Long-term data for Europe

EURHISFIRM

D5.2: Technical Document on Preliminary Common Data Model

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

AUTHOR:

Pantelis Karapanagiotis[1]

ABSTRACT:

The second report of Work Package 5 completes the discussion of the preliminary back-end design concepts of the common data model. The approach of the report is characterized by the principle of least intrusiveness. The proposed solutions respect national idiosyncrasies and allow national centers to advance in a collaborative but independent manner.

The report starts by reviewing the data formats of the countries of the consortium. It draws from identification theory and proposes appropriate principles and requirements for the common model's identification design. It examines the functional and informational requirements for identifying various data items and for linking historical data from within the consortium to external databases with contemporary data. It outlines that the common model's implementation mostly benefits from employing both relational and non-relational technologies to address different issues. It highlights appropriate, subsequent steps for cross-country harmonization, firm-linking, data transformation processes, and data governance.

APPROVED IN 2020 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 and Lukas Manuel Ranft for their valuable contributions during the composition of this report. I am grateful to Angelo Riva for giving me access to the DFIH database. I am thankful to Stefano Battilossi, Stephanie Collet, and John Turner for providing samples from the data collections of Spain, Germany, and the UK. Without them, the analysis of the formats would have not been possible. I would also like to thank Coen Fierst van Wijnandsbergen, with whom I coordinated in some of the implementation topics of the report. I am thankful to Lut de Moor and Lana Yoo for their support while writing this report. I want to also thank Johan Poukens, the comments of whom helped me to avoid many pitfalls and David Smadja with whom I discussed many details of the implementation topics.

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 heterogeneous not only in terms of electronic formats but also in terms of content. *EURHISFIRM* aims to provide a research infrastructure that homogenizes this information in both directions. Work Package 5 (henceforth WP5), in particular, focuses on designing a common data model that supports *EURHISFIRM*'s aim.

The current report presents some preliminary design concepts for the common service. It proposes strategies and solutions for fundamental elements of the common model design. In particular, it sets requirements for the common identification system, for establishing links with contemporary databases, for the harmonization of the national informational content, and for the homogenization of national formats. Furthermore, it evaluates the implications of choosing between different technological frameworks. Last but not least, the report introduces a high-level stepwise process for the implementation phase of the common service.

The topics of the report are approached from an applied perspective. In cases where some general theoretical background is needed, it is only introduced in the context that is relevant for *EURHISFIRM*'s common service. This is, for instance, how the functional requirements of identification and the technological framework evaluations are discussed. The report attempts to provide solutions and, therefore, the discussion of the topics is mostly applied. A few academics points are drawn to facilitate the discussion.

The remaining report is organized as follows. Section II reviews the national data formats of the consortium's data collections. Section III provides a design specification for the identification system. The specification concerns the functional and informational requirements of the system. Section IV evaluates a series of technical, architectural, and business-logic alternatives for the design of the common service. Section V lays the foundation for the harmonization and consolidation of national data. Section VI catalogs methods of transforming national data into the common format and proposes a stepwise implementation process. The last section concludes.

II. Review of national data formats

This section contains a selective description of the data formats that are used in five major application cases; namely France, Belgium, Germany, the UK, and Spain. As suggested in (Riva et al., 2017), the first four study cases are indicative of the national variations that *EURHISFIRM*'s model has to accommodate. The fifth case is not included in (Riva et al., 2017) but also contributes to the discussion of the formats.

The first aspect of the considered variations is the legal one. There are three legal families considered; namely the French civil-law (Belgium, French, and Spain), the German civil law (Germany) and the Anglo-Saxon common law (the UK) families. As it is shown by the following review, this aspect has not extensive implications concerning the data format[2]. The second considered aspect concerns the national

[2] The legal variations have important implications concerning the design, the schematics, and the content of a data model. A price can be calculated or reported differently in different legal systems and different schematics are

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

implementations' progress. Two alternatives are considered. Advanced data model implementations, such as "Data for Financial History" and "StudieCentrum voor Onderneming en Beurs" (henceforth DFIH and SCOB respectively), and data collections organized in datasets, as in Germany, Spain, and the UK. As it is shown in this section, this aspect has strong implications for the data formats. The third considered aspect is the language. This aspect is not prescribed in (Riva et al., 2017) and it has no implications from a strict data format perspective. However, it has important implications concerning identification and harmonization and, therefore, it is also shortly discussed here.

The DFIH and SCOB models have similar implementations and they both contain information that is generated in French civil law countries. To avoid repetitions, the review of their formats is bundled together. Whenever there is a specific need, a distinction between the two reviewed formats is made. The complete data format specification of these two implementations can be explicitly obtained by examining the Data Definition Language (abbreviated DDL) generating the databases[3]. This section reviews only a small subset of the available formats. Instead of recording how every data field is formatted, the review focuses on the fields that convey information regarding the identification and harmonization processes of the Common Data Model (henceforth CDM).

The data formats of Germany, Spain, and the UK are simple to review. The implementations are not as advanced as in the cases of DFIH and SCOB. There are not yet precisely specified relations among data elements and specifications of data fields. Nevertheless, they are included in this review as they can contribute to the discussion of the implication of legal system and language variations to the data format. Besides their contribution to the discussion of the data formats, their inclusion can be helpful concerning the transformation of datafile collections to structured databases. This point is also relevant for the future development of *EURHISFIRM* as any contributions from individual researchers are likely to be similarly formatted.

The format review of this section does not intend to examine only how a variable is electronically stored. In particular for cases without advanced implementations, such an approach is not helpful as the electronic format can be plain text. The review also examines composite information that can be derived by combining two or more data fields by established model relationships. This is of relevance for both the identification and harmonization processes of the common service.

i. Identification

The recording of firm-name fields in the language of the source is a central element of the 'preservation of the original information' design principle. It is also appropriate for identifying entities at a national level. For international identification, the names of the firms have to be translated into a common language. Both the translated name and the original company name are relevant identifying attributes for an

needed to describe such cases. In terms of formatting, however, a price is typically stored as a number irrespective of the background calculation.

[3] The DDL scripts are available for the members of the consortium. They are not included in this report to prevent the duplication of the database schemas.

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

economic entity. The translation of names is not only important for the international identification of entities but can also be helpful to promote the common service to wider audiences.

The founding date field is undoubtedly appropriate for identifying organizations that are founded using some constitutional document. It can record the date on which the constitutional document goes into effect. This is applicable to incorporated firms. For cases of unincorporated partnerships and states that are not founded based on constitutional declarations, this field can record the date that the business operations commenced and the historical founding date of the state accordingly. For administrative counties, the founding date can record the date in which the legislation prescribing the county has gone into effect. For municipalities, it is usual that the founding dates of the corresponding cities are not available. Moreover, the same city can historically belong in the jurisdiction of different states. Should in such cases the municipality be recorded using different entities? What if the debt of the previous administration is recognized by the following one? These problems are indicative that the founding date field might not be definitive for identifying such entities in a historical context.

Belgium/France

The identification of organizations of official and non-official nature is performed by a composite key in DFIH and SCOB. It is based on the name of the organization and on its founding date. This information is organized into two relational tables. The first table assigns an integer identifier to records of founding dates, liquidation dates, sources and an indicator of whether the organization is official or not. The second table uses a composite key that references the identifier of the first table and combines it with the organization's name and founding date. It also records information on the liquidation date and the source.

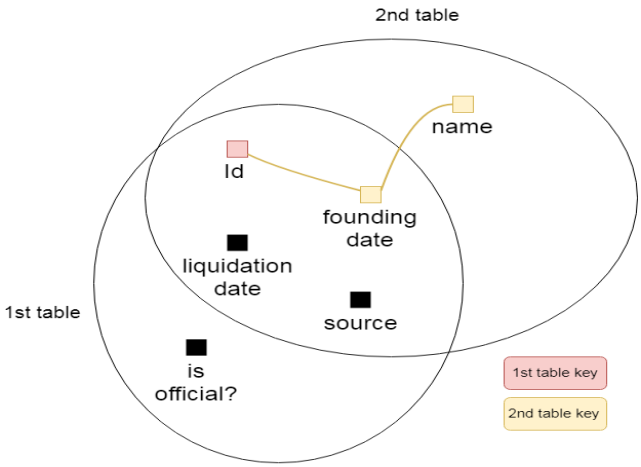

Figure 1. Identification design in DFIH and SCOB

The design is depicted in Figure 1. Each set represents a table. The (relevant) fields are points in these sets. The `id` field has a numeric format. The `founding` and `liquidation` dates are recorded in date fields. The `name` and `source` are recorded as text and the `official` organization indicator as a boolean. The `source` field is repeated into both tables. Potentially, it can be more efficient to store the source in a separate table and use a foreign key to the tables that contain organization data. Moreover, this design

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

allows for associating an organization with a single source. In practice, however, an organization can be found in many sources. The liquidation and founding dates are also present in both tables, which can be technically inefficient. The official organization flag is stored separately from the name of the organization. This requires a join when one wants to retrieve both the type and the name of the organization.

