Long-term data for Europe

EURHISFIRM

D5.1: Technical document on national data models

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

Long-term data for Europe

D5.1: Technical document on national data models

AUTHOR:

Pantelis Karapanagiotis[1]

ABSTRACT:

The report reviews a selection of existing micro-level data-model implementations both from within as well as outside the consortium's countries and identifies best design practices. It proposes preliminary model concepts for *EURHISFIRM*'s metadata scheme and evaluation criteria for assessing the effectiveness of historical, cross-country, company-level data models. Since there is no precedence in designing such models, the report methodologically introduces a conceptual 2-dimensional separation on the information space that *EURHISFIRM*'s model aims to cover and reviews representative implementations from each subpart. The first dimension concerns the time domain. In this dimension, the reviewed models are classified either as contemporary or as historical. The second dimension concerns the cross-country domain. Models here are classified either as national or as international. The analysis constitutes one fundamental block upon which the process of synthesizing national models into a unified European common model builds.

APPROVED IN 2019 BY:

Jan Annaert (Universiteit Antwerpen)

Wolfgang König (Goethe Universität Frankfurt)

Angelo Riva (École d'Économie de Paris)

Goethe Universität Frankfurt , Faculty of Economics and Business Administration,
Theodor-W.-Adorno-Platz 3, 60323 Frankfurt am Main, Germany.
Phone: +49 69 798 33673; +49 69 798 30081
pantelis.karapanagiotis@hof.uni-frankfurt.de

[1] I would like to thank Jan Annaert, Jefferson Braswell, Wolfgang König, Lukas Manuel Ranft, Lut de Moor and Uwe Walz for their valuable suggestions on many of the topics discussed in this report. In addition I would like to thank Stephanie Collet for her support while composing this report. I would like to express my gratitude to Frans Buelens for his detailed guidance through the mechanics of the Belgian data model. Furthermore, I am thankful to both Emmanuel Raviart and David Smadja for their insightful explanations of the extensions of the French model and to Jérémy Ducros for disambiguating fine historical details on its informational context. Last but not least I want to thank the WGIS participants and especially Johan Poukens, Coen Fierst van Wijnandsbergen and Lana Yoo for the inspiring discussions that we had during our meetings.

Table of Contents

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

I. Introduction

Historical firm-level information in Europe is scattered among printed and electronic sources documented in different data formats. Even among electronic sources, the underlying data models are heterogeneous. This situation prohibits potential stakeholders from utilizing valuable sources of information for research and industry-oriented purposes. The overall objective of Work Package 5 (henceforth WP5) is to provide a Common Data Model (henceforth CDM) design that overcomes this obstacle and allows integrated and harmonized access of European historical firm-level data from heterogeneous sources[2].

The current report aims, through reviewing a representative selection of existing micro-level data model implementations, at identifying best design practices and developing preliminary model concepts for the metadata scheme. Initially, the report reviews the most advanced implementations of the consortium. Further, it goes beyond the consortium's implementations and reviews micro-level model implementations outside of the consortium. The review of existing implementations accommodates the design of additional models. It accommodates, in particular, the design of the CDM and of national models in countries of the consortium for which advancements in this direction are still needed.

Methodologically, the report introduces a conceptual 2-dimensional separation on the information space that *EURHISFIRM*'s model aims to cover and reviews representative implementations from each subpart. The first dimension concerns the time domain. In this dimension, models are classified either as contemporary or as historical. Models from both within and outside the consortium are reviewed. From within the consortium, the report focuses on the models of "Studiecentrum voor Onderneming en Beurs" (henceforth SCOB) and of "Data for Financial History" (henceforth DFIH). From outside the consortium, the focus lies with the model of "Center for Research in Security Price" (henceforth CRSP) and its merge with Compustat. The second dimension concerns the cross-country domain. Models here are classified either as national or as international. In this dimension, the model of "European Financial Data Institute" (henceforth EUROFIDAI) which focuses on contemporary, European financial data is reviewed.

The reviewing procedure involved the national communities by interviewing qualified experts[3] and identified national idiosyncrasies as well as overarching design challenges. The documentation of these points in this report develops multidisciplinary insight with respect to possible issues that might arise during the harmonization process.

Last but not least, this report constitutes one fundamental block upon which the process of synthesizing national models into a unified European CDM builds. In this direction, the report proposes a set of points for the methodical evaluation of national data model implementations based on the design principles, needs and goals of *EURHISFIRM.*

[2]. See (Riva et al., 2017, p. 35).

[3]. Frans Buelens guided me through the Belgian data model. David Smadja, Emmanuel Raviart, and Jérémy Ducros introduced me to the French model.

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

The remaining report is organized in a natural way. Section II is the base of the report; it reviews representative models from both within and outside the consortium. It also reviews a European data model with contemporary financial data. Section III discusses special issues with respect to European, historical firm-level data that are important when designing a common data model and introduces the initial metadata concepts. Section IV builds on the review and identifies best practices and points at which there is room for improvement. Finally, section V proposes a set of criteria that constitute the basis of an evaluation framework that brings together the insight of the analysis with the strategic goals of *EURHISFIRM.* The last section concludes the report.

II. Review of existing company-level data models

National model implementations vary among the countries of the consortium. The longest-lived implementation among *EURHISFIRM*-participating countries lies in Antwerp. It is a relational implementation, based on a data model of Belgian firms, with more than twenty years of development. In 2011, the development of a derivative, French-firm based implementation begun in Paris. Since this fork, the two models were developed in cooperation, but independently and occasionally followed divergent paths. In the rest of the countries of the consortium, the implementation of national, firm-level data models is not as advanced as in Belgium and France. In particular, for most countries, there is no implementation of a national model.

The report focuses on data models with existing implementations. Specifically, it focuses on implementations that have advancements in terms of firm/security identification, data collation, input-output interfaces, documentation and harmonization transformations. Many institutions of the consortium, for instance, “Sustainable Architecture for Finance in Europe” (henceforth SAFE) in Germany, “Carlos III University in Madrid” (henceforth UC3M) in Spain, and “Queen’s University of Belfast” (henceforth QUB) in UK have extensive collections of micro-data but do not yet have finished implementations. Data collections in spreadsheets are omitted in favor of focusing on advanced modeling elements. Details on the availability of data and the semantics of the sources in these countries can be found in (Poukens, 2019).

Outside the countries of the consortium, the situation is similar. On some occasions, institutions in countries like Portugal, Sweden and Austria have achieved progress in terms of historical data collection and model implementation. The scope of these projects is mostly focused on stock-market data at a national level. A prominent, advanced national implementation of stock-market data is “London Business School’s Share Price Database” (henceforth LSPD). The LSPD has a security-based model with records of UK stock returns starting from 1955[4]. In most cases, the implementation of national, company-level models has not yet started.

Micro-level, company models with European-wide, historical data have not yet been implemented. This is exactly the gap that *EURHISFIRM* aims to close. However, there are implementations of models with financial data that concern contemporary information. Collating and linking data from different European

[4] See (LSPD, 2019) for an overview of LSPD’s content.

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

countries is indicative of some of the challenges that *EURHISFIRM* needs to overcome in the cross-country domain. For this reason, the EUROFIDAI, which is a prominent and established contemporary European security-level implementation, is also employed as a resource for identifying successfully applied practices that may fit *EURHISFIRM*. LSPD offers an alternative choice of an established security-level implementation. In contrast to EUROFIDAI, LSPD is a national model and cannot be used to identify cross-country modeling elements.

Linking data across European countries is only a part of the problem that *EURHISFIRM* tackles. The historical dimension of *EURHISFIRM* constitutes a separate challenge. The legal, accounting and financial systems of European countries were never in the past as harmonized as today. In this direction, *EURHISFIRM*'s model needs to achieve unprecedented innovations in order to provide unified access of information to its end-users.

Contrary to the European situation, in the US there is an established, widely used model implementation of historical, financial data. CRSP has a long and successful record of supplying academics and practitioners with historical and financial information. Furthermore, its merge with Compustat extends the informational content of CRSP to a company level.

Historical information across US states has been documented in a more harmonized manner compared to Europe. The documentation was done in a single language and the financial, accounting and corporate governance regulations have been, to some extent, centrally regulated by the federal government. As a result, the historical cross-country element that *EURHISFIRM*'s design needs to model is more complicated than that of the US case. Furthermore, the time domain that *EURHISFIRM* targets is greater than that of CRSP. Nevertheless, the CRSP/Compustat model is relevant for identifying best practices for *EURHISFIRM*'s design. There are commonalities between modeling the US and European historical data as in both cases, one needs to deal with evolving legal, financial and accounting systems.

Another aspect of the CRSP/Compustat model that is relevant for *EURHISFIRM* is the merge itself. Already within the consortium, there are two distinct model implementations. Furthermore, the possible future inclusion of countries like Sweden and Portugal in *EURHISFIRM*'s consortium would introduce additional model implementations that the CDM should have to take into account. One important aspect of the common model is the process of linking and collating information across national implementations. For this reason, the study of the CRSP/Compustat merge can be informative of the difficulties that the CDM needs to overcome.

Table 1. Conceptual dimensions and representative models

	National	**European**
Historical	SCOB, DFIH, CRSP/Compustat	*EURHISFIRM*
Contemporary	CRSP/Compustat	EUROFIDAI

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

The list of for reviewed modes outside of the consortium is not exhaustive. The review neither claims to be nor attempts to provide a complete catalog of financial database implementations. The set of reviewed models in this section is representative and it was selected so that the CDM design benefits the most from past experiences. Although a comprehensive, historical, European company-level model has not been implemented in the past, models with national historical, national contemporary and international contemporary data have been already implemented. By focusing on the time domain, *EURHISFIRM*'s design identifies best practices from SCOB, DFIH, and CRSP/Compustat models. By focusing on the cross-country domain, *EURHISFIRM* identifies best practices from the EUROFIDAI model. Refinitiv and Bloomberg constitute alternative choices of implementations that contain contemporary, cross-country elements. The infrastructures of these alternatives are mainly targeting users in financial practice and, therefore, the focus is laid on EUROFIDAI, which is a research-oriented infrastructure. The methodical 2x2 conceptual distinction is presented in Table 1. It aims to account for and present the majority of challenges and modeling issues that are relevant for a comprehensive, historical, European design.

i. The SCOB model

SCOB is a research center founded at the University of Antwerp in 1999 as a joint project of the economics, history, and computer science departments. SCOB obtained archives of the Brussels Stock Exchange and Antwerp Stock Exchange and partially digitized them. It also obtained limited archives from the Liege Stock Exchange and the Ghent Stock Exchange, but have not yet digitized them.

Main Concepts

The Antwerp model is spanned across three main, plus two supplementary, information sets. On the three main sets, a relational model of historical, firm-level financial information is spanned. These information sets coincide with the information sets of *EURHISFIRM* research infrastructure[5], namely financial information, accounting information, and corporate governance. One supplementary information set concerns macro-level data that is needed either in various transformation processes or is commonly used in conjunction with micro-level data for research purposes. The other supplementary information set concerns geospatial data. Diagram 1 gives a high-level, top overview of SCOB's model.

The starting point of the Antwerp model was to design and implement a model with accurate, historical stock data. This concept is the basis of the CRSP model and is also the main topic of the EurHiStock research group. During the initial design process, it became apparent that a model in which stocks are the sole modeling concept is an inadequate description of historical, financial information. Stocks are issued by legal entities and are owned and traded among physical (i.e. persons) and legal entities (firms and other organizations). Stocks issued by a firm can be quoted in different stock exchanges at different prices. Firm characteristics and accounting information are highly relevant in the context of accurately depicting historical financial information. Thus a concept of wider scope than that of the stock is needed in order for one to be able to tackle in a comprehensive manner research questions of economic and historical interest.

[5]. See (Riva et al., 2017, p.4).

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

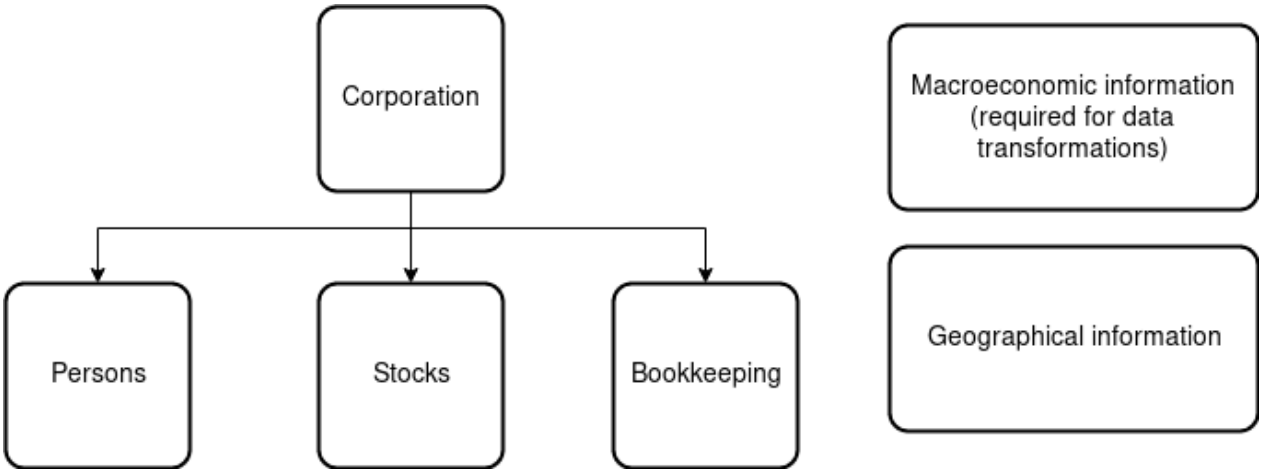


Diagram 1. SCOB model's basic concepts

The central concept of Antwerp's implementation is the `corporation`. In the context of the model, a corporation is not to be understood in the strict sense of a privately owned company, but rather includes also legal entities of official nature such as local and national government. In this respect, the Antwerp implementation closely relates to the main object of interest of the common *EURHISFIRM* model. The corporation concept is related to the `person` concept in multiple ways. For instance, persons hold offices and roles in governments, they own firms and hold managing positions in organizations. Corporations are also linked with a `stock` concept. In the model's context, a stock represents a wider variety of financial instruments than only stocks. For instance, it includes government-issued securities like bonds. The last main concept with which corporations are linked is that of `bookkeeping`. The bookkeeping concept refers to information related to financial reports and statements of corporations.

When relevant, for instance in the case of privately owned firms, `geospatial information` about the headquarters, as well as about where a corporation is acting, is stored. In particular, as far as it concerns firms, the Antwerp implementation contains information for both listed and non-listed firms. Furthermore, some auxiliary `macroeconomic data` like inflation and currency exchange rates are also stored in the relational database.

