EURHISFIRM members (SCOB and DFIH), as well its study of potential overarching systems (Legal Entity Identifier (developed by the Global LEI Foundation (GLEIF): https://www.gleif.org/). As mentioned in Section 2.1, the WGIS and WP5 are working on enriching the LEI for EURHISFIRM's purposes. Currently, this is called "ELEI" and it will enhance the LEI's ability to accommodate to historical data's nuances and characteristics. In addition, WP5 will also investigate external models such as the EUROFIDAI (Institut Européen des données financières: https://www.eurofidai.org/), London Share Price Database (London Business School) and the CRSP (The Center for Research in Security Prices: http://www.crsp.com/) to optimise the EURHISFIRM CDM's interoperability within its own sources, as well as potential future external models.

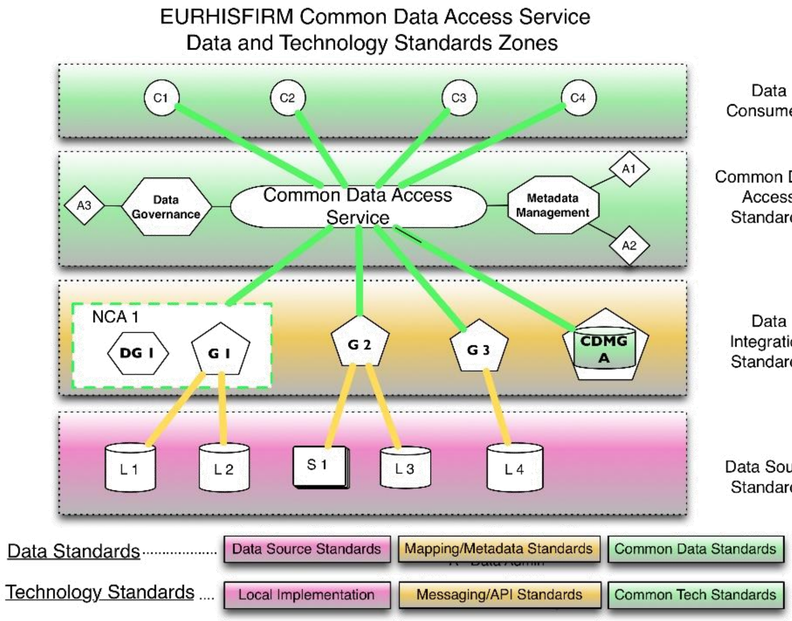

Figure 1: proposed Common Data Access Service (provided by W König (GUF, WP5))

The work from D5.1 and D5.2, completed in 2019 and 2020, further expanded upon this concept of a federated model with the “principle of least intrusiveness”, i.e. retaining as much of each data source’s original characteristics and independence as possible while ensuring their interoperability with one another in the common data model.[17] The preliminary design for this system is described in D5.2.

As mentioned above, WP6 will also study 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. Eventually, based on the models resulting from this design study, the ideal goal would be to increase the interoperability of historical financial data with data on as many other European countries as possible. The semantic analysis performed by WP4 revealed a lot of heterogeneity and national idiosyncrasies in long-term data from a diverse set of countries, however. Harmonisation of data is therefore part and parcel of interoperability. Harmonisation first involves mapping data elements (e.g. columns in a spreadsheet) from various datasets and databases to each other and to a common ontology or data model (schema matching). In preparation for this work, the IT team at PSE is currently evaluating the data vocabulary called Financial Industry Business Ontology (FIBO) created by the EDM Council (https://edmcouncil.org/) in order to standardise the existing Belgian and French data from their Oracle platforms, which would enable their migration to the open-data Wikibase platform. Upon its completion, the EURHISFIRM common data model will serve as complement or substitute to FIBO for historical company data. Next, the harmonisation process will involve deduplication of double entities (e.g. cross-listed securities) based on the matching algorithms and linking tables described in the previous section. Additionally, the DDI family of standards has been adopted as the metadata standard, as previously mentioned. The wide acceptance

[17] EURHISFIRM Consortium, 2020

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

of DDI in the social sciences will increase the RI's compatibility with academic and historical databases. The articulation between FIBO and DDI Lifecycle will be also evaluated.

WP7 works with Data Source Standards (the pink/lowest level in Figure 1) in order to explore the most suitable methods for transforming the sources into exportable outputs. This process takes into consideration the interoperability of the different potential sources and how these would interact with each other in the Data Integration Standards and the Common Data Access Standards levels in Figure 1. Currently, WP7 is working with the DFIH database to export its data using XML formats. The results from this study will enable WP7 to determine the optimal formats for repeating the process for data from other databases that could have different characteristics than those of DFIH. The result of this work will allow the import and export of heterogeneous sources at the Data Sources Standards level for interoperable capabilities at the Data Integration Standards and the Common Data Access Standards levels.