The elements that uniquely identify an organization are the name and the founding date. The names are recorded as they are found in the sources. The founding date is incorporated in the identification system to tackle the difficulties relating to the identification of firms in a historical context. For instance, two firms with the same name operating in different time periods can be correctly identified as separate entities.

Germany

In Germany, data are organized in twelve datasets that are stored in STATA and EXCEL formats. There are monthly observations from 1871 to 1875 and from 1882 to 1914. The data files, however, are not always linked. Some of the files contain aggregates and statistics of other files. An integer identifier is assigned to companies. The identifier does not persist across all files. In some cases, only the name of the company is given. In cases that an identifier is provided, the name is the element that identifies a firm.

Spain

In Spain, data are organized in datasets. There are datasets with monthly and daily index data and a dataset with company-level data. In the time domain, the data spans from 1913 to 1936. The identification of firms is based on their names. In the two index-datasets, the names are preceded by an ascending integer. Data for each company are saved in a different EXCEL sheet. This sheet also contains financial statements and dividend information. In the index datasets, data for each company are organized in different columns.

UK

In the UK, data are organized in datasets. The derived data are sometimes calculated from other data sets either in EXCEL or in STATA. The identification of firms is based on the company's name and the incorporation year of the firm. Every firm-entity is assigned an integer identifier and a string name. There is also a variable that concatenates the company's name and the incorporation year (variable `Company`), which is, of course, unique for every record.

ii. Quantities of issued securities

The number of available securities at a particular time-point is a starting point for a variety of data transformations that are commonly used by both academics and practitioners. For instance, the market capitalization of a stock is calculated as the product of its market price and the number of stocks available in a market. Historical sources do not always report this information in the same manner.

Belgium/France

This is reflected in the format of recorded quantities in DFIH and SCOB. To store an electronic representation that is faithfully close to the original source of information, the models are adapted to embrace how information is recorded in various sources. The storage of quantities is case dependent. When the number of securities in the market is available, it is stored directly. In some cases, nominal values

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

are available instead of quantities. In such cases, the face value is stored. Face values are primarily available for bonds and the number of securities for stocks.

Whether the stored value represents the number of securities or the market valuation of a security is indicated by the stored value of an accompanying quantity-unit attribute. The quantity unit attribute refers to a table that stores units that are used also in other parts of the model. In cases that the stored quantity value represents a market valuation, the model also provides information on the currency of the valuation. The valuations are recorded not only in French Francs but also in many other currencies. Types of currencies are stored separately. Exchange rate information is also stored separately. Last but not least, information on the type of the stored quantity is provided. The type of quantity signifies whether the recorded quantity refers to shares issued, or shares admitted for trading, etc. Quantity types are separately stored and extending the model to include more types is easy to achieve.

Quantities are not stored for each date separately. Rather, the period for which a quantity was issued is recorded using a starting and ending date. This is an economical design for storing quantities. No information is lost by the design. One can always locate the interval that contains a particular date and infer the quantity (or valuation) of a security on this date.

Germany/Spain/UK

Quantity data are not available in German data files. There is information on the level of subscribed capital both in annual records and in aggregate values.

Quantities are, in contrast, recorded in the Spanish data files. Specifically, the files contain the number of outstanding and admitted shares in total.

The quantity information available in the UK datafiles is comprehensive. The way that the data are formatted is straightforward; every type of quantity is stored in a separate table field. The quantities of available stocks are accumulated to also obtain the total number of available stocks. Besides the total number of stocks, the number of ordinary, secondary, preference, deferred and founder's shares are recorded in the datasets.

iii. Security prices

The security price information is one of the most central elements of the common data model. Possibly the most common data transformation is the calculation of market returns. Besides information on relevant corporate actions (for example splits, reverse splits, rights issues, etc), the calculation of returns requires that both price and dividend data available.

Belgium/France

The DFIH and the SCOB designs are sophisticated enough to cover the available price information that is found in the historical sources. The price information is mainly stored in the `notation` table-family. The `notation_price` table has, among other, fields that record opening, closing, min, max, and previous prices. The previous price is not necessarily the price of the previous day and, therefore, the date on which the previous price was marked is also recorded. If the price of the previous date is available, it is recorded separately. A logical field indicates whether prices are stored as percentages. There is also a field that

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

records the traded volumes. As the model is flexible enough to describe multiple stock exchanges, the prices are linked to particular exchanges.

The DFIH model further extends the model's ability to record price data. The table `notation_extra_price`, through referencing the notation table, associates notations with multiple intraday prices at a given date.

Germany/Spain/UK

The data files with German data contain prices as percentages of their nominal values. Multiple prices are not stored. In cases that one company is listed in multiple stock exchanges, only one price is stored. If price information for Berlin's stock exchange is available, then it is used. If not, prices from other stock exchanges are used. Following Berlin's stock exchange, the ones of Frankfurt, Hamburg, Cologne, Leipzig, and Munich are prioritized from left to right.

The company-level Spanish data files contain price information. The minimum, the maximum and the closing prices are recorded. The average of the minimum and maximum prices is calculated. Stock exchange information is not available.

The UK data contain information both on the shares' nominal values. These values are recorded for every type of share that is relevant for each security. There is also available information on the stock exchanges on which securities are traded.

iv. Accounts of financial statements

Financial statement data are also commonly used by researchers in economics and finance. For instance, the accounting performance of firms is used both for evaluating market prices in the cross-section and for predicting future prices.

Belgium/France

Granular financial statement data and aggregates are available in the DFIH/SCOB implementations. The relevant data are stored in the `corporation_bookkeeping`, `bookkeeping` and `aggregates` table-families. The firm-level data are recorded using a multi-table scheme. The main bookkeeping table records the amount of an account of a firm at a given date. This table is linked both with corporation information tables and with bookkeeping items tables. The bookkeeping items tables contain information on the accounts.

DFIH's model can flexibly support multiple aggregation models. For each model of aggregation, the information on aggregate accounts is self-referential. An aggregate account can refer to another aggregate account. This design reflects the structure of the financial reports found in the sources and allows the researchers to select the level of granularity of the data that they use.

Germany/Spain/UK

Granular financial statement data are not available in Germany, Spain, and the UK data files. The German and Spanish data contain records of the companies' subscribed capital. The UK data contain information on the distribution of profit among shareholders.

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

III. Overarching identification system

Does *EURHISFIRM* need an overarching identification system? *EURHISFIRM* intends to offer various important services. Undoubtedly, the cultural aspect of *EURHISFIRM* is of great general interest. Digitization and promotion of historical archives is on its own a valuable service. In terms of research infrastructure, providing the researchers with access to consolidated, harmonized, historical, European data are central among *EURHISFIRM*'s unique selling points.

Local databases support research in economics and finance at a national level within the geographical area of Europe. Indeed, a variety of economic and historical inquiries can be assessed without establishing links between national data. Other inquiries, however, can only be investigated from an international perspective. Research concerning historical international finance in Europe is sparse.

This section discusses the introduction and use of unique, overarching identification for European, historical firm-level data. The objective of the section is to examine general concepts of identification theory in the context of historical, cross-country firm data. In this respect, the section is not an abstract essay on good identification practices. It, instead, assesses the relevance and applicability of these concepts in *EURHISFIRM*'s model. Moreover, it provides a comprehensive set of informational and functional requirements of the identification procedure. The functional requirements address the technical necessities of how the objects of interest are identified. The informational requirements address the topic of which objects are to be identified.