The stock concept is related to `stock exchange` information. The corporation-stock exchange association is a one-to-many relationship. Each corporation can be linked to multiple stocks that are traded in different stock exchanges, even within one country. Fields for opening, closing, maximum and minimum daily prices, stock dividends as well as the respective currencies of the stocks are included in the model design. Besides prices and quantities, events related to the lifecycle of a security (e.g. emission dates, stock splits, etc.) are recorded. Corporations issue stocks in different markets, even within one country. Associations of different stocks with each other can then be obtained through linking with the corresponding issuing corporation. Stocks are traded in different currencies and more importantly, the currency of dividend payments can differ from the trading currency. The model addresses this complication by allowing the recording of the currencies for both the trading of the security and the dividends.

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

There are various bookkeeping information objects that the model uses to record the assets and liabilities of firms. One of the difficulties in any model that targets to describe firms from a historical perspective is that accounting systems are not time-invariant. In the SCOB model, this obstacle is overcome by associating the historical accounts with the current Belgian accounting system. Both the original and the associated contemporary accounts are stored. At an original information layer, accounting data are documented using the accounting system that is used at each date in the printed archives. At a derived information layer, every account of the historical accounting system is translated into the corresponding account of the current accounting system. This, while preserving the original accounting documentation, allows data users to access accounting information in a unified and familiar way.

The accounting information enters the model at two different aggregation levels. Initially, new accounting records are entered at an aggregated level. Then separate accounting bookkeeping items are entered. Aggregation is then computationally performed within the system, and the results are contrasted with the manually entered aggregated values for validation purposes.

The person concept of the SCOB model is used to give a relational representation of the corporation's managerial and ownership structure. Personal information of owners and managers are captured from multiple sources. For instance, the model allows for the title of the person and her address of residence to be stored. In general, there are legal bounds concerning privacy with respect to the extent that this information can be used, but the actual data used in the SCOB database originate from publicly available sources.

The person's relationship with a corporation is represented in the model by a `person function` concept. As mentioned earlier, the notion of the corporation of the SCOB model also includes states. As a result, many of the functions of persons are political. In order to record also political memberships of persons, the SCOB model also documents historical information on political parties, governments, and coalitions. Such information is central in corporate finance and corporate governance economic research.

On another direction, the SCOB model associates persons with each other based on their social relationships. For instance, it allows for describing matrimonial or maternal relationships between persons. The relationship types that the model is able to capture is extendable. For instance, the current version of the implementation also includes relationships such as 'king collaborators' and 'University classmates'.

The macroeconomic information object contains supplementary information to the model. This information is used in conjunction with the information contained described from the main concepts to produce derived data. Derived data are new data that result from transformations that are commonly used for research purposes. Macro data stored in the current version of the implementation include information on population, GDP, inflation, exchange rates, etc.

The corporation identification is achieved with the use of a composite key. The composite key consists of an integer part, which references the entity's name, augmented with a date-record, which captures the date of the latest corporate action that altered the legal standing of the corporation. The integer reference part of the key is used in place of a firm's name in order to allow for firms to have the same name, even if

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

they are for instance founded on the same date. The integer part of the key is not created based on name changes of a corporation, but rather on headquarter and activity location changes. In the context of the model, the headquarter location is referred to as `corporation location`, and activity locations are label as `social locations`. The records for both of these concepts are time-dependent.

Every corporation is affiliated with a single industry at a given date. The industry information follows the classification of firms in industries as found in the financial information archives. The model implementation allows one to easily add new sectors into the model. However, multi-purpose firms, which is a common case today as well as a case that frequently occurs in historical data, are not supported. Associating a firm with a single industry in some cases fails to accurately describe the historical evolution of companies that operated synchronously in multiple sectors.

`Corporate actions` are also included in the model design. Such corporate actions can lead to the introduction of new corporations (e.g. mergers of firms) or removal of corporations (e.g. defaults), but can also be of historical relevance without altering the corporation entity (e.g. capital adjustments, political events, buybacks of shares, etc.). Stock splits, emissions, and other `corporates events` that affect the stocks issued by the corporation receive special treatment and are handled separately from the rest of corporate actions.

The records of the stock concept are categorized using a two-level scheme. The first level records the general security type (e.g. stock, future), while the second level records more specific information (e.g. common stock, preferred stock, founder stock, etc.). The second layer categorization is based on specific contract attributes of the security. For instance, stocks with and without voting rights are represented by different contracts and, hence, are recorded with different second-level type entries.

System Design

Instead of following the mainstream research approach of collecting information in simple data files, the SCOB implementation pursues a long-term, research infrastructure solution. The collected information is organized in a table-based model and stored into a relational database. The source of information is printed archives, which are digitized and inserted into the database manually. An overview of the system design is contained in Diagram 2.

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

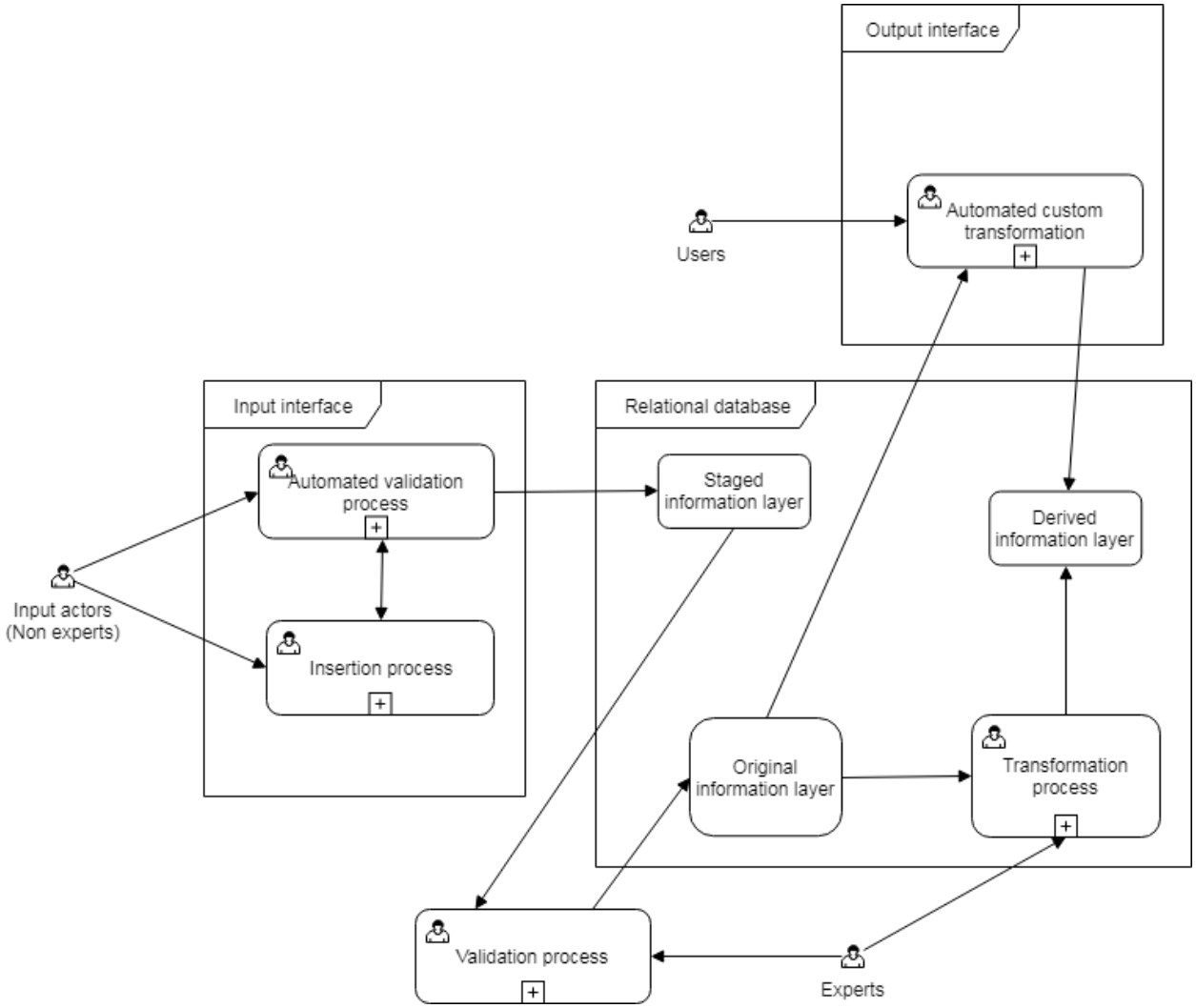


Diagram 2. SCOB system overview

The input process is performed by usually non-expert input actors like student assistants through the use of Graphical User Interfaces (henceforth GUI). The use of the GUIs offers at least two advantages in comparison to direct input to the database through code. Firstly, it limits the database's exposure to input actors only to the parts that are relevant for a particular insertion. Secondly, it introduces a variety of automated validity checks for typos and errors.

The GUIs are the front-end parts of database-driven applications developed by various contributors to the SCOB database. These applications create database connections in the background and provide a safer interface for input operations. In the earliest implementations of the system, data input was performed directly from the database client interface. Such an approach exposed the system to input actors directly. It required from users a better technical understanding of database technologies, and it was more error-prone. With the introduction of input applications, the system exposure during input operations was reduced. In the current version of the system implementation input actors have to use text- and combo-boxes instead of Structured Query Language (abbreviated SQL) code. The applications automatically

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

compose user input into SQL and perform various types of error checking before staging any changes into the system.

The Antwerp implementation uses two input interfaces. The first interface is used to input financial data and the second to input corporation level data. The first interface allows users to input security data, as well as information on prices and traded quantities. The corporation level interface allows input actors to store information on a firm's official location (headquarters), effective location (where the firm actually operates), administration (board of directors), stockholders, accounting data (liabilities and assets), flows (costs and benefits), portfolios (security positioning of a firm) and banking information (intermediation and loans). These fields are not mandatory as organizations with legal structures other than that of privately owned traded firms are stored.

The new data that are created through the input applications are not directly saved into the main database tables, but instead, they remain into auxiliary tables in a staging phase. New information is committed to the main database tables only after it is thoroughly checked from experts. The insertion process already reveals a fundamental principle of the SCOB design; that is, only information of validated quality is allowed to be inserted into the main model data.

On the output side, one can obtain information either directly from the original, historical data stored in the database, or from the transformed data. The transformed information uses the original data on accounting, dividends, and prices and adjusts them based on macroeconomic conditions and corporate events to provide harmonized access. For instance, original dividends might be in foreign currencies and transformed dividends adjust them to the national currency by dividing them with the contemporary to the payment date exchange rate. Another example concerns the calculation of returns. If one calculates returns directly from the original data, the resulting values will not accurately reflect the actual returns of the underlying securities in cases in which stock splits occurred during the used dates. For this reason, transformed data on returns take into account corporate actions that affect securities.

There are many output applications with functionality targeting specific data selection cases. As an example, there exists an application that facilitates index construction. It offers the user the ability to construct personalized indices based on a nested series of choices. The user can choose the type of securities and the exchange market in which they are traded. Having already specified the last two options, the user further specifies the sub-type of the securities she is interested in. Further, she chooses the types of the firms she is interested in (e.g. firms with headquarters in Belgium, or colonial firms), and select specific firms from this pool. There is an option to select also firms based on where they operate. Then the user chooses the dates of interest and lastly the weighting scheme (e.g. market capitalization based weighting, equal weighting, etc.). The output process shows that access automatization is among the fundamental design principles of SCOB.

Another design principle that is ubiquitous in the SCOB model is the storage of a copy of the original text from which the information is extracted. As the printed archives are digitized, the lines containing information that is relevant to the model are stored alongside the extracted information. This both benefits to some extent the verification of information accuracy, as anyone can always cross-validate the relationally stored data with a copy of the text from which they were elicited, and offers the opportunity

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

of future data transformations that use as starting point the original, digitized source of information. One should note that the verification of information accuracy is only partial when one compares stored copied information with relational data, as there might also be errors in the original data layer. As shown in Diagram 3, which depicts the validation process, if one wants to check the data against the original information, she has to have access to the printed archive.

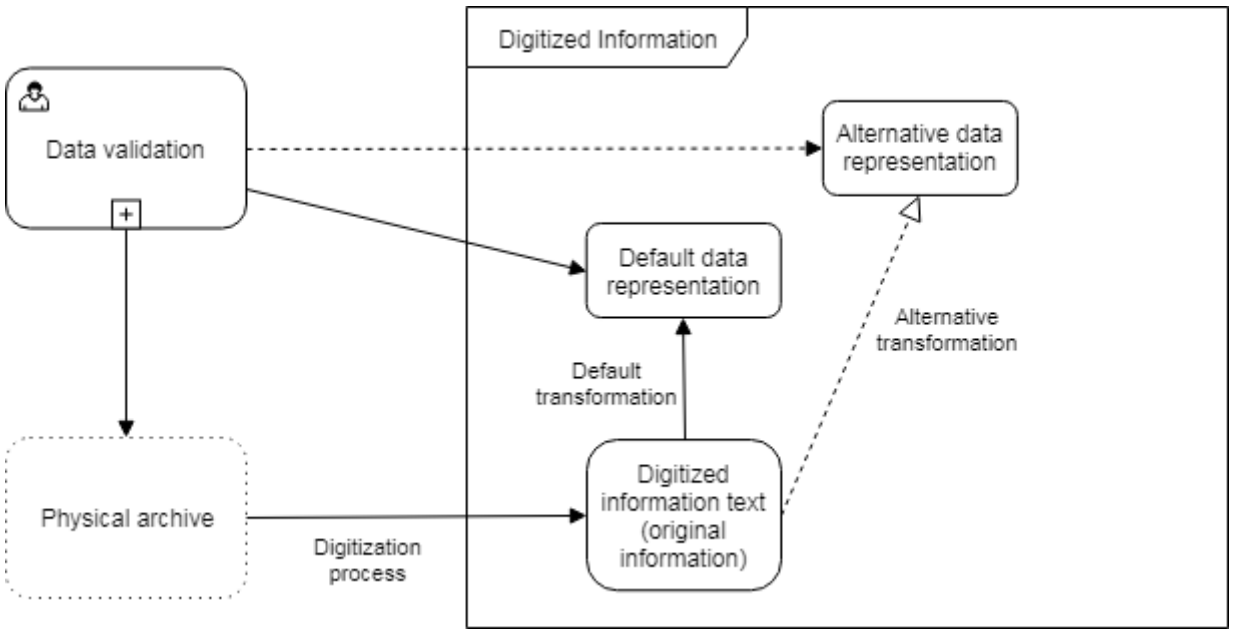


Diagram 3. Validation against source in SCOB

From a technical point of view, these three principles lead the design to conceptually adopt a three storage layer system. The first one can be labeled as the `staged information layer`. It consists of temporary tables with data that can be potentially part of the main database in the future, as long as they satisfy specific requirements. The second layer can be called `original information layer` and contains relationally linked information as formatted and documented in the source archive. Despite that original information is indispensable for the long term health and data accuracy of the system, it does not always constitute a fertile environment with respect to research activities in economics and history. For instance, the original sources do not contain returns of securities. This issue is addressed by adding a third, `derived information layer`, which contains commonly used research data transformations.