2.4 Increase data re-use (through clarifying licences)

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

WP8 has also completed a survey (observing compliance with GDPR) with interested stakeholders in academia, business, and policy in order to understand the users' needs and interests concerning long-term financial company-level data (see D8.2)[18]. As these survey respondents come from diverse fields, understanding their data needs will allow the RI design to consider its long-term utility across domains, ensuring the data's open access and reusability. WP9 will also study and recommend infrastructure policy and architecture that would optimally support the sustainability, and therefore the reusability, of the data. Additionally, WP11 will also explore on the ways that digitised historical financial data can be used to promote and deepen European research, culture, and heritage.

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

3. Allocation of resources

Regarding financial resources, the costs for conforming to the FAIR principles established within the EURHISFIRM project will depend on the final setting of the RI and the business model selected, which will be completed within the framework of WP10. The business model (for long-term funding of the infrastructure) is currently being designed by WP10 and will be decided by the EURHISFIRM governance. So far, WP10 has examined existing social science research infrastructures for this purpose. It has also completed D10.1 as a preliminary exploration of business models and governance structures that could be

[18] Adams, Campbell, Coyle, & Turner, 2019

applicable for EURHISFIRM. The final designs will be selected with consideration of stakeholders' preferences.

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

WP11 will also study the methods of using and promoting the cultural value of the infrastructure. The working hypothesis on this policy will be established by WP9.

4. Data security

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

These policies will be studied and detailed in future versions of this document.

5. Ethical aspects

There are two main ethical issues that arise in the EURHISFIRM project's scope: 1) ownership/intellectual property rights and 2) compliance to the General Data Protection Regulation (GDPR)[19].

Following the passage of the GDPR, data providers must comply with the ethical handling and storage of data which may be personally identifiable. This policy applies to EURHISFIRM in the following scope:

- How (storage [will it be stored in any networks or computers linked to networks], how will it be accessed by the scientific team/how is it shared)
- Why/objective (including is it for commercial use)
- Future use of the data (will the data be reused into other surveys and/or other research)
- Users' rights (how to delete it, access it, who to contact for questions)

WP3 has started work on these topics regarding EURHISFIRM's data sources. The studies so far indicate that for EURHISFIRM, there is generally little possibility on infringement of data privacy rights, since EURHISFIRM concerns data on persons who are not living (and these protections apply normally upon natural and living persons). However, the WP will aim to identify specific cases in which the law may still

[19] European Commission, 2018

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

be applicable. The basis for these rules will be the EU laws.[20] The report D3.2: Ethics, which will be completed by the end of December 2020, will provide further details on the ethical issues.

Additionally, WP8 (interaction with users), observed GDPR compliance in its work, as mentioned above in section 2.4.

Note on this data management plan

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

[20] EURHISFIRM consortium, 2020

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

References

Adams, R., Campbell, G., Coyle, C., & Turner, J. (2019). *EURHISFIRM D8.2: Report on surveys results.* Queen's University Belfast, Belfast.

Adams, R., Campbell, G., Coyle, C., & Turner, J. (2019). *EURHISFIRM D8.3: Report on interviews results.* Queen's University of Belfast, Belfast.

Annaert, J., & Buelens, F. (2017). (University of Antwerp) Retrieved from SCOB (Studiecentrum voor Onderneming en Beurs): http://www.scob.be/

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

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

EDM Council. (2018). *FIBOpedia*. Retrieved from https://spec.edmcouncil.org/fibo/fibopedia/master/2018Q4/FIBOpedia.html#

EURHISFIRM Consortium. (2020). *D1.10: Second yearly progress and strategy report to the General Assembly.* Various.

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

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

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

Global Legal Entity Identifier Foundation. (2019). *Introducing the Legal Entity Identifier (LEI)*. Retrieved from https://www.gleif.org/en/about-lei/introducing-the-legal-entity-identifier-lei

Hautcoeur, P.-C., & Riva, A. (2018). (Paris School of Economics) Retrieved from Data for Financial History (DFIH) Database: https://dfih.fr/

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

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

Poukens, J. (2018). *EURHISFIRM D4.2: Report on the Inventory of Data and Sources.* University of Antwerp, Antwerp.

Poukens, J. (2019). *EURHISFIRM D4.3: Report on the semantics of data and sources.* University of Antwerp, Antwerp.

Poukens, J. (2019). *EURHISFIRM D4.4: Report on data and sources documentation and quality assessment.* University of Antwerp, Antwerp.

Poukens, J. (2019). *EURHISFIRM D4.5: Report on EURHISFIRM documentation standard.* University of Antwerp, Antwerp.

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