This applied approach is sometimes limited by the ambiguity that has not yet been resolved concerning the business model of *EURHISFIRM*. As argued in (Karapanagiotis, 2019), the modeling space is interconnected with the organization structure and the business goals. Different models may better suit different business structures. In particular, when it comes to the identification system, the requirements are mostly neutral. However, there are "implications which are beyond the technical ones; for example, the recommendation for the administration system for an identification system to be decentralized but controlled (i.e., the responsibilities for maintaining the system should be parsed among a number of participant groups) raises a number of political issues regarding who runs the identification system and how the costs of running the system are apportioned" (Paskin, 1999).

i. Functional requirements

The two main sources of theoretical concepts used by this section are (Berman, 2013) and (Paskin, 1999). Not all the concepts discussed in these sources are relevant for the discussion in this section. The focus is laid on concepts that are relevant for the identification of economic entities in the sense of the definition given in (Karapanagiotis, 2019). Whenever relevant, the notion of an economic entity is specified more concretely to facilitate the presentation of the ideas.

Identification in the context of this report is the process of unambiguous labeling that specifies entities of interest. The labeling concerns objects of the logical domain of *EURHISFIRM*'s model. The labels are intended to be used by *EURHISFIRM*'s system. The labels are referred to as identifiers. The following

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

paragraphs are indented to be used also as a reference and, therefore, are self-contained. They are not ordered by significance, but lexicographically.

Capacity/Periodical labeling elements

Capacity is a necessary characteristic of the identification system. It requires that the system has enough distinct labels to assign to the identified entities. A compound identifier design that has elements that are updated periodically satisfies this property. For instance, the year and the month during which the identification of the entity occurred can be part of the identifier.

The periodical elements in the labels insert intelligence in the identification system. The intelligent part of the identifier, however, does not reference attribute values of the underlying objects to be identified, rather some information that is unrelated to the identified object and it is only used to increase the capacity of the system. Therefore, the waterfall effects, i.e. the necessity to alter the identification system when the underlying attributes of the identified object are changed, do not apply in this case.

Completeness

An identification system is complete if a label is assigned to every object of interest. Completeness is a theoretical property that is unlikely to be fully satisfied in *EURHISFIRM*'s context. For example, non-listed, private firms that are not recorded in available historical sources cannot be identified.

A conditional version of the completeness property is a more appropriate guideline for *EURHISFIRM*'s identification system. The identification system is conditionally complete if a label is assigned to every recorded object in the historical source. Even in this relaxed form, the completeness property is to be understood as a guideline or goal and not as a strict requirement. In terms of firm identification, completeness is a goal that can be reasonably targeted. Targeting completeness concerning the identification of natural persons can be inappropriate for *EURHISFIRM*'s system.

Computability

An identifier is computable if it can be derived from the data and metadata of the identified entity. Computability can be thought of as the counterpart of intelligent identification. An intelligent identifier allows the extraction of metadata from the identifier, while a computable identifier can be obtained when the metadata are available.

The computability attribute can be helpful in the design of a two-level identification system in *EURHISFIRM*. That is, a system in which national research centers assign a local identifier to harvested data and, then, global identification and deduplication are performed at a European level. By agreeing on common identification metadata, computability can provide guarantees that the entities identified by different national research centers are compatible. Such an approach will also reduce the identification costs at a European-wide level. In terms of implementation, hashing algorithms can potentially be of relevance in providing computability.

Documentation and quality assurance

The standardization of the identification system should document protocols for the identifiers and their usage. Specifically, standardization for establishing the identifier, the assignment of identifiers to entities,

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

and the governance of the system are needed. Changes and updates to the system should also be thoroughly documented.

Granularity

Granularity refers to the ability of the system to identify entities at various levels of resolution. In terms of the CDM, the levels of resolution can be perceived as being different levels in logical hierarchies. An identifier can be used to identify a legal entity or a natural person. At a finer granularity level, an identifier that identifies a legal entity can refer to either a non-governmental organization or a privately owned firm.

Some aspects of the model, as for example the corporate governance relationships, require that identified economic entities are related to other economic entities. In the case of corporate governance, the relationship between entities is general enough to include both natural persons and private firms. In other cases, as for instance in cases of financial relationships, the relations can be restrictive. Historically, physical persons do not issue stocks. A granular identification system is appropriate to describe a variety of such cases.

Immutability

An identifier is immutable if the label cannot change. In cases of migration to alternative identification systems, the identifiers are preserved without modification. For instance, in the Center for Research in Security Prices (commonly abbreviated as CRSP)/Compustat-merge, the identifiers of both systems are offered for the merged data. This can also be relevant when connecting the identifiers of the existing implementations of the consortium with the common data service.

Intelligent and non-intelligent identification

An intelligent identifier is a label that, besides identification information, carries metadata information of the identified object. In contrast, a non-intelligent identifier is a label that contains no inherent information., other than identifying an object in a given identification system. The identifiers used by the European Financial Data Institute (known as EUROFIDAI) are an example of an intelligent identification design. The Legal Entity Identifiers and Digital Object Identifiers (abbreviated LEI and DOI respectively) are examples of non-intelligent identification. SCOB and DFIH identification systems are also non-intelligent.

Non-intelligent identification is used by many modern identification systems. Its advantage over intelligent identification designs is that its definition is independent of the underlying attributes of the logical entities to be identified. In case that unforeseen events change the attribute values of these entities, the identifier remains unaffected. Although unforeseen events may not be relevant in *EURHISFIRM*'s historical context, a non-intelligent identification system is still advisable for the core of the identifying label. Future technological innovations and expansions of *EURHISFIRM*'s model might include data elements with attributes that are not currently considered in the design.

Meaningful parts in the identifier's standard can be advantageous when it concerns information that is not related to the identified entity. For instance, this form of standardization can facilitate the capacity of the identifier.

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

International identification/interchangeability

An international identification system is not restrained within national or regional boundaries. Its identification capability is potentially global. This property is fundamental for *EURHISFIRM*'s system. An overarching identification system is feasible only if it can specify companies that originate from different European countries.

Furthermore, an international identification system assigns identifiers that are interchangeable across European countries. A particular label identifies the same entity in every national context. If a label is assigned to an entity at a particular country, then this label also validly identifies the entity in every other European country. This property is of particular importance in distributed system designs.

Label validation

This is not a functional requirement, but it is a recommended practice. The property refers to providing the ability to systems and humans to check the validity of the identifier's label. This is typically achieved through one or more check digits. The standard described in ISO/IEC 7064:2003 gives a recommended specification of check systems.

Modularity/Neutrality/Autonomy

The property refers to the attribute of the identifier to be system-independent. An identification system that does not depend on a particular system architecture promotes its sustainability. Identification designs that are built in a modular fashion and are connected with plugin characteristics to the system are more isolated against system changes. Such a modular design facilitates the overall evolvability of the system. For instance, if in the future the consortium decides to migrate from one database technology to another, a modular identification system requires only a few, if any, modifications to be compatible with the new technology.

Persistence

Persistent identifiers have labels that unambiguously specify entities for indefinitely long. This property is to be interpreted within the context of the lifespan of the CDM. As *EURHISFIRM* aims to provide a sustainable, long-lived research infrastructure, the identification system should satisfy the persistence attribute for time-frames that span more than half a century. In any case, the identification labels cannot be reused to specify different entities even in cases where the originally identified objects cease to exist.

Readability

Although that readability is not an inherent identification characteristic, it is relevant in *EURHISFIRM*'s context. An identifier may be presented with different syntaxes. For instance, a five-digit postcode can be syntactically presented, among others, as XX-XXX or XXXXX. The first presentation style is friendlier for human-readability than the second. An identifier has a readable representation if it facilitates human interpretations. A readable representation is useful for *EURHISFIRM*'s end-users.

Conducting research is not always confined to applying statistical methods in groups of data. There are occasions in which information about particular data items need to be examined. In such cases, researchers need to select specific items by inputting the corresponding identifiers into a computer

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

system. A readable representation of the identifier not only facilitates the optical inspection of the identifiers but also reduces the possibility of mistyping.

Another possibility to address readability issues is to provide an abbreviation through which users can easily refer to data items. A motivating example can be found in contemporary stock data. International Security Identification Numbers (shortly known as ISIN) are commonly used to uniquely identify listed stocks at an international level. For humans, it is easier to refer to a stock using its ticker symbol. Although the ticker symbol is not enough to identify a security in an international financial system scope, within a particular market, it is sufficient to pinpoint specific securities.

Reconciliation

An identification system is reconcilable if it provides a mechanism for merging its identified entities with similar entities identified by other identification systems. This property is important for *EURHISFIRM*'s system. For instance, the identified, historical, economic entities are expected to be linked with contemporary ones. Such a feature shall allow researchers to link historical with contemporary data.