The design of the model supports the construction of multiple stock indices. Users can in principle design and construct their own index using data from the original information layer. Indices that are commonly used are stored in the derived information layer of the databases. An illustration of how flexible is the construction of indices in the SCOB model is given in (Annaert, Buelens, & De Ceuster, 2012). In this publication, new monthly return series for Belgian owned equity using stock market data from Brussels for the period 1832-1914 are constructed. Indices are constructed using three different weighting schemes, namely using relative market capitalization weights, equal weights, and price weights.

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

Content access and documentation

At the moment, access to the database is restricted to people that are involved with the SCOB project. Remote access to the database is not granted. There is no official, public documentation. Instead, there is an internally used manual that documents parts of the database and the accompanying applications. Its contents only partially cover the system functionality and sometimes the supplied information is outdated. There is also no established procedure that describes how and under what conditions external researchers can access the data. Restricted access and limited documentation are two important factors that hinder the potential of the research infrastructure to establish itself among the research community.

ii. The DFIH model

DFIH is the research institution behind the development of the French model. It brings together people with expertise in history, economics, information technology, and data science. Through its operational partnership with SCOB, DFIH introduced the French model as a fork of the Antwerp implementation. As such, the DFIH implementation shares a lot of model and system design characteristics with SCOB's implementation. This section presents DFIH's model incrementally rather than repeating elements that are also found in SCOB, i.e. it focuses on the characteristics that are distinct in DFIH's implementation.

In its eight years of development, the French implementation introduced a series of innovations in the original SCOB design. Some innovations were driven by exploiting advancements in machine learning technologies, while others were introduced in order to accommodate the idiosyncratic needs of the French data. The adoption of artificial-intelligence-based technologies is an integral part of *EURHISFIRM*, closely related to the primary objective of Work Package 7[6]. The adjustment process to French idiosyncrasies is indicative of the difficulties and raised when an implementation designed for a particular information space is applied to another one. This is also relevant for the common data model, as its design might as well be directly adopted by countries that do not currently have national model implementations.

Content expansion and adjustments

The Paris implementation is also a relational model that maintains the triple information-layer structure of the SCOB model (staged, original and derived data) and at the same time expands it content-wise. For instance, it allows for a wider range of prices in spot markets. Intraday transaction prices are also stored when information is available in the printed sources. Moreover, the model introduces new security types, like (call) options and certificates and allows the storage of relevant extracted information from the historical price lists. It adds more materials, units, and taxes on top of those of the SCOB model. The DFIH model also calculates and introduces in the derived information layer market capitalization data. Last but not least, the model adds database bookkeeping fields to all the tables. This extension, in particular, facilitates maintenance by introducing versioning features into the model.

The DFIH model standardizes the listed sector information based on Eurostat's statistical classification of economic activities in the European Community (henceforth NACE). The model offers a proposed mapping between historically listed classifications and NACE, but also allows end-users to use and create different mappings between this classification and listed stocks. Such implementation targets both end-users that

[6] See (Adam et al., 2019).

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

are oriented toward ease of use and end-users that follow a bottom-up approach. The first group of users can choose the built-in categorization, while users from the second group can create customized categorizations that fit their needs. As in the SCOB's case, the model restricts the association of firms with a single sector.

There are also efforts to expand the model's content in other directions, which at the time of writing of this report are still work in progress. Such endeavors are relevant for the development of the CDM and are briefly summarized here. The DFIH model standardizes countries and city names and links cities and regions in France with geolocation coordinates. Although this extension works well with current maps, matching geolocations with historical maps poses a currently unsolved challenge. Historical maps do not always offer precise depictions of historical locations of cities, and it is sometimes infeasible to attribute exact coordinates to some locations, in particular, if these cities do not exist anymore.

When it comes to validation of information against original sources, current work in the Paris implementation aims towards adding to the model the ability to store scanned images of printed archives and linking them with the corresponding digitized text. As in the SCOB implementation, the digitized text of the original information is also stored. The digitized text is central to the development of an updatable model as it is more suitable for applying new transformations to the source information. The scanned images, however, are more appropriate when it comes to information validation. As long as printed-to-text digitization processes are more error-prone than scanning, scanned sources offer a better alternative for validating information against the original source.

Artificial intelligence-based input process

One significant modification in the DFIH system concerns the input process. DFIH adds to the manual input process of SCOB an Optical Character Recognition (henceforth OCR) driven input process. Diagram 4 shows how this modification alters DFIH's system overview. There is of course still human-based digitization of sources, which is usually outsourced. Besides the acquisition of information, the human-based input is used to train, cross-validate, and compare digitized input from OCR technologies.

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

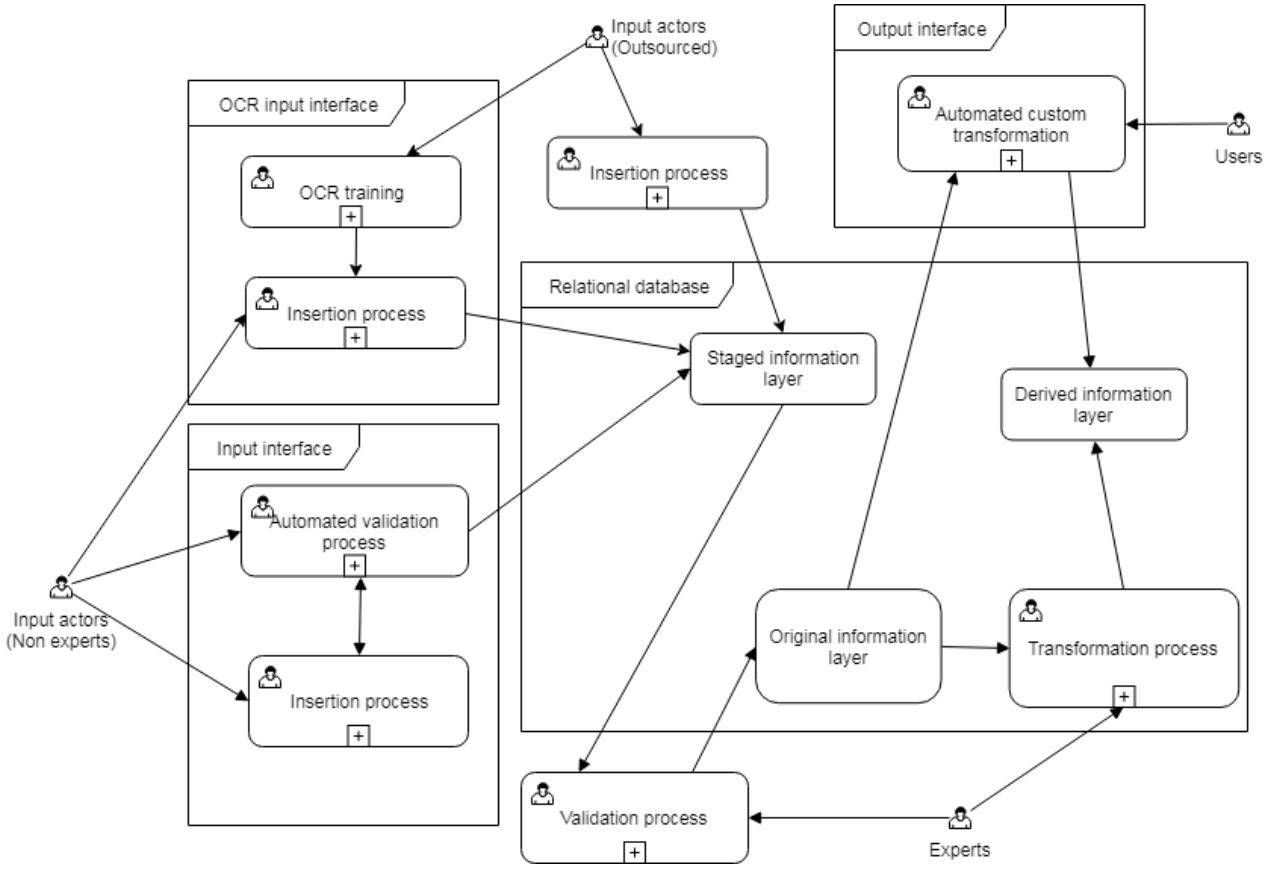


Diagram 4. DFIH system overview

The OCR input, as the manual input process, is GUI oriented. The GUI provides basic steering to the underlying OCR software, and its users can be individuals without expertise in economics and/or history. Users at this stage are responsible for guiding the OCR software to focus parsing the right parts of the document and to classify the parsed parts into the corresponding staging tables. Similar to the case of the SCOB model, the staged input information is validated by experts before being pushed into the other database layers.

The introduction of the OCR input process allows for the database to rapidly grow in terms of content. Conditional on the technological feasibility of this approach, the introduction of such technologies is a cost-effective, viable alternative to rapidly digitize historical information in other national implementations.

Moving into semantic technologies

Current work opens the DFIH implementation to a semantic-based redesign. Parts of the relational implementation are ported into semantic-based technologies, expanding the model in non-relational directions. The semantic part of the implementation is intended to work complementary to the relational part. The former part is more flexible in terms of browsing and navigational interactions with the users and supports the building of a broad user community around the project. The relational part is more robust, as it is already in development for many years.

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

One central characteristic of the adopted semantic approach is that it is based on web-oriented technologies. These technologies have mark-up input and output interfaces embedded in their design. This ameliorates the need to develop separate applications to handle input and output. Moreover, less training is required for the users of these interfaces as they can simply be accessed and operated through any web-browser.

The semantic re-design of the model requires the adoption of standardized vocabulary. The DFIH implementation references the Financial Industry Business Ontology (henceforth FIBO). The implementation does not aim to a full FIBO adoption. It instead focuses on borrowing financial terms from it and applies them to the corresponding DFIH model concepts.

Content access and documentation

Content access is in principle open also for people that are neither part of *EURHISFIRM* nor DFIH. The official website contains an online request data form. It also offers restricted but user-friendly, web-based, access to the data, and the corresponding scanned documents. The online software also offers the ability to produce standard graphs. Furthermore, there is also a proposed, standardized way to cite the DFIH infrastructure. Remote access is possible. The website offers help on the data items of the database, but the documentation is still under construction. All of the above points hint towards the orientation of the DFIH to establish itself as one of the leading historical research infrastructures in Europe.

iii. The EUROFIDAI model

EUROFIDAI is a research infrastructure funded by the French National Centre for Scientific Research (abbreviated CNRS). It aims to develop databases and provide academic researchers with financial data. The scope of EUROFIDAI is narrower than that of *EURHISFIRM.* It is primarily focused on financial instruments and not on firms. Nevertheless, the cross-country, European attributes of the model can be helpful for the CDM design.

Content

It provides both daily time series and high frequency, intraday data and covers stocks, indices, mutual funds, and exchange rates. The origin of the data is predominantly from European stock exchanges, but the implementation also provides financial information for markets in Asia and Oceania. The scope of the review in this section stays on the European part of the design.

The implementation currently provides daily stock data from 37 European countries. It covers both traditional stock exchanges as well as electronic markets. However, it contains only common stocks and not preferred stocks or bonds. For France in particular, time series start from 1977 and for the rest of the European countries three years later. The model calculates and provides a variety of market index data based on the stored stock data items. The indices are calculated at a European, national, and sector-level. Furthermore, risk-factor specific indices are provided. In addition, EUROFIDAI gives access to commonly calculated indices by other data providers and offers customized index composition on user demand.

Starting from 1980, the implementation provides spot exchange rate information at two levels. At a detailed level, the user has the ability to access exchange rate quotes from different suppliers and/or places. At a higher level, EUROFIDAI selects and provides a single currency pair as a rate. This two-level

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

approach enables researchers to customize the selection of information to their needs. For instance, researchers that are interested in competition issues in the exchange rate markets may access detailed information, while researchers that focus on international macroeconomic topics may as well choose to use the single exchange rate selection.

The current EUROFIDAI implementation provides mutual fund information both from Over the Counter (henceforth OTC) and organized markets. Besides the composition and returns, the service provides detailed background information on funds and fund managing characteristics.

The design also covers a variety of corporate actions that are relevant to stocks. These actions include name changes, capital structure changes, splits and reverse splits, assimilations, company reorganizations, listings, delisting and suspensions, dividend payment and others. The corporate events are classified using a two-level nested system. According to (EUROFIDAI, 2015), corporate events are classified into 28 classes that are then grouped into 13 general classes.

The sector classification is fairly limited and focused mainly on the stock exchange perspective. The design supports classified stocks into ten sectors that roughly describe the activities of the issuing firm. The sector classification is EUROFIDAI specific and does not adopt or use any standard.

The EUROFIDAI implementation calculates derived factor data. The size, value and momentum factors, which are commonly used in finance after their introduction by (Fama & French, 1993) and (Carhart, 1997), are calculated on an annual basis and provided out of the box. This is indicative of EUROFIDAI's orientation towards supporting innovative research in finance.

Identification design

Significant effort has been put in unifying the identification of securities. The global identification scheme uses EURIFIDAI codes. The EUROFIDAI codes are uniquely assigned to a combination of instruments and trading places. They enable tracking of an instrument throughout its trading history. In this way, the model is able to track securities, regardless of whether there were organizational changes or even changes in their International Securities Identification Numbers (henceforth ISIN). In addition, the identification method provides information on the securities' trading places.

A EUROFIDAI code is a 15 digit literal with a polymorphic interpretation that depends on the first number. The first number signifies the type of the instrument, i.e. if it is a stock, an exchange rate, etc. The following two digits define further the nature of the instrument. For all instrument types except exchange rates, the next seven digits constitute a unique instrument identifier. For exchange rates, the fourth digit is always zero, and the next six are codes for the base and the counter currency of the rate. The last five digits contain information on the data provider for exchange rates and on stock exchanges, currencies and geographical areas for all other instruments. The EUROFIDAI interface allows users to match identifiers with ISINs.

Services

EUROFIDAI's primary service is data provision. This service is organized into two parts. One part is readily available online. Subscribed researchers may directly access commonly used financial information on European stocks and other securities directly. The second part concerns mostly researchers that require

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

more specific information, such as subscription rights, forward exchange rates, warrants, and more. This information is not provided online, and users can obtain it only upon request.

EUROFIDAI's business model extends beyond plain data provision. The current implementation also offers technical support and scientific support services. Technical support ranges from mere assistance to download the data to specific tailor-made database construction and high performance, volume computing. The scientific support services concern data customizations such as calculation of non-standard indices, portfolios, and asset pricing factors upon user demand.

Documentation of data items is available. Moreover, the EUROFIDAI web site offers a lot of information with respect to the content of the databases. Information is not restricted only to a mere description of data items but also on detailed information on the construction of indices and asset pricing factors. The last form of documentation increases the usability of EUROFIDAI as it offers essential information to academics that need to know if the construction of a particular item is relevant for their research.

iv. The CRSP/Compustat model

The last model reviewed in this section is the outcome of the merge of the CRSP and Compustat models. CRSP provides only financial data, similar to EUROFIDAI. Compustat is a data vendor that supplies accounting firm-level data. The combination of CRSP and Compustat is the closest modeling example concerning the nature of historical information that *EURHISFIRM* aims to describe. However, drawing analogies between the CRSP/Compustat merge (henceforth CCM) and *EURHISFIRM* should be done with caution; US historical data is in many ways a more harmonized information space in comparison with European historical data.

CRSP is one of the leading research infrastructures for historical financial data worldwide. Merely the availability of historical financial data through CRSP for the US constitutes an essential factor that drives most of the contemporary empirical research in finance to be US-data oriented. When conducting such research, it is often informative to match security level data with issuing firm characteristics. For instance, the book-to-market factor requires both a stock-based and an accounting-based valuation of the firm. Matching databases that do not share a common identifier is complicated even when it is conducted by specialized experts. This is the business gap that the CCM product targets.

Identification design

The initial CRSP and Compustat databases have neither a common identifier for securities nor for firms. The merged database is a CRSP product that also contains Compustat data items, among which also the Compustat's identifiers. The resulting CCM data can be accessed by using either the original CRSP or the Compustat identifier.

CRSP uses `PERMNO` to identify stock issues. PERMNO is a five-digit, permanent identifier that is uniquely attributed to each security and unifies security access throughout its entire history. The identifier can be used to locate a particular security among different CRSP data files and irrespective of potential name and capital mutations. Similarly, CRSP identifies companies by using `PERMCO`. PERMCO is an integer that is uniquely associated with every firm that issues stocks covered by the CRSP implementation.

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

Compustat's permanent company identifier is called GVKEY. GVKEY is a six-digit number that is used to uniquely identify both public and private companies. Compustat's permanent issue identifier is called IID. Each GVKEY can have multiple IIDs and to properly identify a Compustat security both the GVKEY and the IID are needed.

Database merge

The CCM linking process was performed in two phases[7]. The first phase established an initial link between CRSP and Compustat's contemporary data. The merge of the two databases was based on linking the PERMNO and GVKEY identifiers. Since the former is a security identifier and the latter a company identifier, the resulting matches are characterized by a many to one relationship. The second phase extended the initial link to match historical company-level data. The security history was used as a guide to generate historical company linking.

The CCM implementation functions as an intermediate linking table rather than a complete database merge between CRSP and Compustat. It constitutes a link between the fundamental data of the two databases, e.g., identifiers. It does not contain all of CRSP's security or Compustat's accounting level data. The linking process involved both the expertise of CRSP's researchers and programmers. There were three types of challenges that were addressed during the linking process.

The first challenge concerned differences in the information spaces of the two databases. CRSP data concern only companies that are listed on the NYSE, NASDAQ, AMEX, and ARCA stock exchanges. Compustat's data cover all public and some private companies. Furthermore, in cases of mergers, one can find disagreements between CRSP and Compustat's surviving companies.

The second challenge concerned the time domain and frequency of the data. CRSP data goes back to 1925, while Compustat data goes only back to 1950. Moreover, there are time series discontinuities because of periods in which companies' stocks cease to trade. In addition, the frequency of records is different between CRSP and Compustat. CRSP's data are reported daily or monthly, while Compustat's data quarterly or annually.

The actual link between CRSP and Compustat entities may as well cease to exist during a fiscal year rather than at the end of it. The CCM link gives the end-users the ability to determine which occasions constitute a valid match for the time series range that they are interested in. A researcher may choose to include matches if the time series range is within the link range. She may also choose to include matches if the end of the time series data is in the link range. Alternatively, she can choose to include matches if any part of the time series data is within the link range.

The third challenge was related to differences in the modeling spaces of the two designs. CRSP is security-based, and Compustat is company-based. The primary entity identifiers for the two databases do not aim to identify the same entities. This is the reason that the resulting CCM link is a one-to-many relationship.

[7] See (CRSP, 2018a, p. 19).

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

For instance, a firm that is uniquely identified by a GVKEY in Compustat issues both shares with voting and without voting rights, which are identified as different securities in the CRSP database.

The established CCM links between the databases do not assume a unique form. Types of links with different linking characteristics are offered by the product. For instance, CCM gives its users the ability to choose if the established link is verified by CRSP researchers, or if it is only based on the comparisons of values of historical identification fields that are common in the two databases. In total CCM offers eight different link types, which reveals the orientation of the product in providing researchers with the flexibility to choose the data that fit better their research requirements.

Services

As CCM is a product of CRSP, it aims to cover a particular need in the scope of CRSP's wider business model. CRSP business model focuses on delivering specialized research products like CCM. Among the products that CRSP offers are historical stock prices and indices, US treasury date, and real estate data.

The content of the CRSP products can be accessed in multiple ways. There are two main content access types; namely, access for research and access for investment purposes. CRSP also provides data access tools. These include a Windows GUI called CRSPSift, text files of different formats, command-line tools. Furthermore, content can be accessed through low-level Application Programming Interfaces (henceforth API). A comprehensive guide on this form of access is provided in (CRSP, 2018b).

As far as it concerns research data access, researchers can access the content either directly by a subscription to CRSP, or indirectly by a subscription through a third-party partnership, as for instance through Wharton Research Data Services (abbreviated WRDS) or Computing in the Humanities and Social Sciences (abbreviated CHASS). The last two partnerships offer additional, web-based tools to facilitate access to CRSP's data.

When it comes to investment index data access users can access the CRSP's content through third-party data distributors. Many leading data distributors, as for instance Bloomberg and Thomson Reuters, have partnerships with CRSP.

CRSP documentation is one of its strongest points. The variables are documented in high detailed and the interested researcher is able to obtain information about the exact nature of the data item either directly by CRSP's documentation or indirectly by supporting documents of third-party partnerships. CRSP documentation does not cover only data item definitions and index construction but also includes topics on the methodology followed to create a particular product. For instance, the data guide of CCM (see (CRSP, 2018a)) describes the procedure through which the CRSP and Compustat databases were merged.

III. Special points and Metadata concepts

This section highlights points of special attention when it comes to the design of cross-country, historical data models, and introduces some initial CDM metadata concepts. Both the special points and the metadata concepts are crucial for an effective and comprehensive design of a common European model. In comparison with a design that captures only contemporary firm-level data, a comprehensive model of historical data is founded on concepts that are more flexible than how one understands a firm today. Had

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

the contemporary perspective of firm characteristics and functions been imposed on the CDM, it would have constituted an anachronism with adverse effects on the quality of the information provided by *EURHISFIRM*.

Most of the special points are distilled from the designs of the models that were reviewed in section II. Many of these points are already incorporated in some of these designs, but there are also points that come from missing features. For instance, some particular points concerning cross-country historical data are not relevant to the reviewed models. This section only logically organizes the points. The comparative overview of the solutions proposed from the reviewed models to these special issues is postponed until section IV, where best practices and possible extensions are discussed.

Besides identifying special design points, this section also identifies some of the most crucial elements for the CDM design. These elements are fundamental concepts that commonly appear in the reviewed models. The common elements are then used to introduce some preliminary metadata concepts for the CDM.

i. Special points

The special nature of the points summarized in this subsection comes either from the time dimension of the information space or from the cross-country dimension and in some cases from both. There are also points concerning the taxonomies of any model covering such an information space. Taxonomies, of course, are not special to this information space; models over any information space require some form of term standardization. The taxonomy points included in this subsection, however, are specific to modeling historical, financial, European firm-level information.

Cultural aspects

EURHISFIRM's sources of information are various historical printed documents. Besides the informational content that these documents contain, they constitute items of unique historical value. Among *EURHISFIRM's* goals is the promotion of European cultural heritage by collecting, digitizing, and providing electronic access to historical sources. This is an innovative design objective that characterizes *EURHISFIRM*, as most historical, financial data models neither contain, nor promote the cultural heritage aspect of the sources.

Digitization and promotion of historical sources are more advanced in arts and humanities. For instance, many museums also offer in a digitized format parts of their exhibitions. *EURHISFIRM* design can use experiences, support, and infrastructures that have already been developed in other fields to provide this functionality. The "Digital Research Infrastructure for the Arts and Humanities" (abbreviated DARIAH) offers services that *EURHISFIRM* can potentially utilize in promoting the cultural heritage aspect of its design.

The main objective of DARIAH is to promote and support digitally-based research and teaching in arts and humanities. The First Virtual Competency Centre of DARIAH works on providing e-infrastructures that allow sharing community-developed data. Among its goals is the provide services for large national archives and interoperability across locations and languages. Ongoing working groups in DARIAH target

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

the preservation of media arts and Urban heritage. *EURHISFIRM* can explore the possibility of cooperating with DARIAH concerning the preservation and dissemination of scanned archives.

Currency

Currency requires special treatment even when one restricts her attention in a particular country. The historical value of a currency is time-variant. Depicting accounting or financial valuations in long historical periods in nominal currency values reveals only partially the economic standing of the objects of interest. Value normalization, however, can be achieved through various methods and by using different price indices. The appropriateness of a method depends on the needs of particular research and is typically performed by economists. Furthermore, research might be focused on small historical time windows, for instance, to perform case studies or exploit natural experiments, for which the inflation effects can be less important. Researchers in such cases might be interested only in nominal values. Therefore, both nominal prices and values, as well as macroeconomic *price indices* are relevant for describing historical financial information.

In the cross-country dimension, values are documented in different currencies. Even for research topics that concern relative values, as for instance asset returns, currency differences can be important in long time frames as the inflation rates between countries might evolve significantly different. In addition, cross-country research topics, for which the center of interest is placed in variables described in absolute terms, require a way to transform values expressed in individual currencies to comparable ones. Thus, macroeconomic *exchange rate information* is essential for the design of a European model.

Accompanying exchange rate data are essential for commonly used data transformations even for national implementations. For instance, the construction of stock indices requires the correct incorporation of stock dividends. There are historical occasions that stock dividends are paid in a currency that is different from the currency in which the stock is traded. This generates the need for currency conversion. The actual exchange rate that it is used for a particular trade is subject to highly idiosyncratic microeconomic conditions, and it is generally unavailable. However, historical exchange rates at a macroeconomic level are generally available and can be used as a proxy for currency conversion to facilitate the construction of stock indices.

Evolution and characteristics of firms

What is a firm? Even from a contemporary, static perspective, there are multiple answers to this question. From a legal perspective, a firm is an association by which persons are united for business purposes. Within economics, there are alternative theoretical approaches to answering this question. The contractual perspective views a firm “as complex set of market contracts, only distinguished from ordinary spot market contracts by the continuity of association among input owners” (see (Foss, 1993, p 130)). This approach is supplementary to the legal definition. It is also compatible with *EURHISFIRM’s* approach, which aims to model information in the organizational and ownership spaces of a firm’s structure.

As an administrative entity, the firm might operate in various sectors at the same time. Moreover, the sectors are likely to change over time. For example, according to (Chandler, 1992, p 490) after the second Great War, machine companies became first movers in mainframe computers and over the counter drug companies became first movers in the new antibiotic prescription drugs.

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

As a legal entity, a firm signs contracts with its suppliers, distributors, employees, and often customers. Taxonomies that classify suppliers can be relevant for the CDM. Information that connects suppliers and beneficiaries with firms can be a valuable source of information for researchers. The objectives of the firm and other organizational attributes are also of importance. Not every firm is solely for-profit; some firms have non-profit objectives. Some firms have multiple objectives, among which there are also some with non-profit targets. Some firms consist of a single plant, shop or office, but others are multi-unit.

Identification of securities, persons, firms and other organizations

Historically, not much emphasis was given to the unique identification of companies. The growth of both the number of actors in the financial system and the complexity of the relations between them gradually raised the need for better identification of legal entities for supervision purposes. The "Global Legal Entity Identifier" (henceforth GLEIF) develops a legal entity identification system to address this need. *EURHISFIRM* needs to solve a similar identification problem in a historical context, and previous work on the identification of companies from GLEIF is potentially relevant for the CDM.

The situation is similar when it comes to the identification of historical, financial securities. Today the identification of securities is well established and advanced, but historical identification is more convoluted. ISINs were only introduced in the late 20th century. Before their introduction, security identification was made either by the ticker of the security or by a local identification number. Even in the latter case, the historical identification of securities through this number is not robust. There are cases in which the number assigned to a security was later reused for another security after the trade cessation of the one for which it was initially assigned.

The identification of economic entities other than companies and securities is also relevant for *EURHISFIRM*. Firstly, the identification of persons is relevant to depict the corporate governance components of a company. Secondly, the identification of organizations of official nature is also relevant as historical data concern also financial instruments that are issued by local and national governments. Not much work has been done towards identifying official organizations. However, the similarities of these organizations with the companies indicate that identification can be based on similar reference metadata. This is also the approach that is used by the SCOB and DFIH implementations. As far as it concerns person identification, the situation is much more complicated. This identification issue is yet to be resolved for contemporary information and *EURHISFIRM* is facing it in a historical context. The SCOB implementation has achieved some progress in this direction with the help of historians that specialize in the topic.

Language

In contrast to the US experience, designing a European financial data model requires special attention to the language of documentation. Both financial and accounting information is historically documented in different languages. Previous work on the semantics of the sources, see (Poukens, 2019), shows that many financial and accounting concepts have semantic analogs in various documenting languages. This becomes more and more prevalent as the dates move towards the end of the twentieth century. However, in particular for earlier dates, there is a subtle difficulty that is relevant only for the design of multiple-language, cross-country, historical models.

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

Linguistic term equivalence today does not necessarily imply historical concept equivalence. Differences in legal and accounting standards can potentially lead a term that linguistically corresponds to another term in a different language to semantically represent a different concept. For instance, the German term 'Vorstand' linguistically corresponds to the English term 'Board of directors'. In a corporate governance context, the two terms, however, have historically different charges. The German term, which is legislated already in the Prussian Joint-Stock Company Law of 1843, has more standardized meaning and is of more central importance than the English term, the legal meaning of which was not defined or provided at a state level but instead only clarified at an organizational level in the companies' articles of association.

The last observation signifies the importance of also storing linguistic information at a data or metadata level. For legal, economic, and accounting terms that specialists agree on the uniformity of the concepts they represent, variable unification and common storage of the underlying information is an effective approach. Terms for which there is disagreement among specialists about the equivalence of concepts can be potentially stored in inter-related but separate data structures and fields accompanied by the specific linguistics found in the sources.

Legal System

The description of the legal system requires special attention in both the historical and the cross-country directions. The legal regime for corporate governance, accounting, and financial instrumentation continuously evolved during the last two centuries. A review of the historical developments that are relevant for the countries of the consortium can be found in (Poukens, 2019). Besides this historical variation, there is also variation in the cross-country dimension. There are three influential legal families in Europe; the French civil law family, the German code family, and the UK common law family. This paragraph examines the implications of variations of the legal systems on the CDM design.

One crucial aspect of *EURHISFIRM* is that of the evolving, cross-country accounting regulations. Historically, different accounting systems were used, and the records were kept using accounts in different languages. Empirical research in economics that uses accounting data can only be performed if the historical accounts are appropriately transformed so that they are comparable. In contrast, it is more appropriate to have access to the original records for case studies and historical research.

In conclusion, research needs indicate that the model's design should be flexible enough to cover both original, historical information as well as provide harmonized data. Harmonization in *EURHISFIRM*'s context spans over two dimensions. In the time dimension, there are harmonization needs for within-country variations of regulations. In the cross-country dimension, there is a need to provide transformations with output that is comparable between different European legal systems.