Representative administration

The identification system has administrative needs. The nature of *EURHISFIRM* dictates that the administrative aspects of the identification system are of decentralized nature. Technical maintenance can be assigned to a particular research center, however, the administration of the identification-design and is controlled by all the participating members. Participating members should meet the needs and represent all interested parties. The representation can be based on geographical regions. The participating members have responsibilities and roles throughout the lifespan of the identification system. The coordination of the design is performed in international standardization meetings with representatives of all participating members.

The administration of the system comes with operational costs. These, for instance, include the registration, maintenance, and deduplication of identifiers. Necessary funds to cover these administrative costs should be allocated to the research centers that are involved in the process.

Security

An identification system is secure if it is resilient against malicious attacks. In particular, the identifiers of the *EURHISFIRM*'s system should be protected against irreversible corruption. Modifications in the identification data should be only performed under authorized access and recovery copies should be regularly saved.

Scoped identification

This property concerns the scope of the identifiers. The scope of the identification system is defined in terms of the entities it specifies and is delimited by other identification systems that specify other objects. The scope of the CDM is to specify historical economic entities. The most central elements of the scope are the identification of firms, governmental and non-governmental organizations. Natural person identification is also desirable to be included in the scope.

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

A well-defined identification scope has concrete definitions of the classes of the objects to be identified as a prerequisite. The three main classes of entities that are interesting for *EURHISFIRM* are legal entities of official nature (e.g. municipalities), legal entities of non-official nature (e.g. privately owned firms), and physical entities. In particular, for the legal-entity-objects, precise and granular definitions are needed to define the identification's scope. The scope of the identification of economic entities can be facilitated by including an intelligent element that signifies to which of the three broad classes of entities the identified objects belongs. This can also be helpful for readability.

Existing identification systems, for instance, the LEI-based identification design of the Global Legal Entity Identifier (abbreviated GLEIF), already identify legal entities of non-official nature. However it is atemporal, i.e. their scope is limited in the identification of currently operating entities. The CDM's identification scope is intertemporal. The identification of entities does not concern a particular date, but instead historical time-spans.

Determining the scope of identification is a property that has organizational side-requirements. Specifically, the parties involved in the assignment and administration of the identifier and the roles in the accompanying metadata governance must also be specified.

Storability

The identification system is accompanied by storage costs. Appropriate infrastructure for storing the identifiers and the defining metadata is necessary. A comprehensive infrastructure also allocates resources for the storage of identifiers of related identification systems. This is compatible with the linking tables approach of DFIH described in (Karapanagiotis, 2019).

The technical administration of the system's storage can be cost-effectively performed at a centralized, international level. This does not contradict the need for international, organizational administration. The decision making and the steering of the system can still be decentralized. The implementation of the decisions can be performed from a single technical, administrative authority.

Uniqueness

The mapping of identifiers and logical entities is uniquely defined. In other words, every label identifies one and only one logical entity. The same label cannot be associated with two different logical entities. However, two different identification systems may as well point to the same logical entity. For instance, an LEI identifier and a Value Added Tax Identification Number (knowsn as VATIN) can point to the same company.

This property is essential for the CDM, as it is for any data-driven system. It is a necessary property to associate logical entities with their digital manifestations. The minimum requirement is that the CDM implements at least one unique identification system. The usage of multiple parallel identification systems can be beneficial in terms of linking *EURHISFIRM*'s entities with information outside the system's scope. Of course, the more identification systems one supports, the greater are the implementation and maintenance costs.

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

ii. Informational requirements

This section examines the informational requirements for identifying three main classes of economic entities. These classes are natural persons, legal entities of official nature (e.g. local governments) and legal entities of non-official nature.

The Working Group of Identification and Standards (henceforth WGIS)[4] has decided on the meeting of the 17th of October 2019 that identification of persons and official organizations (specifically municipalities, national and local governments) will begin after the standardization of firms. This should not be considered as changing the scope of *EURHISFIRM*. The common data model will not address these aspects of its content the same way as it addresses the companies. According to the aforementioned meeting, the standardization of the common model focuses initially on organizations of non-official nature. The corporate governance component, as well as the non-standardized elements of the official organizations of *EURHISFIRM*, can be implemented using alternative approaches that are not based on standardization of the attributes of the elements. The discussion in section IV.i is relevant to this point.

Natural persons

The identification of physical entities[5] is the hardest identification case among the three classes. Even in a contemporary context, the unique identification of physical entities is complicated. An identification design that stays close to the guidelines of III.i, requires information that is typically hard to obtain even today; e.g. biometric data. Even worse, it is unfeasible to obtain such information for most historical physical entities.

The SCOB implementation has achieved progress with the identification of persons. The identification of persons is performed by historians with expertise on the genealogy of Belgium. The identification is achieved by manually comparing and combining information from various historical sources. In particular, information of persons is obtained by the listed companies' registries (i.e. the *Recueil financier*), the government gazette that records the deeds of incorporation of legal entities and from civil registers.

Typically the information associated with a physical entity that is available in the yearly books of listed companies consists of the name and the occupation of the person that is involved in the firm's management and ownership structure. In some cases, the municipality or the address of residence is also available. This information is not adequate to unambiguously identify a physical entity, in particular for greatly populated countries like Germany and France. *EURHISFIRM*, however, aims to provide its users with data on the corporate governance of firms and therefore completely ignoring this informational aspect is not appropriate. For the moment, the SCOB's approach that involves human expertise in genealogical trees is the most viable if not the only feasible one.

Three informational fields that are commonly available in the printed sources are

1. the first name,

[4] See (Karapanagiotis, 2019, p. 26) for more information on WGIS.
[5] The term physical entity is used to distinguish persons from other organizational entities that have legal nature. This categorization is typically found in legal documents.

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

2. the last name of the person, and
3. her occupation.

Although these fields are not enough to provide unambiguous identification, they, in conjunction with the year that the source of the information was first published, can provide the core fields for establishing preliminary identification. The preliminary results should then be refined by experts. Information on the birthdate and Nation of birth are important for identification; these attributes, however, are not typically available in the standard sources that *EURHISFIRM* uses as input. The use and dissemination of such information can be subject to personal data protection laws for persons that are alive.

Legal entities of non-official nature

This is the most studied class. Previous work on the identification of contemporary firms is indicative of the informational requirements for the identification of entities of this class. For the identification of contemporary entities of this type, it suffices to have information on

1. the name of the legal entity and
2. the registered office address of the entity.

It is also desirable, but not necessary, to have information on the source from which such information is obtained.

EURHISFIRM, however, also has a historical dimension. Legal entities are to be identified not at a single time point but for a long historical period. Therefore, in addition to the two aforementioned informational requirements, one should also add

3. the founding date (incorporation) of the legal entity.

The founding date information is available in historical sources (see (Poukens, 2019c)). The founding date of a legal entity is the date that the legal entity came into existence. For instance, after a corporate action that changes the legal status of a firm, a new data entity is recorded using a separate founding date. The new entity can be connected to the entity that existed before the corporate action. Concerning identification, however, the two entities are separate objects.

The identification designs of the implementations of DFIH and SCOB are based on the names and founding dates of the companies. Other attributes that can be helpful for the identification of firms are the legal form of the firm, its liquidation date, and the industries in which the firm operated. These attributes are typically found in the input-sources of *EURHISFIRM*.

Legal entities of official nature

In some cases, entities of official nature have similarities with those of non-official nature. For instance, one may argue in favor of using constitutional declarations as incorporations dates. The problem is that, firstly, constitutional declarations are not available for all the sub-classes of these entities. For example, municipalities are not typically formed using constitutional declarations. Should the date of the belonging state be cascaded to the local authority? What happens in cases that local authorities move

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

organizationally from one encompassing authority to another one? Should such a situation give rise to a new model object?

These are questions that are more appropriately answered within context. For instance, many of the wars that occurred in Europe during the 18th, 19th and 20th centuries ended with significant changes in national borders. Assigning new, but interconnected, identifiers to entities of official nature that had regime changes is relevant for *EURHISFIRM*. Besides wars, there are cases that, due to some organizational reform, a municipality passed from an encompassing regional administration to another one within the same national entity. However, new identification for such cases is not of the same significance.

This is indicative that, as in the cases of physical entities, human expertise is essential for the common model implementation. Besides technical requirements, *EURHISFIRM*'s identification requires experts in the fields of economics and history that are able to channel the technical support to correctly identified objects. Expertise on these topics can probably benefit from locality specialization, which implies that national research-center-needs should be also considered in the allocation of identification resources.

The DFIH implementation has elements that allow it to incorporate entities of non-official nature in its design. These are found in various places of the corporation table-family. For instance, identifiers are assigned to both states and cities in the corporation table. The identifiers are linked with the `coroporation_name` table. Country names are prefixed by the string 'ETATS : '. For example 'ETATS : FRANCE' and 'ETATS : PRUSSE'. Cities are prefixed by the string 'VILLES : '. One can potentially retrieve information only on cities and states by querying the database for names that start with the aforementioned terms. The same information can be selected even easier by using the `public_status` field of the corporation table. This is a logical field that indicates whether an entity is of official nature.

This form of design is appropriate to support the future needs of DFIH and SCOB. Furthermore, it is compatible with the top-level overview model proposed in (Karapanagiotis, 2019). An element that can be added in this design is that of granularity. A separate storage table for official elements would allow, for instance, the names of the entities to be saved without the prefixes. Moreover, the specialized table of official entities can contain, or be linked to tables with, fields that are only relevant for official entities.

iii. Standards and Governance

The design and maintenance of an identification system have both administrative and technical standardization requirements. In fact, this is true for *EURHISFIRM*'s system as a whole rather than only for the identification system. Moreover, the distributed nature of *EURHISFIRM* essentially mandates an analogously distributed standardization mechanism. To solidify the longevity and generality of the research infrastructure, such a mechanism has to be international and representative. All interested parties should be able to contribute to the specification of standards.

When it comes to identification, in particular, the body of the standardization committee should consist of national representatives of the research infrastructures that are members of the consortium. Such a structure promotes the neutrality of the identification system. The design is commonly agreed to facilitate the needs of all national implementations and it does not focus on a particular one. Since some countries may decide to distribute the responsibilities of the national systems across many research centers, while

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

other countries may decide to centrally build the system in an administrative center, representation at a national level is more appropriate to ensure equality among participating countries.

Such a distributed structure is particularly relevant if a two-level identification system is adopted. Issued identifiers that are compatible with the agreed standard can be assigned regionally. Deduplication and final global identification are then performed at a CDM level. In cases of conflicts, the global identifier assigned by the common system can be sent back to the national centers. In turn, national centers can update their systems.

The global storage and quality assurance of the CDM-identifiers are performed at an international, centralized level. The administration mechanism of the identification system is decentralized. However, technical maintenance, administrative and technical support is the responsibility of the common model and is performed centrally. Centrality can provide more guarantees concerning the international nature of the system. The administrative vehicle that is used for this purpose should also be responsible for providing know-how and know-why to consortium members that are in the process of developing new implementations or adapting national solutions to the common standard.

IV. Modeling topics evaluation

This section discusses a collection of modeling issues that are relevant to the common service design. The first evaluation topic is technical. It examines the appropriateness of the application of two technological frameworks in *EURHISFIRM*. The second topic is a cost-benefit analysis of the adoption of SCOB/DFIH elements to the design of the CDM. The third topic is a derivative of Work Package 4's work. It discusses the adoption of metadata standards by the common model. The fourth topic is fundamental for the common service. It proposes a strategy for and discusses the requirements of linking historical data to contemporary databases.

i. SQL and NoSQL based models

The discussion in this section is not to be perceived as a general comparison of Structured Query Language (known as SQL) and NoSQL (Not Only SQL) based technologies. Such a comparison in an abstract context is not helpful for the scope of the report. The interested reader may consult (Kleppmann, 2017) for a general comparative approach. The points raised in this section are specializations of the points of the last reference. Implementations of models based on different technologies serve better different goals.

There are advantages and disadvantages in both fixed schema and schemaless solutions. *EURHISFIRM* aims to be a world-class research infrastructure and at such a scale it is will mostly benefit by adopting both solutions for different parts of its design. On the one hand, the heterogeneity of information in historical sources, even within a single national jurisdiction, suggests that non-relational solutions are appropriate. This is particularly relevant for countries that develop new implementations and they do not have to support legacy systems. On the other hand, both SCOB and DFIH have currently relational implementations. The common data model should take into account both of these points. This section discusses how the strengths and weaknesses of the two approaches can be utilized by the common access service.

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

There is substantial feature heterogeneity both in SQL and, even more, in NoSQL technologies. For this reason, this discussion is decomposed to particular characteristics rather than general technology groups. One of these characteristics is the flexibility of the model's schema.

Fixed-schema and schemaless designs

The flexibility of the system does not only refer to how information with different data formats can be injected into the system but also on the structure of the model. The structure of the model can be based on a fixed-schema design or on a schemaless design. SQL technologies are fixed-schema based while NoSQL technologies are not based on schema designs. The heterogeneity of data across countries is indicative that a pure fixed schema design might be too rigid for *EURHISFIRM*'s purpose.

Fixed-schema technologies offer predictable data content. They impose the format and relationships within the data elements. Data insertion has to be conformant with the rules of the schema, which implies that when data are sought, the resulting set has a predefined format. Fixed-schema technologies are rigid and less easy to update. Inserting data that do not conform to the original design requires changing the design. Schemaless technologies are descriptive and easily updatable. Depending on the implementation of the technology, they impose less format and content constraints and allow the inclusion of data with variable attributes.

Two CDM design goals are model evolvability and data harmonization. When it comes to choosing between fixed-schema and schemaless technologies these goals point to opposite directions. The need to have an updatable model, that can easily adapt both to unexpected informational content found in currently available sources and to informational content expansions to new sources, points to a schemaless solution. The need to offer harmonized data, that can be easily accessed and used both from scholars and practitioners, points to the direction of fixed-schema technologies.

A potential solution that addresses both the model evolvability and data harmonization needs is to split the database model into two parts, with each part addressing one specific need. The schemaless part can be used to enable seamless model evolvability and the fixed-schema part to hold the harmonized data. Furthermore, the flexible schemaless part can be more front-end, search-oriented, allowing the community around *EURHISFIRM* to introduce new content and validate the existing one. The harmonized part can be more back-end oriented and offered as a solution or service to institutions, organizations, and firms that desire to have access to pre-harmonized historical data.

The immediate issue that is raised is the communication between the two sub-systems. The design has to establish an internal process of transforming the flexible inflow of information coming from the schemaless, front-end system to the fixed-schema, back-end one. If this is not designed in an appropriate manner, there is a danger of ending up with two completely separate data models. The use of compliant identification between the two sub-systems can help avoid this pitfall.