Taxonomies

Standardization of terms and controlled vocabularies is not a topic that is special to the European historical, financial information space. Any model that is designed to be utilized by heterogeneous groups of people that were not involved in the implementation process has to use terms that are descriptive and indicative of the information that its data items contain. This paragraph gathers the taxonomies that are relevant when developing a model over *EURHISFIRM*'s information space.

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

On a top-level overview, *EURHISFIRM* needs to standardize, describe, and give examples of its fundamental modeling entities and relations. All the models reviewed in section II contain terms to describe entities of financial instruments. The CCM, SCOB, and DFIH models contain terms that describe financial statements and companies. Last but not least, SCOB and DFIH also provide terms for describing persons, governments, and corporate governance.

Drilling into the financial instrument concept, terms that classify security types are needed. In this direction, the reviewed models use different classifications. SCOB and DFIH use in some cases non-standard terms to describe security concepts. Recent work in the DFIH project borrows terminology from FIBO. Securities are typically classified in industries and sectors. For instance, EUROFIDAI classifies stocks into ten sectors. In contrast to EUROFIDAI however, *EURHISFIRM*'s primary design focus is the firm. Therefore it is more relevant for the industry and sector classifications to be performed at a firm level. Securities can then appropriately inherit the industry attributes of the issuing firms.

When it comes to financial statements, there are two different standardization levels. Each national implementation needs to standardize how harmonized accounting data are accessed. The common data model then standardizes how European harmonization of accounting data is implemented.

In terms of corporate governance, there are two areas in need of standardization. The corporate actions of vocabulary and the classification of organizational types. As *EURHISFIRM's* model extends not only to privately-owned companies but also to non-governmental organizations and governments the classification needs to be inclusive of these organizational types

For supplementary data, standardized vocabulary on fundamental macroeconomic terms and geographical areas can be useful.

ii. Metadata concepts

The standardization of concepts and terms is a major subject of the Working Group on Identification and Standards (henceforth WGIS). The WGIS is an organizational instrument that *EURHISFIRM* established on the Executive Committee meeting of the 25th of April, 2018. It is formed by the leaders of Work Packages 1, 4, 5, 6, and 7. The purpose of the group is to assist the standardization and identification topics of the common model. On the one hand, it brings together *EURHISFIRM* experts with different backgrounds and involves them in the standardization and identification design processes. On the other hand, it acts as a coordination device among different countries and promotes inter-package communication. Discussion on the specification of metadata concepts is ongoing.

This subsection identifies the fundamental modeling elements that are commonly found in the reviewed implementations and coincide with the information space that *EURHISFIRM* covers.

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

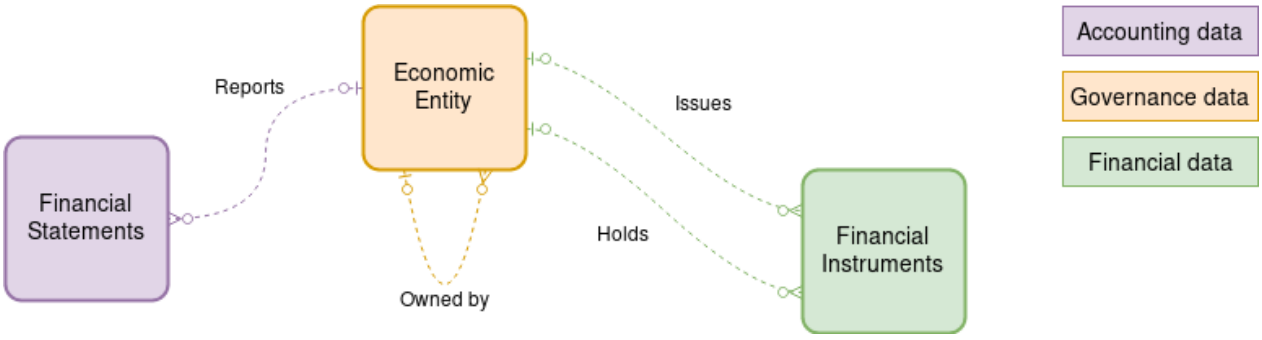


Diagram 5. Top overview of common model concepts

A top-level overview of the fundamental common model concepts is given in Diagram 5. The diagram contains three entity concepts and four relationship concepts. The top overview concerns the three basic informational sets that *EURHISFIRM* covers. This abstract overview does not contain geospatial and supplementary macroeconomic information. The former is related to all the entities in the overview and is left for more detailed model specifications. The latter is supplementary to the CDM, and it is excluded in favor of clarity of the presentation.

Entities

The three entities of Diagram 5 are the main classes of abstract objects that the model contains. Specializations of these classes represent concrete data objects of the model. Table 2 contains the definitions, examples, and some notes on the core classes of the overview.

Table 2. Core classes' definitions

Concept	Definition	Examples	Notes
Economic entity	A non-empty collection of societal units that bears economic activity in the form of transforming input to output or in the form of decision making that affects the economic activity of others.	Examples of economic entities are physical persons, companies of any organizational form, national, regional and local governments, and non-governmental organizations.	The economic entity concept encompasses both the concepts of organizations and persons that are found in SCOB and DFIH models. This global scope entity concept can be useful when describing corporate governance information. A company can be owned by physical persons, firms, and other organizations, as well as governments. The economic entity concept covers all cases.
Financial statements	Formal, typically written in a structured manner, records that convey the economic activities, financial	Examples of financial statements are balance sheets,	

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

	performance, and the economic standing of an economic entity.	cash flow statements, and income statements.	
Financial instruments	Contracts that specify monetary, financial, or commodity flows between *economic entities*.	Examples of financial instruments are money, bonds, stocks, forwards, futures and options.	The instrument concept corresponds to the stock concept found in SCOB and DFIH. In terms of content, both of these models contain already information on securities other than stocks, as for instance government bonds. The instrument concept aims to reflect this broader scope of available information.

Relationships

These concepts capture the interactions between the entities of the modeling space. Depending on the implementation technology used, these concepts can be either modeling constraints or also themselves stored data items. Table 3 contains the definitions, as well as some examples and notes of the relationship-concepts.

Table 3. Relationship's definitions

Relation	Definition	Examples	Notes
An economic entity reports zero or more financial statements	An association of a recorded economic standing with a particular economic entity as documented at the time of the recording.	For example, firm A reported a balance sheet statement at date YYYYMMDD referring to date YYYYMMDD, or proprietorship B reported a cash flow statement at date YYYYMMDD referring to date YYYYMMDD[8].	This is a general relation term that corresponds to the information context captured by the bookkeeping tables of SCOB and DFIH models. In the CCM merge, this relation concerns the informational content of Compustat. The Compustat model covers all the 10-K form company fillers.

[8] There are two dates that are relevant for recording the issuing of reports. The publication date of the report typically follows the date that the records refers to. Both dates are relevant for researchers. For instance, it is common for researchers in economic policy to focus on the dates that the information is announced instead of on the dates that the information refers to. Similarly, investors can only update their beliefs on the standing of a company when the reports are made public. Practically, the publication date is difficult to trace, however, the CDM should be able to deal with both. I want to thank Jan Annaert for pointing out this subtle, but important distinction between the two dates.

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

An economic entity is owned by zero or more economic entities	A description of the ownership structure of an *economic entity* at a particular date.	For example, a physical entity is not owned by anyone. Firm A is 51% owned by physical entity B and 49% owned by unknown entities that hold equity-based traded assets (e.g., stocks).	This relation describes the ownership structure relationships. It is general enough to allow for companies to be owned by physical persons, by other companies and organizations as well as states.
An economic entity issues and holds zero or more financial instruments	An association of a specific contract with two economic entities, namely the entity that offers the contract and the owning entity, at a particular date.	For instance, firm A issues bonds B, which are 60% held by physical entity C and 40% held by Firm D.	

IV. Best practices and possible extensions

Accounting data harmonization

As far as it concerns the harmonization of accounting systems, the CDM needs harmonization across time and across countries. The approach used in the SCOB model to harmonize in the time dimension can be extended to cover the cross-country dimension needs.

Besides storing the original historical accounts found in the printed sources, SCOB maps these accounts to contemporary Belgian accounts and also stores the resulting mapped accounts. The mapping is performed by historians that specialize in accounting systems. The CDM can, in turn, map the national accounts into harmonized ones. The harmonized accounts should provide users with uniform access to historical financial data. The CDM may propose its own financial statement standardization or adopt an existing one. The latter option is cost-effective as it requires less allocation of implementation resources. Moreover, it has the extra benefit of promoting *EURHISFIRM* accessibility. Users familiar with the adopted system are faced with less learning costs when they initially use *EURHISFIRM*'s data.

The International Financial Reporting Standards (henceforth IFRS) offers a single set of international accounting standards. The European Union has adopted the IFRS standards for all listed companies. Any company with securities that are traded in an EU stock exchange is required to prepare consolidated statements in accordance with the IFRS guidelines.

Documentation

Documentation is an integral part of the successful adoption of any software system. This is particularly relevant for software systems that provide convoluted services. Complicated functionality that exists in some software can be rendered inaccessible if it is not accompanied by sufficient documentation.

Both EUROFIDAI and CRSP offer extensive documentation. The offered documentation concerns holistically the system design and not only mere explanations of data objects. Online information and data

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

guides explain how to use user interfaces and API's. They provide detailed explanations of the information that the data fields describe. They also contain explanations of how links between different data sources were established and of how various data transformations were performed.

The models of the consortium fall behind at this point. This point is of particular importance when developing a common model that places academics and researchers in the center of attention. For *EURHISFIRM* to be successfully established as one of the leading research infrastructures for historical, financial firm-level data, resources in documenting linking procedures, data items and data transformations should be allocated in the implementation phase of the project.

Firm, instruments and market relations

This paragraph concerns three associations; namely, the relation between firms and financial instruments, the relation between instruments and stock exchanges, and the relation between firms and markets.

A given firm can issue zero or more financial instruments. This relation is natively captured by both the SCOB and DFIH models. EUROFIDAI is security-centric, and this relation is not included in the model. One of the significant challenges in the CDM design was to establish this link accurately. The end product also contains a one-to-many relationship between firms and securities (see subsection II.iv).

An instrument can be traded in one or more stock exchanges. This relation is not relevant only in a contemporary context, but it is also observed historically. Multiples stock exchanges were historically common even at an intra-national level. These markets can, but do not necessarily need to, be antagonistic. As described in (Hautcoeur & Riva, 2012) there were two complementary markets in Paris during the nineteenth century. The Parquet market was a mainstream security market organized by official brokers, while the Coulisse was an innovative forward market driven by independent brokers. Despite their specializations, the traded securities in these markets sometimes overlapped.

Both the Antwerp and Paris implementations allow storing information on firms' securities that are simultaneously traded in multiple markets. CRSP also covers the information space of the major US stock exchanges, and its design captures this one-to-many relationship.

The informational context regarding the stock exchange market can be further extended. Firstly, stock exchanges were historically either private or official entities. Similarly to firms, information on the managing directors and also the organizers (e.g. market makers, brokers, etc.) of the stock exchange can be included in the model. The legal structure in which the market was operating is also of research interest. For instance, the Parquet market had historically monopolistic rights for some classes of securities. Moreover, particularly legal issues and or legislation changes can be documented. Legislation changes like fairness-of-prices clauses, protection against broker defaults, price manipulation, and fraud laws potentially offer researchers environments to study quasi-natural experiments.

A firm can be active in one or more markets. This relation is the most difficult to describe as the market definition is context-dependent. Macroeconomics research might focus on the aggregate European market, or in aggregated national markets. Microeconomics research might as well focus on county-level markets or even at more restricted market definitions. For example antitrust and merger analysis in industrial organization research defines markets based on distances from coordinate locations.

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

The SCOB and DFIH models contain information on where a firm is active. In a historical context, this is particularly relevant for colonial firms as it is possible that a firm has its organizational jurisdiction in one country and its business activity in another one. Besides this, in the CDM space there is a need to describe firms that operate simultaneously in many European countries and cities.

Identification design

One difficulty in *EURHISFIRM*'s data collection process is the lack of properly identified information. Sources that contain firm information do not have a particular identification system. Interested readers can search for companies based on their names and other characteristics. Early sources that contain financial information also do not an identification system. More recent sources assign numerical codes to securities but the identification design remains inadequate. For instance, identifiers are not exclusive and the same codes are used for different securities.

Computer science helped to both theoretically and practically advance the design of identifiers. *EURHISFIRM* can adopt contemporary identification theory and apply best practices to avoid the pitfalls of historical identification designs. A proper identification system allows researchers to confidently refer to identified entities, not only within the scope of *EURHISFIRM*'s system but also when connecting information with external systems. This is relevant for the identification of entities in the long-run; proper identification design can protect future researchers from facing problems that are similar to those that are faced in *EURHISFIRM*'s data collection procedures.

The reviewed implementations adopt quite different identification approaches. For instance, and DFIH adopt a design that uses non-intelligent identifiers, i.e. strings that serve only as unique labels and reveal no more information. In contrast, EUROFIDAI adopts an identification design with intelligence, i.e. a system in which identifiers reveal meaningful information. Designs with unique non-intelligent identifiers and well-form metadata are used in many modern identification systems. This trend is mainly due to experiences from problems that arise when the embedded information in an intelligent identifier changes. For instance, GLEIF adopts a non-intelligent system for the identification of firms. The Digital Object Identifier (henceforth DOI) adopts also a non-intelligent system for the identification of intellectual property. More details on this specification can be found in (Paskin, 1999).

Industry taxonomies and its relation with firms

Industry classification is of central interest in economic research. Typically microeconomic research projects focus on particular industries and thus classification of firms in industries is relevant for the CDM. There are two alternatives through which the CDM can standardize industries. Either, as in EUROFIDAI, it can define its own classification, or it may adopt an industry classification like NACE.

The second approach is a better alternative for two reasons. Both arguments are parallel to those of the accounting standard adoption case. Firstly, the adoption of a standard with which many economists are already familiar reduces the learning costs from the user side and facilitates *EURHISFIRM* endorsement. Secondly, home-grown sector classification is accompanied by the necessity of allocating resources in this task. If not enough resources are allocated to the task, the CDM faces the risk of ending with limited sector classification as in EUROFIDAI.

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

Both national model implementations use a one to one relationship between firms and sectors. This, however, might be inadequate in describing not only multi-purpose firms but also the historical evolution of firms and industries. The CDM should include a one-to-many relationship between firms and sectors to capture both of these aspects.

Multi-layer storage classification

The design of SCOB and DFIH are implicitly using a three-layer storage system (see section II.i). The original information, the staged information, and the derived information layer. The staged information is typically stored in separate tables. Original and derived information is stored in mixed tables on some occasions. A model that makes more explicit this three-layer separation can be beneficial in terms of implementation. The layer separation attributes modular characteristics to the design, which makes the long term support of the system easier. For example, separating original and derived data minimizes the side effects to the derived part caused by any alterations originating from updates of the original information part of the system.

The multilayer storage classification can be standardized within the model by using storage prefixes. For instance, in a relational model implementation, tables that contain original information can be stored in one schema, and tables with derived information can be stored in another schema. Such a model design makes controlling user access rights easier. For instance, at the original information schema, user access can be further limited in comparison with the derived schema.

The three-layer classification can be further expanded by introducing additional (sub-) layers for the model objects that belong to specific informational (sub-) spaces. For instance, a sub-layer that contains only the scanned archives of the original layer can be introduced. Another relevant layer can be one that contains supplementary data, like macroeconomic information.

Open-source technologies

Both SCOB and DFIH databases are built using commercial software. Their implementations also use technology-specific functionality. This renders migration to alternative database systems difficult; a point which is relevant both when discussing the adoption of the current model by other interested institutions and when discussing the transition from national implementations to the CDM.

In principle, system migration is feasible. The model implementation does not depend on the adopted technology and can also be implemented in alternative open source technologies. There are not free conversion tools available. There are however many IT companies that offer conversion services. As a proof of concept, the author developed a regular expression based script that translates the proprietary based Data Definition Language (henceforth DDL) of the DFIH schema to DDL of an open-source relational database. The script is using an open-source license, but it reveals sensitive information about the schema of the Paris model, the disclosure of which could lead to security vulnerabilities. Therefore it is only confidentially distributed among *EURHISFIRM*'s members.

Recent developments in the DFIH model employ technologies that are open source. Wikibase is an open-source family of software tools and libraries for handling structured data. Wikidata is an open knowledge

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

base project built on top of Wikibase tools. Among other topics, current work focuses on mapping information from the relational implementation to the Wikidata implementation.

Original information preservation

A strong design principle of the Antwerp model is the digitization and conservation of the text from which the data is elicited. This promotes the adaptability of the model to future, unforeseen needs, as there is always the opportunity to introduce new data relations and/or transformation using the initially digitized text. Moreover, it enables users with sufficient access rights to validate the relationally stored data against the digitized text. As long as the stored digitized text precisely corresponds to the text found in the printed archives, this verification is equivalent to verifying the stored data against the original sources. Besides that, the digitization of archives is a cost-intensive procedure, and maintaining the original information reduces the need of effort duplication in the direction of re-digitizing the same archives.

Current work in the DFIH model attempts to take this principle one step further. Instead of storing only the digitized text, the model allows storing the scanned images of the sources. This enhances the validation capabilities of the system, as it gives the opportunity to end users to access a less error-prone digitized depiction of the original archives.

The principle of maintaining the original source of information has proved its value in practice through its adaptability in both French and Belgian models. Its value can also be illustrated in the context of *EURHISFIRM.* The transformation process of national data models into a common one requires constructions of data fields that are not useful or relevant for national models. For instance, harmonizing accounting information across various national models requires depicting accounting information in a system that is potentially different from all the originating, historical accounts. Since historical accounting information is stored in the original information layer of both Antwerp and Paris models, it can be directly used to bring the accounting data into the accounting format adopted by the CDM. Note that such a procedure does not need to access the accounting information in the derived information layer of Antwerp and Paris implementations. This procedure allows for the parallel storage of both national and international accounts. It also gives the ability to interested users to access accounting information not only in their European harmonization state but also in national formats.

The last example is not exhaustive. In abstract terms, the existence of an original information layer enhances a model's extendibility and adaptability. Harmonization of data and other data transformations can occur at any stage of the model's lifecycle without the need of revisiting the printed archives.

User interfaces

Both SCOB and DFIH have homegrown user interfaces for input and output. Most of the code of these interfaces is written in Java, hence they can be easily extended to cross-platform access tools. Furthermore, DFIH has already implemented some web-based user access components. As far as it concerns output, user interfaces have the advantage that they restrict the database model exposure to end-users. As far as it concerns input, they can also act as automated validation tools.

CRSP offers a variety of user interfaces either directly or through third-party partnerships. CRSP offers a Windows-specific GUI and command-line tools for cross-platform access. Partnerships with WRDS and

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

CHASS enable users to have access also using web-based tools. Last but not least, CRSP offers access through an API (see (CRSP, 2018b)). Systems designs that provide API access have two main advantages.

The first advantage is relevant to the business model of the underlying design. Providing an API enhances ease of access for specialized end-users. Specialized end users can either be end-users with programming skills, or programmers of third partner partnerships that build applications on top of the data products. Previous work of Work Package 8, see (Adams, Gareth, Christopher, & John, 2019), focused on academics as end-users. Data scientists and economists often develop software to conduct their research. For instance Excel, Stata, and R, which all have programming capabilities, were rated as the most favorable access tools. For researchers with programming skills, API access allows automatization and parametrization of content access. For third party programmers, API access allows for design and build modular derivative applications and products.

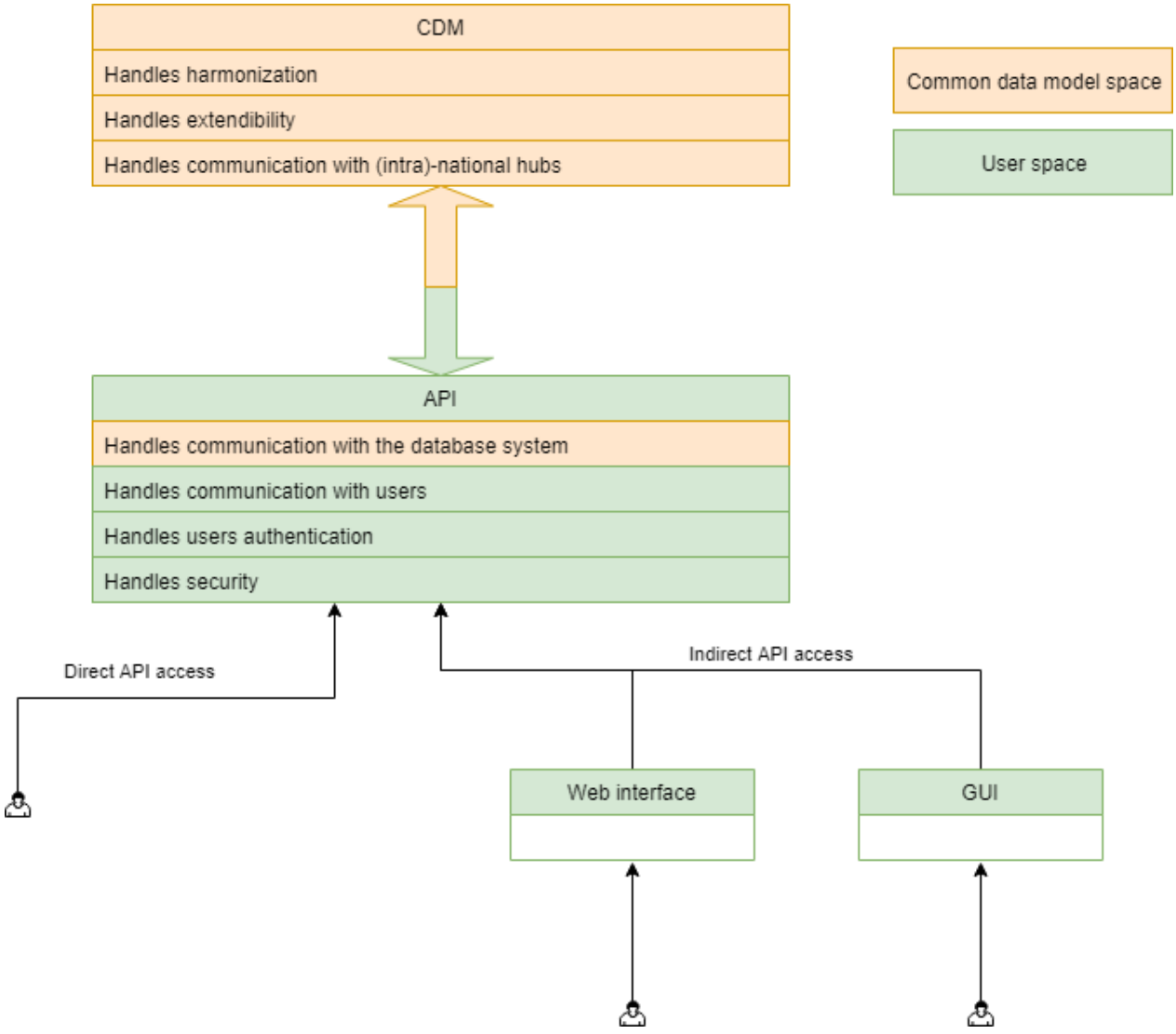


Diagram 6. An API driven encapsulated data access design

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

The second advantage of such a design is that APIs can also act as an isolation mechanism for the underlying database model. The consortium databases, as well as *EURHISFIRM*, can benefit by adopting such system design. A system design, as in Diagram 1, potentially allows for changes in the underlying database model without breaking the userspace. Every time that the database model is extended or updated, the functionality of the procedures in the API can be suitably adjusted to fit the model changes. The names of the procedures, i.e. the elements of the API that are in userspace, can remain invariant.

In principle, such a design allows the model to be extended without any need from the side of end-users with legacy access systems and codes to change their setups. In practice, this isolation mechanism works not always perfectly. CDM extensions that alter the underlying model significantly might force the deprecation of some API functionality and the introduction of new processes. Even in such cases, adopting a versioning system for the API allows both end-users and the system to check access compatibility.

Validation against source

The idea of providing the users the ability to check the validity of information against the source originates from the SCOB model. SCOB uses digitized textual data to represent the original source within the model space. Current work in the DFIH model uses also scanned images to represent the original source. This enhances the ability of the users to validate the implementation's information as digitization through scanning is less error-prone in comparison to manual digitization.

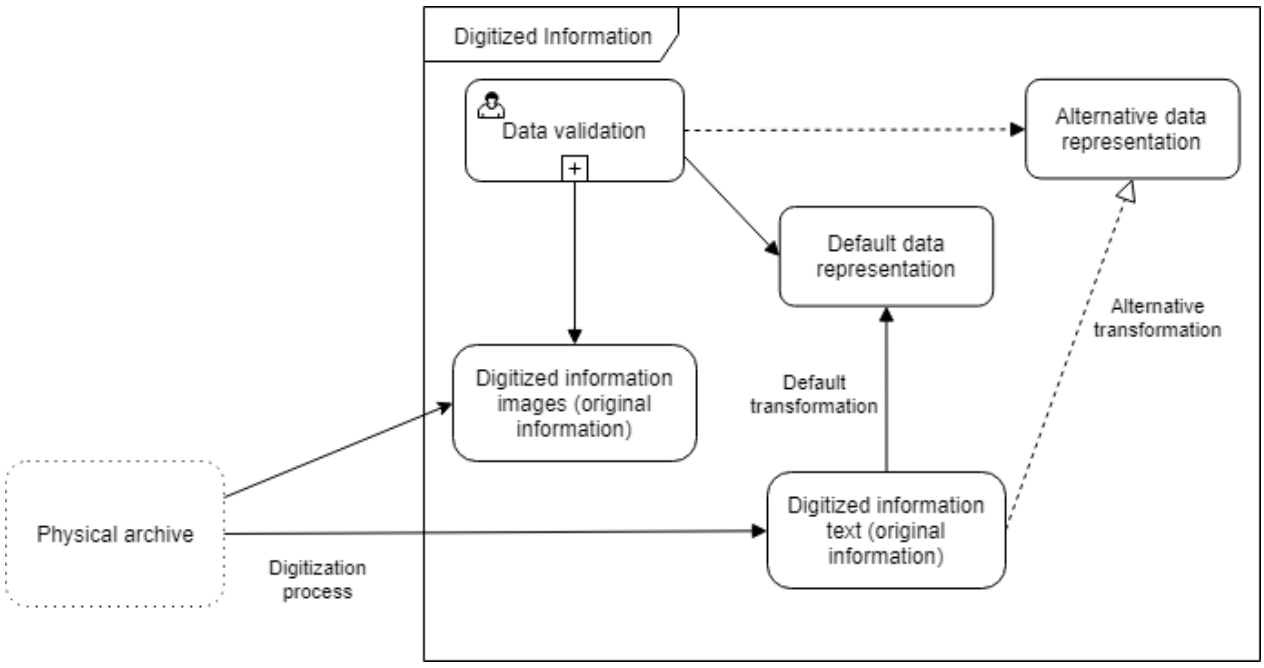


Diagram 7. Original information validation extension

The extension of including the scanned archives into the modeling space is depicted in Diagram 7. Users are able to more efficiently validate the accuracy of the information that they are using within the digitized information set. The need for resorting to the physical archives is further reduced with such a design.

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

This extension can be conditionally useful for *EURHISFIRM*. Among *EURHISFIRM*'s objectives is the promotion of the cultural heritage aspect of the historical printed sources. Their digitization and provision to the interested researchers as well to the general public undoubtedly benefit from the adoption of the extension of Diagram 7. However, any model implementation must be legally compliant. The extent to which, as well as the terms and conditions which can allow scanned sources to be provided by the CDM implementation is the subject on ongoing work of Work Package 3 (see (Riva et al., 2017, p 31)).

V. Evaluation methodology

This section proposes a set of evaluation criteria for firm-level model implementations with historical financial content. On the one hand, the criteria can be used to evaluate the national model implementations of the consortium's countries. They can also, of course, be used to evaluate model implementations outside of the consortium. On the other hand, the evaluation criteria can be used as fundamental principles for the CDM design.

The criteria are composed based on two building blocks. The first block is the findings of the previous sections of this report. Specifically, the review of the existing implementations led to identifying common patterns that have already been used from previous models. Then, the comparative overview of the implementations identified some of the best practices, the common concepts, and the weak points of the underlying models. The second block is *EURHISFIRM's* needs and design principles. Indeed, some criteria have no previous analog in the reviewing models, but instead, they are mandated by *EURHISFIRM*'s design.

The goal of this section is to set the directions across which future work on the CDM is to be evaluated. The topics are approached from a comprehensive perspective of *EURHISFIRM's* design rather than from a strict data model design. They are approached in this manner for two reasons. The first one is to highlight that the data model design is endogenous with respect to *EURHISFIRM's* design. Different design-goals lead to different data model designs. The second is to promote inter-work-package communication and exchange of perspectives. Future work of work-packages may find this report informative on the effects that their decisions may lay on the CDM and vice-versa. With this in mind, the evaluation topics are discussed, but solutions are only partially specified. Future work both from WP5 as well as from Work Packages 3 (Legal and ethical design), 6 (Data connection), 8 (Interaction with users), 9 (Infrastructure policy and architecture), 10 (Business model and governance) and 11 (Cultural heritage) have significant contributions in the development of the CDM.

Concerning, in particular, the processes and requirements discussed by Work Package 9, the evaluation criteria are mapped to the corresponding characteristics of (ISO, 2011) and (ISO, 2008). The mapping aims to act as a conductor between the CDM's attributes and infrastructure processes. For brevity, I will refer to the Quality in Use Model of (ISO, 2011) as QUM, to the Product Quality Model of (ISO, 2011) as PQM and to the Data Quality Model of (ISO, 2008) as DQM.