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

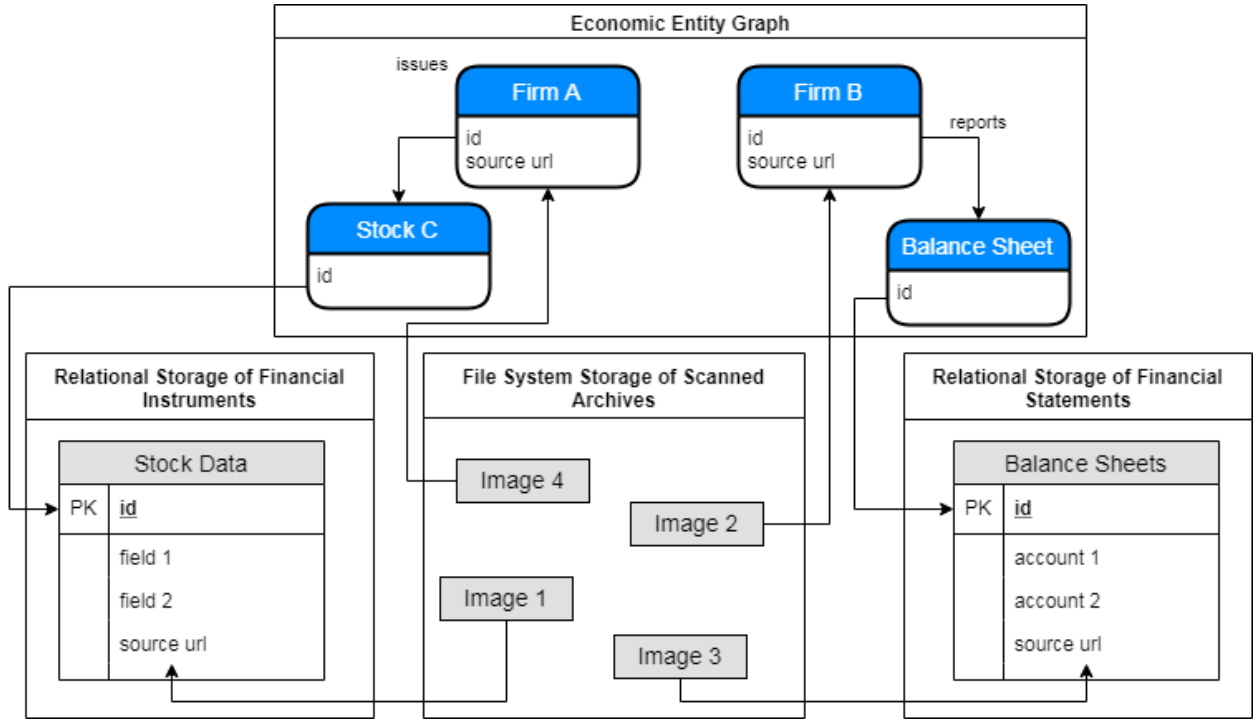

Figure 2. An example of a system with multiple storage technologies

Figure 2 exemplifies how a system with three different storage technologies can be organized. The depicted system uses three different storage technologies. A NoSql graph technology to store high-level entity data, a relational technology to store time series and financial statement data, and a filesystem technology to store images of sources. The graph technology conveniently depicts information about the relationships between the entities. The users are enabled to traverse the data directly based on these relations. The relational technologies give table representations of information that is used in research. Such a representation is typically expected by statistical software. The filesystem storage contains the images of the scanned archives. In combination with a web-server, it gives a viable alternative for providing end-users with scanned files for validation purposes.

User-driven information can be inserted in the back-end design only if it is conformant with the schema and content-wise valid. The first requirement can be checked automatically, but the second needs to be checked manually by *EURHISFIRM* experts.

Storage locality, duplication, and query performance

Every performance statement made in this section is based on commonly acknowledged attributes of the discussed technologies. One should always keep in mind that accurate performance comparison can only be performed by benchmarking. In the implementation phase of *EURHISFIRM*, the statements of this section can serve as a guide of the alternatives that should be examined.

NoSQL technologies typically duplicate data and require more storage. Relational technologies avoid data duplication by splitting data across different tables. The second approach is typically performing better when it comes to updating data, as the update occurs in a single entry. Updating duplicated data requires that the updating operation is performed at multiple database points. Data duplication, however, has a

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

performance advantage concerning read operations. With continuous storage designs, there is no need to join data from different tables that possibly require multiple disk reads.

The relative intensity of read and write operations is the point that determines whether one or the other technology is more appropriate for a system. If a system is read-intensive, then a NoSQL approach should be considered as an implementation alternative. If a system is write-intensive, then a SQL implementation should be also considered. The services that *EURHISFIRM* offers are not time-critical. The time spent in retrieving data from a data provider, such as CRSP, is a small fraction of the total time spent on a research project. The performance considerations in terms of storage locality are not of central importance for *EURHISFIRM* and, therefore, both technological approaches can be employed by the system.

Traversing research relations

EURHISFIRM's system offers micro-level data. Among these, there are data on financial statements and financial instruments. Modern research in asset pricing uses time series of stock returns and factors derived from financial statements to price risks and assess the cross-section predictability of returns (see for instance (Campbell, Giglio, Polk, & Turley, 2018)). Data harmonization and homogeneity are required for such research topics. This indicates that a fixed-schema implementation with predictable types of results sets is appropriate for storing at least a subset of *EURHISFIRM*-provided data.

Besides research in finance, *EURHISFIRM*'s micro-level company data can be used in empirical corporate finance and network analysis. Such research topics sometimes require selecting information stemming from the relations between entities rather than from the attributes of the entities themselves. For instance, one may inquire about the stock returns of the firms for which there were natural persons involved either as managers or shareholders that went to the same college with a natural person that is part of a national government. Such queries can be formulated both in SQL and in graph programming languages, such as Gremlin. The deeper the level of recursion gets, the more awkward becomes the SQL formulation. Graph query languages are more suitable for retrieving information of this nature.

Scaling and performance

EURHISFIRM system has more storage and computational requirements in comparison with the DFIH and SCOB implementations. Instead of storing data for one country, the system has to support data from multiple countries. Moreover, the *EURHISFIRM*'s system will be available to the public and, as a result, will have to accommodate greater request-traffic. Finally, the cultural aspect of *EURHISFIRM* requires storing thousands of image files. To succeed, *EURHISFIRM*'s system has to be scalable.

There are two distinct notions of system scaling. Horizontal scaling refers to using multiple machines to distribute the load. Vertical scaling refers to using a single-machine with greater capabilities. Single-machine-based systems are easier to implement and maintain, however, machines that can support such an approach can be costlier than multiple lower-capability machines. There is no dominant solution concerning scaling.

The appropriateness of a solution depends not only on the system's traffic-size but also on the traffic-characteristics. A system that serves 1000 1kB requests per second has different scaling needs from a

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

system that handles a single 60MB request per minute, despite that both systems serve on average the same size of traffic per minute.

The scope of *EURHISFIRM*'s targeted services dictates that the system is designed to be horizontally scalable. This is a frequently occurring characteristic of modern systems with great storage and computational needs (see (Sutter, 2005)). In terms of *EURHISFIRM*, horizontal scaling can be achieved either through partitioning a single service or by modularizing heterogeneous services.

Besides scaling, modularization is advantageous concerning the evolvability and maintenance of the system (see also (Karapanagiotis, 2019)). A service that can be provided in a modularized manner is the provision of scanned sources. Concerning scaling, this service can benefit from the know-how of the Digital Research Infrastructure for the Arts and Humanities (henceforth DARIAH). Instead of setting up a scanned archive provision system from scratch to store and make publicly available the scanned sources, *EURHISFIRM* can provide this service through DARIAH's infrastructure. Provided that the network addresses of the scanned archives are persistent, the database can store only the network Uniform Resource Locator (shortly known as URL) that is associated with each data-chunk.

When it comes to main data provision services, single-machine data-locality is usually the most-performing solution. Even without considering the cultural aspect, the amount of information that *EURHISFIRM* aims to provide possibly requires scaling across many machines. The design of NoSQL technologies is such that it allows a system to be more easily scalable. Duplicated data can be distributed across different machines. Similar partitioning functionality is implementable in the relational world. Instead of using a single database schema to store the data, one may use multiple connected schemata that are spread across different machines.

ii. Extending current or developing new implementations

A central characteristic of the initial *EURHISFIRM*'s design is the distributive nature of the CDM. The proposed design is minimally intrusive to existing implementations, respects national idiosyncrasies, and gives the ability to national research centers to grow collaboratively but also independently. Towards this orientation, the CDM is initially designed as a supranational structure that coordinates, consists of, and is lead by the national research centers.

The collaboration between the national centers and the harmonization of information at the European level can potentially lead to the convergence of national implementations to a universal, European design in the future. As *EURHISFIRM*'s research infrastructure matures, it gradually addresses more national idiosyncrasies. Such a procedure can create a fertile environment for the centralization of services. At the moment, such a centralized solution is unfeasible. Initially, *EURHISFIRM* adopts a distributive design. This allows the system to be operable and able to provide services faster. In the long run and conditionally on the agreement of the consortium's members, the design can be updated to incorporate more centralized features.

The overall assessment of (Karapanagiotis, 2019) indicates that the national models, specifically DFIH and SCOB, are successful in providing extensible, updatable designs that cover financial, accounting, geographical and corporate governance information of historical economic entities. A distributive

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

EURHISFIRM design benefits by extending the current implementations. It addresses the weaknesses of national implementations by introducing standardized vocabularies for the described concepts.

The national implementations can also benefit from the adoption of these vocabularies in their systems. Although this is suggested, it is not required. National implementations are not controlled by the central design. For participation in the common service, it is required that the designs conform to the common design. The implementation details lie however in the jurisdiction of national centers. A local implementation can choose to adopt the introduced vocabularies or can establish a mapping from the locally implemented metadata model to the common model's metadata concepts. The metadata scheme and communication between the local implementations and the common model is established by the standardization committee.

The adoption of national implementations is limited by the scope-differences between the national and common models. For instance, the common model has to link firms across countries, a feature that does not exist in national models. Therefore, the common service has to develop new solutions in such cases. These solutions can extend the SCOB and DFIH designs.

Undoubtedly, *EURHISFIRM* can benefit from using elements of advanced national implementations. The adoption of design elements from national implementations is not, however, equivalent to complete system replication. As national centers develop their implementations independently, so does *EURHISFIRM*'s hyper-structure. The choice of used technologies, for instance, can be different from that of the national implementations. Moreover, the adoption of national implementations can be fragmented. For example, the common system may choose to adopt the financial statement and financial instrument designs of the national models and use a different approach for depicting relations among economic entities.

iii. Applying or modifying metadata standards

This section distinguishes four levels for which the specification of metadata standards is relevant for *EURHISFIRM*. The first one concerns the printed sources. The second one concerns the community-contributed datasets. The third one concerns the national data. The last one concerns the data of the common model.

A general discussion on metadata standards and how they related to *EURHISFIRM* can be found in (Poukens, 2018). The discussion here assumes familiarity with previous work and does not repeat the introduction of standards. The goal of the section is to assess the most suitable approach concerning the usage of metadata standards in the common model, i.e. the last of the aforementioned units.

Concerning the first level, (Poukens, 2018) proposed the adoption of the Data Documentation Initiative (henceforth DDI) 3.2 for the documentation of the printed sources. The standard was chosen based on its appropriateness to document data sources. Subsequently, the chosen standard is used in (Poukens, 2019a) to produce metadata for a representative sample of sources. When it comes to the second level, (Poukens, 2019b) proposes DDI 2.5 to be used during the upload of datasets from individual researchers of the *EURHISFIRM* community. The earlier version of the standard is chosen due to its simplicity in comparison with the subsequent one. The contribution of datasets is intended to be also performed by researchers

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

that are not affiliated with *EURHISFIRM*. Convoluted documentation requirements in this unit can be discouraging for many researchers. For the third level, (Poukens, 2019b) proposes DDI 3.2 for documenting variables at a national harmonization level. The collection of elements found in version 3.2 is a superset of the collection of elements found in 2.5. Furthermore, version 3.2 allows the possibility of grouping and mapping source variables to variables of the common model.

Concerning the fourth level, adopting different metadata standards would only raise compatibility issues. Following the analysis of Work Package 4, the common model metadata standards should also be documented using DDI 3.2. This will allow the smooth integration of the work of (Poukens, 2019a), as well as the future documentation of printed and contributed sources to the common data model. The documentation of harmonized variables at a national level can be grouped and mapped to variables of the common model.

iv. (Backward) linking to contemporary databases

This section discusses the informational requirements for connecting *EURHISFIRM*'s historical data to contemporary databases. Four study cases are considered. Three of the studied databases, namely EUROFIDAI, Bloomberg, and Refinitiv are mentioned in the proposal. The London Share Price Database (henceforth LSPD) is also examined as it is relevant for the linking exercises of Work Package 6.

The linking process

EURHISFIRM's model is firm-centric[6]. Firm entities are also central in Bloomberg, and Refinitiv' models, so it is natural to inquire about the ability to link between these infrastructures and *EURHISFIRM*. In contrast, the models of LSPD and EUROFIDAI are security-centric. They contain limited information on firm entities and, therefore, linking *EURHISFIRM* with them can be more complicated. Previous work of WP5 in (Karapanagiotis, 2019) is indicative of the linking strategy that can be employed in these cases. The strategy consists of two main axes.

Firstly, linking is essentially reduced to an identification problem. Once firm entities of the two heterogeneous sources have been unambiguously identified, linking between the sources becomes trivial. Adopting an identification system that is compatible with the intersection of information found in two systems can greatly facilitate the linking process. The identification design proposed in section III takes into account this consideration.

Secondly, the linking process is best performed by moving backward in time. The process starts by attempting to link the most recent data available to *EURHISFIRM* with data from the contemporary databases. The proposed strategy goes firstly for the low-hanging fruit. It is an appropriate approach because more recent printed sources are of better quality[7] and the contained information in these sources is more compatible with contemporary standards. Once the linking of recent data is achieved, the process

[6] This does not mean that other instruments, such as commodity derivatives, or macroeconomic data are excluded from the model. The model is primarily built with firm-level data in mind but it includes information that goes beyond firms whenever is relevant.

[7] Some of the physical historical archives are partially destroyed or the have deteriorated. This is reflected in the scanned images and makes the extraction process more challenging.

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

proceeds to the next step. The next step involves linking of data that directly precede the data already linked. This linking step can take advantage of the previously performed linking. An entity of database A linked to an entity of database B for a given year is likely to be linked to the same entity also for the preceding year. The procedure continues inductively by linking entities further into the past, until the time-span covered by one of the databases is exhausted.

A similar backward stepping approach was employed by the merge of CRSP with Compustat (see the (CRSP, 2018)). Moreover, merging CRSP, which is security-centric, with Compustat, which is firm-centric, has many analogies with merging *EURHISFIRM* with LSPD and EUROFIDAI.

Bloomberg

Bloomberg is a financial software and data provider. Its services target mainly providing information to support decision-makers in business, finance, and government. Its platform gives access to low-latency, global financial services. Besides the main services targeting practitioners, Bloomberg offers its platform also to universities for academic reasons. This aspect of Bloomberg features data that for some countries span more than 20 years.

Although from the research infrastructure aspect Bloomberg's time span is limited, the comprehensiveness of the provided information makes it a fine case to be used when it comes to linking with contemporary data. Company information can be obtained by using the DES command in the Bloomberg terminal[8]. The DES command returns a comprehensive set of formatted output in the Bloomberg terminal. Among the information contained in the output, one can find the company's name, a description of the activities of the company, the company's contact information (e.g. the registered address, the corresponding registration date, telephone, and fax numbers and the website), company management profiles, information on the stocks that it issues, information on the stock exchanges in which these stocks are traded, shareholder information, and geographic segmentation of activities in terms of sales.

The output of the DES command is not appropriate for use as input in an automated linking operation. Bloomberg offers, however, data-export functionality. Data can be exported in EXCEL using an add-in provided by Bloomberg. Such a data-format can possibly be more helpful for automating the linking process.

The EXCEL plugin is based on BQL, which can be used to perform calculations directly on Bloomberg servers and reduces the needs concerning data traffic. Besides EXCEL, pulling data using BQL can be performed via one of the available Bloomberg APIs[9].

EUROFIDAI

EUROFIDAI mainly provides information on financial instruments. It also provides some information regarding the issuing company. According to (EUROFIDAI, 2015), the name, the website, the legal form, the domicile, the expiration date, and the expiration reason of the issuing company are available. The

[8] Bloomberg information can only be accessed using specialized software. Examining the details of how this is done is out of the report's scope. The interested reader is referred to (Bloomberg, 2014)
[9] There are C++, Python, and R flavors available to the users.

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

starting dates of available data are not uniform across countries. The companies' addresses are not stored in the database. The earliest starting date is 1977 for France. Germany and UK data are available starting from 1986.

The name and the expiration date are the main attributes based on which the linking process can be based. The domicile can be indicative of which historical, national data the linking process should focus on. The names of the companies are recorded in latin characters without using accents and other special marks. In contrast, the names of the companies are recorded in national languages in the printed archives. The linking process requires either transforming the names found in the sources into plain latin character format, or vice versa. The first option is better as the plain latin character transformation can be more convenient for end-users. *EURHISFIRM* aspires to build a multinational community and there is no guarantee that software that deals with accents and other special marks is present in the systems of the potential end-users.

The legal form is documented in EUROFIDAI-specific categories. If it is to be utilized in the linking process, a mapping between the legal form categories standardized by *EURHISFIRM* and EUROFIDAI is needed. This should also be taken into consideration during the standardization of *EURHISFIRM*'s legal form types. In cases of doubt, the information available on the company website, as well as the expiration reason can be helpful for verifying the validity of the results of any automated linking process.

LSPD

According to (LSPD, 2019), LSPD provides data of companies traded in London's Stock Exchange starting from 1955. LSPD contains the stock exchange code assigned to the issuing companies. The stored identifiers (variable G31) can be used to link the identified firms with entities of company databases like Extel's EXSTAT and MicroEXSTAT. These databases contain financial statement data of British listed and non-listed companies. Their content is the UK analog to the US content of Compustat. The companies' addresses are not stored in the database. In principle, they can be retrieved through the existing link with Extel's EXSTAT and MicroEXSTAT databases.