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

i. Ease of access and user-need orientation

Are the model's services adjusted to target user-needs?

This point concerns the user-space exposure of the CDM. In terms of the QUM, this point maps to the effectiveness characteristic (4.1.1). The exact nature of the interactions with users also depends on *EURHISFIRM's* chosen business model. The FAIR data principles can be a useful guide when it comes to user-need orientation. A data model that is implemented on top of the FAIR data principles requires data to be findable, accessible, interoperable, and re-usable. These principles, in conjunction with *EURHISFIRM's* cultural heritage perspective (see also evaluation point V.ii) are compatible with business models in which data access is open.

Model designs based on open data constitute suitable platforms for information access from the general public. General public access, in turn, can be helpful in promoting *EURHISFIRM* as a concrete, European research infrastructure. With data available online to everyone, *EURHISFIRM* is not an abstract formation of an institution offering specialized services, but rather a service that every European citizen can access and appreciate.

Open data access models are related to public good business models. *EURHISFIRM* may not act as a data vendor, but instead, its research infrastructure can be based on the services it provides on top of the historical financial data. Research in European economic history requires data cleaning and harmonization. Moreover, there is information that is used in research that is not available in the original data. It can be obtained only after applying the necessary data transformation(s). For instance, returns are not given in the original data, but they can be derived from them.

Providing data harmonization, creation of indices, and other on-demand transformations targets specialized scientific audience. The informational context of these services goes beyond the cultural aspect of *EURHISFIRM*. Such services, therefore, can also be provided under subscription. The output of these services can be stored in the derived information layer of the CDM. Such a practice is practically efficient, as for instance there is data transformation, like calculating returns, which are commonly used by researchers.

User-need orientation is not exhausted with the services that the CDM supports. Another essential aspect is content access. In turn, content access has two major components; namely, one that concerns content access methods and one that concerns content accessibility.

Is the model's content conveniently accessible?

The first component is the availability of access methods that fit particular user-needs. It is unlikely that a research infrastructure with *EURHISFIRM's* scale and scope can cover every user-need by providing a single access method. General, public-type of users, for instance, might prefer browsable content. Researchers, on the other hand, might as well prefer access with parametrization or programming capabilities. Therefore, it is important that the CDM's access platforms are suitably targeting all the interested audience. This point essentially corresponds to the satisfaction characteristic (4.1.3) of QUM.

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

Is the model sufficiently documented?

The second component is the existence of extended documentation. Documentation of data items, service provision, access methods, and derived data construction enhance the user-centric orientation of systems. Tutorials, online courses, and video presentations are also tools that can potentially help establish the CDM in the scientific community. Documentation is central for satisfying both the recognizability and learnability sub-characteristic of the PQM's usability characteristic (4.2.4).

ii. Cultural heritage and public good aspects

This point is to be co-evaluated with the legal compliance point. In cases that copyright protection laws of physical archives are applicable, the promotion of cultural heritage through the online exposition of scanned archives might conflict with the legal compliance criterion.

Does the model emphasize the cultural aspects of the sources?

The cultural heritage criterion widens *EURHISFIRM*'s scope to more than a research infrastructure of historical, financial data. The cultural aspect of *EURHISFIRM* is an important promotion vehicle for its positive reception, public opinion endorsement, and adoption. Therefore, any CDM design needs to incorporate this aspect and facilitate its development.

In this respect, *EURHISFIRM* can add another dimension to its design and content by documenting the original sources. The data collection processes in both the Antwerp and Paris models use a variety of historical archives. Both models already contain data fields that describe the printed sources from which the information is derived. As museums provide accompanying information to their exhibits, so can *EURHISFIRM* provide to interested users historical information (i.e. metadata) supplementary to the scanned archives.

This point is not only relevant in a cultural heritage promotion context. Bibliographic information on the sources, especially when these are linked with scanned versions of the archives can provide researchers with a valuable tool for evaluating the informational content. *EURHISFIRM* can utilize parts of the output of Work Package 4 not only as an input for the common model definitions but also as accompanying metadata of the printed archives.

Besides the above point, the inclusion of such a service in *EURHISFIRM* is also beneficial in terms of the credibility, precision and traceability characteristics of DQM. The end-users of the *EURHISFIRM*'s services will be able to validate the information that they use against a scanned copy of the original information source.

iii. Harmonized information access

This point concerns explicitly one of *EURHISFIRM's* potential services mentioned in subsection V.i. Harmonization can be one of *EURHISFIRM's* workhorse end-products. Simple data access in a user-friendly, browsable manner may as well be freely available. Harmonized data provision that is more appropriate for usage in data science projects and research in economics might as well be offered under subscription. Harmonization services can be offered both on per user demand basis and by supplying preconstructed data sets.

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

Does the model promote derived data generation?

The multilayer storage classification approach (see section IV) can be adopted by the CDM. At an initial stage, such adoption could conceptually separate the standardization of historical information and the standardization of harmonized data in the CDM's design space. At a secondary stage, such a storage classification can be also helpful in specifying access types and organizing user access.

The accounting data harmonization discussed in section IV is indicative of the flexibility of the multilayer storage classification. The historical accounting data belonging to the original information layer can be used to promote *EURHISFIRM*'s cultural heritage aspect. The generated contemporary international accounting data belonging to the derived information layer can be part of the harmonization services that *EURHISFIRM* provides.

Promoting derived data generation is not a static concept. For example, the CDM cannot affect whether the international accounting standards or any other standards it potentially adopts change in the future. The CDM can affect however its adaptability to such cases. The multilayer storage always retains a copy of the original data. If accounting standards change, the transformation procedure can be adapted, and the generated accounts can be accordingly updated. This is an approach that allows the system to provide its services in contexts beyond those initially specified. Such an attribute essentially corresponds to the flexibility sub-characteristic (4.1.5.2) of QUM.

Harmonization is one of the cornerstone requirements of *EURHISFIRM*. One of the unique selling points of *EURHISFIRM* is the provision of harmonized access to historical, European firm-level data. Existing implementations of the consortium already provide access to firm-level data at a national scope. In contrast, there is no implementation that provides harmonized access to firm-level data at a European scope. In terms of the PQM, the centrality of this unique selling point is reflected in the functional suitability characteristic (4.2.1).

iv. Integrated identification

The identification issue is central in most if not all data models. The investigation of the possibility of an overarching identification system is also part of WP5's future work. Identification is also related to the ease of access point V.i and specifically with the findability attribute of the FAIR data principles.

Does the model offer cross-country entity identification?

The identification systems in the consortium's flagship Antwerp / Paris implementations are based on pairs of firm names and founding dates. This system is embedded in the design of the respective models. Identifying a new firm from source information is entirely within the model's territory. This is advantageous when it comes to issuing new identifiers, as it is a reasonably uncomplicated approach. However extending this approach in a potentially decentralized, multi-country model is not straightforward. The CDM has to accommodate identification needs that span and intermingle over many European countries. Firms from one country have historically operated with various legal formats in other countries. Identifying the firms and the links between them requires a design that is able to address such special issues.

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

To maintain the country-level model flexibility on issuing identifiers, the CDM identification system can be split into two steps. Initially, each country issues local identifiers using a standardized common format. Before issuing the common identifier, there is a deduplication process, i.e. a process to specify if some other country has also issued a temporary identifier for this entity. The deduplication is performed at a cross-country level and at the end of this process a unique, CDM identifier is issued. Country models are then potentially updated accordingly. Such an approach allows national implementations to advance in data collection independently and promotes unique, cross-country firm identification. Deduplication and identification processes are complex and the setup of such services requires a sufficient allocation of resources. The resulting deduplicated data should satisfy the consistency characteristic in terms of DQM.

Does the model's identification follow best identification design practices?

The topic of unique identification is not new in computer science literature. (Paskin, 1999) discusses the properties of unique identification drawing experiences from the practices adopted by the DOI. Chapter 2 of (Berman, 2018) focuses more on practical unique identification issues in Big Data systems. *EURHISFIRM* can benefit by applying concepts and practices that proved to be useful in the identification design of relevant systems. A system that provides not only internal but also external identification of entities is related to the recoverability attribute of DQM.

The GLEIF identification system is designed using best practices of identification theory. It offers a Legal Entity Identifier (henceforth LEI) based identification system based on the standard (ISO, 2019). The LEI standard is designed to provide unique and unambiguous identification of legal entities. The latest published version of the standard specifies banks, companies, debt issuers, security issuers, stock-exchange listed entities, and sole traders among the eligible legal entities for which identification is supported. The LEI identifier is a non-intelligent, alphanumeric code of 20 characters long. Identification is a central part of WP5's future work and a more elaborate discussion on best practices and the specification of identification is postponed for future deliverables.

Is the model's identification compatible with other systems?

A fundamental business-driven principle of *EURHISFIRM* is its interoperability with contemporary database systems. In this direction, *EURHISFIRM* can benefit from adopting an LEI-based identification system. Historical company data from the second half of the 20th century contain many firms that either currently operating, or were operating until recently. Such firms are already identified in contemporary databases. Using an identification system such that of GLEIF potentially offers a solution for connecting the historical information content of *EURHISFIRM* with content from other data services that have also adopted the same or any compatible identification regime. This interoperability solution requires either that the identifiers that *EURHISFIRM* issues share the same namespace with the adopted identification system, or that there exists a mapping between *EURHISFIRM*'s and the LEI-based system namespaces.

As flexibility for issuing identifiers in each country is of central importance, so is the flexibility of issuing identifiers at the CDM level. To avoid conflicts with existing identifiers, a unique prefix from the adopted identification standard can be reserved for *EURHISFIRM* issuing purposes.

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

Is the model's identification compatible with changes in the companies' legal structures?

The proposed adoption of the identification system comes with an additional advantage. Incorporating an LEI-based identification system offers a framework for handling the changes in the status of firms. Corporate actions that change the legal status of a firm join two entities into a single one, or describe the cease of operations of a firm can be successfully handled by LEI based designs. Likewise, LEI based-systems can handle changes in other secondary firm data.

LEI-based unique identifiers of international scope can be inserted into the system in two ways. Either in a mapping table that contains multiple identifiers and their relation to each other or at a system level. In the first case, there is a local identification that is handled by the system, and the common model identifiers are used only for entity deduplication. This approach is adopted in the Wikibase implementation of the DFIH database. The alternative system-wide approach is to use the identifier at the database level also for record identification. Both approaches in principle lead to the same entity deduplication results, which is the main objective of *EURHISFIRM*. In terms of implementation costs, there is no technology that gives the system-level approach out of the box, which probably renders it more demanding in terms of implementation labor hours. Interoperability with external systems

Is the model connected to contemporary information?

The model design should allow for connections to existing central databases (for instance EUROFIDAI and LSPD). As noted in the integrated identification subsection V.iv, linking historical data to recent data is an essential cornerstone of the CDM design. The system should not be designed to be anachronistic and introspective. (See proposal, 5.2, p36). A carefully chosen identification mechanism can promote inter-system connectivity. This interoperability point corresponds to the compatibility characteristic (4.2.3) of PQM.

Besides identification, *EURHISFIRM* can also provide interoperability as a service. Analogously to CRSP's practice to offer CCM as an end product, *EURHISFIRM* can form partnerships with third-party data providers and link historical data to contemporary data. For the entities that were successfully matched, the CDM can store linking information in the form of collections of pairs of CDM and external models' identifiers.

The CRSP/Compustat experience shows that there might be disagreements among experts when it comes to linking entities across models. A similar situation can potentially be encountered in any *EURHISFIRM*'s linking process, and it should be anticipated that interoperability services are accompanied by various degrees of confidence. Providing services with various degrees of confidence gives the opportunity to individual researchers to choose the degree that better fits their needs.

v. Legal compliance

This criterion concerns the main focus of Work Package 3. There are two essential points at which the CDM intersects with the legal and ethical design. Both of these points correspond to the compliance characteristic of the DQM model.

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

Is the model copyright law complaint?

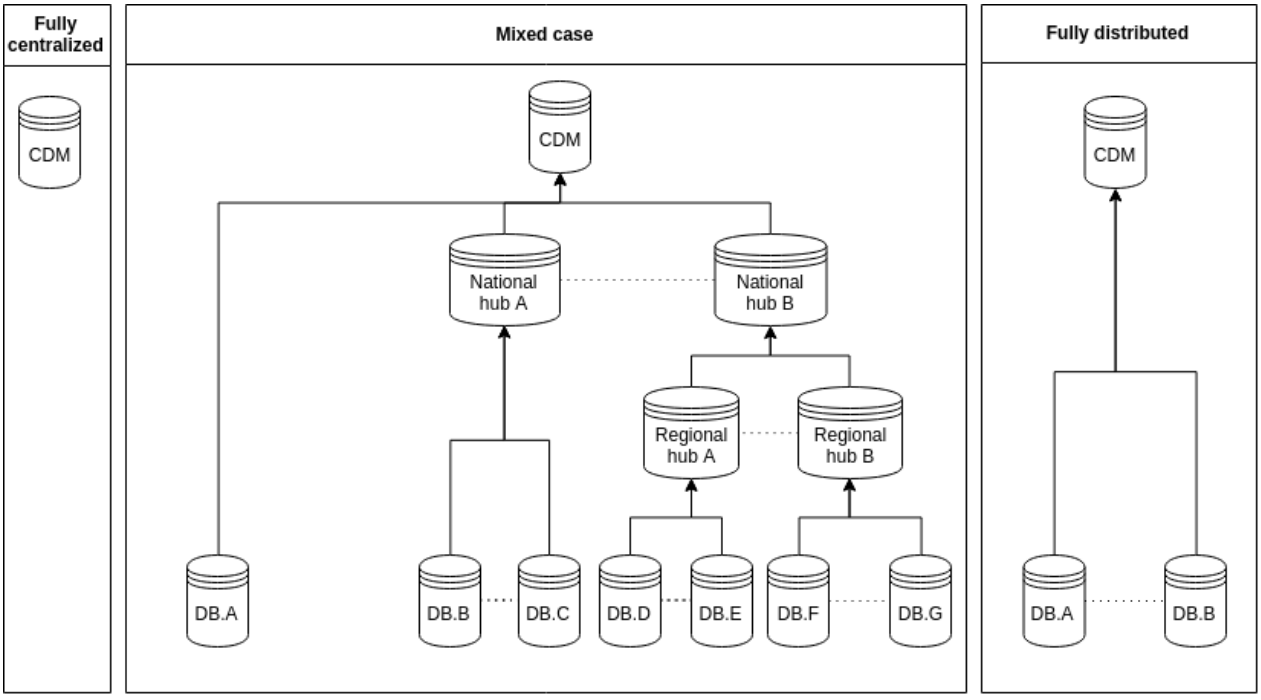
In some countries of the consortium, the historical archives that constitute *EURHISFIRM's* primary information sources are subject to copyright laws. This can potentially limit the information space that the national implementations in such countries can cover. The CDM should have the flexibility to accommodate both restricted and unrestricted national implementations.

Is the model privacy-law complaint?

One part of the information space of *EURHISFIRM* is that of companies' corporate governance. This part potentially concerns information about historical physical persons that were involved in various organizational levels of firms. Both the national implementations and the CDM design should only provide information that complies with privacy laws.

vi. Optimal decentralization degree

The degree of centralization is a defining underlying point in the design of the CDM. It is also strongly connected with the architectural design of Work Package 9. Diagram 8 conceptually depicts systems with various degrees of centralization. The decentralization discussion in this section primarily focuses on the logical, and not physical, decentralization aspects. On the one hand, the fully centralized system consists of a single system node. On the other hand, the fully decentralized system consists of many nodes, each representing a participating party in the common model and each potentially having a separate model implementation. The mixed-case consists again of many nodes with potentially different implementations. In this case, however, the system hierarchy has more than two levels. Nodes are not always on the same horizontal hierarchical level, rather a multiple-country, or even cross-country, implementations are consolidated in sub-system hubs and then further integrated into the common data model. Choosing the right decentralization degree is connected to the efficiency characteristic (4.1.2) of QUM.



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

Diagram 8. Various decentralization designs

Is the model implementation cost-efficient?

Cost efficiency is included in the performance characteristic (4.2.2) of PQM. The fully centralized system is maintenance-cost efficient. There is a single model implementation that needs to be maintained and updated. The maintenance costs rise with the number of different model implementations in decentralized systems. Each separate implementation requires IT-experts to support it. In decentralized designs, updates and extensions are more time, labor and resource-intensive as the changes have to be diffused in both the CDM and the various sub-system implementations. Moreover, connecting decentralized sub-system implementations to the common model creates the need for setting up access requirements and communication protocols. Lastly, seamless operation of the common model requires that the communication interface with the sub-system is not broken, i.e. that all changes on a sub-system level are communication conformant and assimilable from the CDM.

Fully centralized systems are also simpler to implement. For instance, if there exists only a single system node, there is no need to establish communication protocols and standards between the parts of the system. In application development terms, the service does not need to handle subsystem communications, which typically come hand to hand with encryption and data protection requirements.

In decentralized designs, one should also consider data duplication and service atomicity issues. This point relates to the connection between local implementations and the common model. The common model can potentially retrieve data directly from local implementations or store a copy of the data and provide this copy to the end-users. In the first case, there is no danger of users ending up with different versions of the data. Either a user retrieves data directly from a local implementation or by using an atomic operation through the common model, she will end up with the same information. In the second case, if local nodes have updated versions of data that the common model does not, the informational content that a user receives is access-node dependent. Such a system design can potentially lead to data races.

Is the model implementation flexible to adjust to country-specific needs?

An advantage of decentralized designs is that they allow for more flexibility to local implementations. Local implementations might as well go beyond the common model standards and implement region-specific features. Provided that there are sufficient resources available, such flexibility can enhance the advancements in local systems and subsequently also in the common model. A similar approach is also used in the standardization of C++ compilers. The C++ standard defines a minimum set of compilation requirements. Compiler vendors typically offer functionality that surpasses the scope of the standard. Many of the successful extra features are then adopted by subsequent versions of the standard.

When it comes to the business side of the model, there are successful business cases for both centralized and decentralized data services. The centralized case corresponds to a typical data vendor service. An example of a decentralized data service design is WRDS. The service is based in the University of Pennsylvania and has been active for more than twenty years. It provides harmonized, single interface access to services such as S&P Capital IQ, CRSP, NYSE, Thomson Reuters, Bureau van Dijk, Global Insight and OptionMetrics.

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

The centralization degree is by far not only a technical point. The political processes in other areas of the European unification are indicative that a decentralized and federated architecture offers an appropriate organizational model for many European Union institutions. A decentralized system, despite that it is accompanied by greater technical and organizational costs, gives local implementations the flexibility to develop independently and avoids the need to establish a politically challenging, central data governance mechanism. A federated common data design is feasible under the current situation of the consortium's countries and also paves the way for closer future cooperation.

vii. System adoption and community building

This issue is the Work Package 10's object of primary focus. System adoption and community building are not per se technical evaluation points. However, the business model is tightly connected to key model design attributes. A common model that does not support user interactions, user's ability to propose content modifications, and the user's ability to promote their own work are not compatible with community-based business models.

The connection of the CDM and the business model at this point can be illustrated when one considers two simplified scenarios; on the one hand, a fully public good, community-based business model and on the other, a fully proprietary, data vendor business model. It is likely that *EURHISFIRM*'s final design will borrow elements from both sides of this spectrum. The public good business setup, among other characteristics, may contain a data model with information that is available entirely for free on the web. The security considerations of such a design are primarily focused on protection against web-scraping and data extraction attacks. The extent of such considerations is substantially limited in a proprietary business approach. The less actual information is freely available online, the less protection against data harvesting is needed. On the contrary, security considerations related to access rights, copyrights, and data protection are much more prominent.

One important component of *EURHISFIRM's* research infrastructure is the promotion of digital preservation and accessibility of sources with significant historical value (See Work Package 11, *EURHISFIRM* proposal p 49). This component is indicative that *EURHISFIRM* has aspects that are of public good nature and points towards the direction of such business models. Besides this, comparing *EURHISFIRM* informational content with that of CRSP, EUROFIDAI, and other contemporary information data vendors, one can easily observe that *EURHISFIRM*'s capacity in providing data updates is limited in comparison with them. CRSP and EUROFIDAI provide contemporary data and can update their content based on financial developments daily, while *EURHISFIRM* aims to supply historical information.

Although the last observation is not indicative of the inferiority or superiority of one infrastructure over some other, it points out that a pure data vendor business model might be inapplicable for *EURHISFIRM.* An intermediate approach that involves both public excellent characteristics and harmonization, pre-processing data services can be more appropriate for *EURHISFIRM*.

Does the model accommodate community-centric processes?

To support its public good aspects, *EURHISFIRM* can follow a community-based approach. A community of economic and history scholars, as well as practitioners and data scientists, can offer a viable, financially

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

sustainable approach for promoting, maintaining, and extending *EURHISFIRM's* public good aspect. A fundamental requirement for the growth of such a community is an online platform through which the members of the community can interact and contribute.

Essential interactions points of the community are

- The ability to validate the information against the sources. Users should be able to check the validity of the data against the scanned versions of the original sources. A modification process should also enable changes in the stored information in cases that users find inconsistences. Any modification proposal can be stored initially at the staged information level of the CDM. Subsequently and conditionally on *EURHISFIRM* experts' approval can be incorporated into the original information layer.
- The ability to contribute derived/transformed data. Researchers might, for instance, want to contribute the generated index data that they constructed using an innovative approach. This functionality can be limited to a verified subgroup of the community with sufficient access rights. This practice is limitedly adopted by the consortium's SCOB and DFIH implementations. The output produced from the family of desktop output interfaces for various research projects was, on some occasions, subsequently added into the databases. The CDM can also allow storing the user constructed data and possibly the code that was used in order to construct them. The user-generated data can be stored in a separate (sub-) layer of the derived information layer and the source code to accompanying repositories. When appropriate, either the transformed data can be included in *EURHISFIRM*'s main service, or the methodology that was used to derive the data can be adopted by *EURHISFIRM*'s system.
- The ability to browse information in a human-readable format. This point should meet the needs of historians whose research is based on reviewing information on particular documents. Furthermore, this point might be relevant for other users that search isolated information for non-statistical research purposes.

Does the model accommodate incentive mechanisms in favor of its adoption?

Meeting the platform requirements cannot on its own guarantee that a community will successfully grow around *EURHISFIRM*. In the long-run, the success of the community endeavor will be primarily based on the participation of people that are not directly employed by the project. A mechanism to attract people outside of the consortium is the introduction of appropriate incentives.

A model characteristic that can act as an incentive mechanism is the provision of data and research catalog functionality. This can be a side aspect of the data model. The system may enable researchers not only to be able to access *EURHISFIRM*'s data but also to distribute the datasets resulting from their work to other users. Other researchers may then properly cite the datasets constructed and saved in this way. Such functionality facilitates the promotion of researchers' work by increasing web visibility, academic credibility, and the number of their citations. In this direction, *EURHISFIRM* can benefit from available cataloging infrastructures such as GESIS and be part of a larger trend towards increased scientific output transparency, which involves initiatives like Harvard's data-verse.

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

Does the model provide users with safe access to its services?

This point is based on the freedom from risk characteristic (4.1.4) of QUM and to the security characteristic (4.2.6) of PQM. Any form of access to and interaction with the system should guarantee the safety and privacy of the transactions with the users. In particular, the community-based part of *EURHISFIRM*'s system will involve storing some personal information of its users. Some elements of this information can be publicly available. For instance, users that use *EURHISFIRM* supporting platforms for the promotion of their work can choose to release some private information. Other elements of this information are meant only to be kept private and *EURHISFIRM*'s system, as any online platform, should securely handle the personal data.

viii. System extendibility, updatability, and sustainability

Extendibility, updatability, and sustainability are points that are relevant to the operational smoothness and longevity of the CDM. In this context, extendibility refers to the ability of the CDM to include information from sources other than the ones currently considered by the consortium. Updatability refers to the ability of the model to upgrade or modify the information obtained from the currently considered sources. Sustainability refers to the relevance of the CMD design for supporting *EURHISFIRM*'s research infrastructure in the long run. In terms of PQM, these points partially correspond to the reliability (4.2.5) and maintainability characteristics (4.2.7).

Is the model extensible?

This point is central to the CDM design. In future stages of *EURHISFIRM,* there will be a window for more European countries to join the consortium. The model should be designed in a way that takes future expansions into account. Ideally, the model should be adaptable enough to describe legal and accounting standards from countries that are not part of the consortium.

The extendibility of the system does not only depend on its ability to describe the informational context of potential future participants. It also depends on its ability to allow information with different data formats to be injected into the system. Countries outside the consortium Portugal and Sweden have already progressed in terms of data collection. In contrast to the implementations reviewed in this report, the data on some occasions are organized into spreadsheets. To accommodate the inclusion of such cases into the consortium, the CDM should both be able to flexibly depict the newcomers' legal systems and assimilate their data formats.

Is the model updatable?

Populating the CDM with data is not a one-off procedure. There is an initial chunk of information that the common model can assimilate from the existing national implementations. However, this is only a small part of the information that is available in the printed archives. As printed archives from various countries are digitized and parsed by the OCR system gradually, the CDM has to possess the ability to assimilate the incoming information.

Ideally, the CDM could also be able to provide feedback information to the OCR system. Information already contained in the CDM can be used as input data for training the OCR libraries. Moreover, already validated data offer a benchmark for validating the accuracy of the OCR output.

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

Is the model sustainable?

EURHISFIRM infrastructure contains elements that suggest its relevance on a long-term horizon. Firstly, economic research in European, historical firm-level data is currently a relatively unexplored area due to data unavailability. *EURHISFIRM* has a central role in changing the status quo in this area and facilitating research in the forthcoming years. Secondly, *EURHISFIRM* digitizes and saves part of the common European cultural heritage. This point is also indicative of the long-run nature of the project.

The CDM should also reflect the longevity targets of the project. Any design solutions should be compatible with *EURHISFIRM*'s role not only in a contemporary context but instead also with its future placement in the research community of economic history.

VI. Conclusion

This report introduced a methodological 2x2 separation on the information space that is relevant for *EURHISFIRM*. One dimension of the separation concerned time and the other one concerned the cross-country elements of the common model. Representative models for contemporary cross-country, historical national and contemporary national data from both within and outside the consortium were reviewed.

Based on the reviewing process, the report developed concepts, ideas, methods, and arguments that specify central design elements of the CDM. The first-level output of the reviewing process was the identification and collection of special points and initial metadata concepts that are highly relevant for the CDM's design. On a second level, the report focused on how the reviewed implementations tackled difficulties associated with specific points of interest in the design space. The output of this comparative overview demonstrated best practices from which the CDM design can benefit by adopting them. It also demonstrated weaknesses in the current implementations. These weaknesses are augmented by possible extensions, the adoption of which can also be beneficial for the CDM.

The report concludes by deriving from the preceding analysis a set of evaluation criteria. The methodology proposes nine essential criteria of central consideration in the design of European, historical, financial company-level data models. Each criterion is subdivided into a set of primitive questions that can help assess the performance of the design in this direction.

In terms of overall quality assessment, the analysis of the report indicates that the consortiums' implementations are successful in providing extensible, updatable descriptions of the national, historical, financial company-level information spaces. The vocabulary used for many of the main modeling concepts, however, is obscuring this fact. The main weaknesses of the consortiums' implementations concern user access related issues. Implementations outside the consortium are more user-oriented and facilitate their adoption by providing extended documentation, user interfaces, and access possibilities.

VII. References

Adam, S., Bouvier, S., Coüasnon, B., Guerry, C., Lemaitre, A., Paquet, T., … Swaileh, W. (2019). *D7.1: Software libraries developed*.

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

Adams, R., Gareth, C., Christopher, C., & John, T. (2019). *D8.2: Questionnaire results*. Belfast.

Annaert, J., Buelens, F., & De Ceuster, M. J. K. (2012). New Belgian Stock Market Returns: 1832-1914. *Explorations in Economic History*, *49*(2), 189–204. https://doi.org/10.1016/j.eeh.2011.11.004

Berman, J. J. (2013). *Principles of Big Data*. Elsevier. https://doi.org/10.1016/C2012-0-01249-5

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, *52*(1), 57–82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Chandler, A. D. (1992). What is a firm?: A historical perspective. *European Economic Review*, *36*(2–3), 483–492. https://doi.org/10.1016/0014-2921(92)90106-7

CRSP. (2018a). *CRSP/Compustat merged database guide*. Retrieved from http://www.crsp.com

CRSP. (2018b). *Programmer's guide*. Retrieved from http://www.crsp.com.

EUROFIDAI (2016), Data Description Guide: European and Asian Corporate Events Daily Database.

Fama, E., & French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*, 3–56. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.139.5892&rep=rep1&type=pdf

Foss, N. J. (1993). Theories of the firm: contractual and competence perspectives. *Journal of Evolutionary Economics*, *3*(2), 127–144. https://doi.org/10.1007/BF01213830

Hautcoeur, P. C., & Riva, A. (2012). The Paris financial market in the nineteenth century: Complementarities and competition in microstructures. *Economic History Review*, *65*(4), 1326–1353. https://doi.org/10.1111/j.1468-0289.2011.00632.x

ISO. (2019). *17442:2019 Financial services - Legal entity identifier (LEI)*. Geneva.

ISO. (2008). *25012:2008 Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model*.

ISO. (2011). *25010:2011 Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models*.

LSPD. (2019). *London Share Price: lspm201812 & lspd201812 Reference Manual*.

Paskin, N. (1999). Toward unique identifiers. In *Proceedings of the IEEE* (Vol. 87, pp. 1208–1227). https://doi.org/10.1109/5.771073

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

Riva, A., Annaert, J., Köning, W., De Jong, A., Jajuga, K., Turner, J., … Katsanidou, A. (2017). *EURHISFIRM Proposal*.

VIII. List of abbreviations

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

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