Company names are recorded in variable N9. The names of issuing companies are also available (variable G33). This field restricted to be at most 36 characters long, therefore the linking process should consider that abbreviated or truncated names might be stored in this variable. Company status changes, such as name changes, are recorded in variables N1 (start date) and N9 (finish date). However, the same variables are used to record alternative events such as allocations and deallocations of SEDOL numbers. Variables N5 and N7 record justifications for the changes and they can be used to clear the ambiguity with respect to which event occurs. The (LSPD, 2019, p.27) contains an illustrative example of how the evolution of companies can be followed in the LSPD database.

LSPD contains information on the date when ordinary shares were first included in the database (variable G7). This variable does not always correspond to the company's founding date. The type of event that corresponds to the date is recorded in variable G8. Variables G9 and G10 contain analogous information for the date that a share was last quoted and the removal reason.

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

Refinitiv

Refinitiv is an infrastructure that provides financial data and services globally. The content, targeted audience, and approach are similar to that of Bloomberg. The main platform of Refinitiv is called Eikon[10] and it comes in two flavors; a Windows desktop application and a web-browser based version. The data provision, however, is performed also via Elektron. Elektron is platform-independent and its API supports feeds, direct to desktop applications, and cloud applications.

Through Eikon, one can find information about the company's name and commonly used abbreviation, the description of the business activity of the company, the issued securities, the initial public offer and the first date of trading. Refinitiv also offers an EXCEL add-in, which can be used to export data to spreadsheets.

Linking overview

This section examined the possibility of linking with four external databases. The considered databases are different in many respects. Two of them have company-centric models (Bloomberg and Refinitiv), while the other two are security-centric (EUROFIDAI and LSPD). Three of them, namely EUROFIDAI, Bloomberg and Refinitiv have contemporary data, while LSPD has historical and contemporary data[11]. Lastly, two of them are predominately industry-oriented (Bloomberg and Refinitiv) while the other two (EUROFIDAI and LSPD) are research-oriented.

The presentation of informational requirements was split into four separate study cases; one for each considered database. The presentation was organized in this manner to facilitate the exposition of linking requirements. It aimed at providing the reader with a starting point of what kind of information can be used in the linking process. This should not be considered to be suggestive of how the linking process is to be organized.

In contrast with the organization of the presentation, an actual linking implementation can be most successful if it is holistic. It should consider information from all four different targets at the same time. For instance, if an entity from *EURHISFIRM* is linked with one from Bloomberg, then the information on the ISIN and the tickers of the issued securities can be acquired. Such information can be informative when linking with EUROFIDAI or LSPD, which are security-centric and have less information on firms in comparison with Bloomberg.

In terms of implementation details, linking across heterogeneous financial databases is an active research topic. Automated methods based on attribute matching are employed in many cases. For instance, (Rodriguez-Lujan & Huerta, 2016) proposes a methodology involving machine learning to join the CRSP/Compustat merged (abbreviated CCM) database with the Institutional Broker's Estimate System (known as IBES). The approach there considers the matching problem both at a schema matching and at

[10] An introductory guided tour on Eikon can be found in (Thomson Reuters, 2018)(Thomson Reuters, 2018)(Thomson Reuters, 2018)(Thomson Reuters, 2018)(Thomson Reuters, 2018)(Thomson Reuters, 2018).

[11] The classification of data to historical or contemporary is relevant to *EURHISFIRM*'s perspective. The concept of historical data, therefore, hear refers to data from dates earlier than the 1980s.

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

an entity matching level. The entity matching methodology is especially relevant for *EURHISFIRM*'s matching processes.

V. Harmonization process

Harmonization of data is central to *EURHISFIRM*. Research at a national level is indispensable as it can facilitate the design of future, and study the effects of past national policies. However, the historical, economic impact of each European country separately is lacking in comparison with that of the US. A collection of non-harmonized, national data is underachieving for *EURHISFIRM*. Harmonization provides researchers in many fields with the opportunity to study new or old questions from an alternative perspective.

This argument can be illustrated using a simple example. The equity premium puzzle is an (ongoing) discussion in the macroeconomics and finance literature. The term is attributed to the inability of a commonly used, representative-agent economic framework to explain the historically observed average premium of stock returns over less risky securities (typically bonds). Early work on the subject can be found already in (Mehra & Prescott, 1985). On average, the annual equity premium is estimated to be around 6,3% when historical US data are used. This estimate is incompatible with reasonable parameterizations of the framework. One line of argumentation towards the resolution of the puzzle is that of selectivity bias of the US historical market data. The US companies were the most successful during the 20th century and the stock valuations reflect this on their returns. From the estimation of the premium in (Dimson, Marsh, & Staunton, 2012), it is evident that equity premia are considerably lower outside the US. Using data from 19 countries, 13 of which are European, the authors estimate equity premia that are less than 4%.

This section examines the harmonization of data stemming from sources of different consortium's countries. The harmonization of data within a particular source is examined in the documents of WP4. The analysis of this section considers five basic cases; namely Belgium, France, Germany, Spain, and the UK. It adds the topic of cross-country firm linking on the topics of the accounting system and implementation advances heterogeneity that are prescribed in the (Riva et al., 2017).

i. Cross-country firm linking

Cross-country firm linking is central to the services intended to be provided by *EURHISFIRM*. This section discusses both a technical and a contextual aspect of linking.

Technically, linking companies across national jurisdictions is more challenging than linking companies within a particular jurisdiction. Within a confined jurisdiction, a firm is typically a single economic entity. Even in occasions that this is not true, all the information concerning subsidiaries are found in national sources, they conform to the same legal rules, and are usually recorded using the same format. Having information at different time-points or from different sources requires matching the firms' identifying characteristics of these observations.

Multinational firms, however, operate under multiple, different national authorities. The organizational patterns of these operations are not uniform. There are cases in which multinational entities acquire local enterprises to support their operations. In other cases, multinational firms from one country establish

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

subsidiaries or branches in other countries. Attributes other than firm names, for instance subsidiary information, are also relevant for matching firms at an international level. Linking across nations can be facilitated by a procedure of automatic attribute-matching of data from different sources. Manual supervision and verification from experts, however, are central to this linking process.

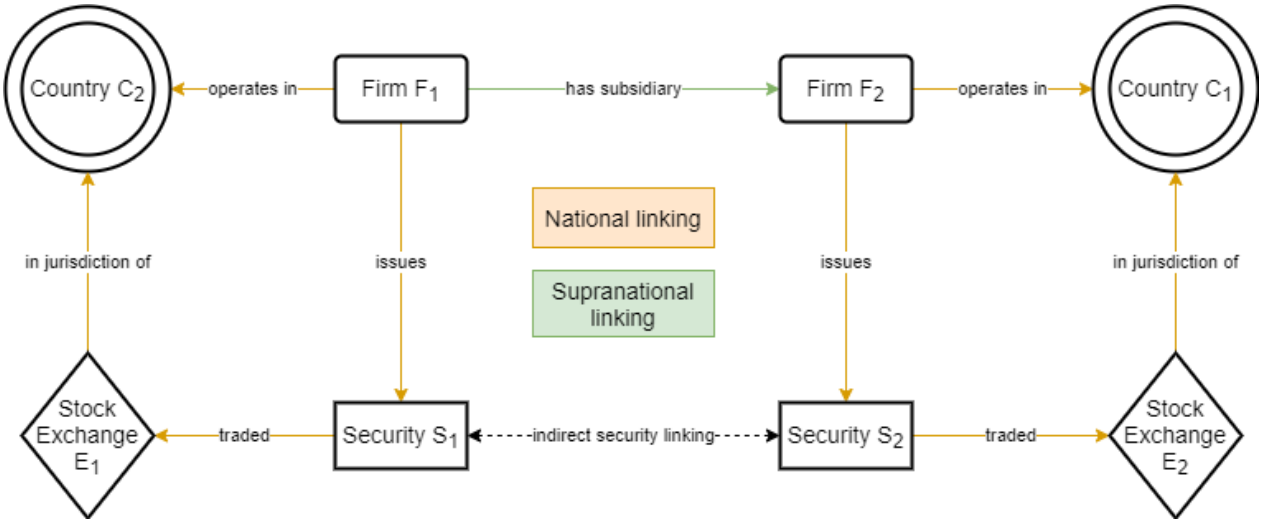

Figure 3. Firm linking at national and international level

Contextually, a successful linking process automatically creates new, multilevel content for *EURHISFIRM*. At an initial level, the established relations between the parent firm and its subsidiaries already add to the informational content of *EURHISFIRM*. With such relations established, linking securities that are traded in different stock exchanges that are located in multiple countries comes for free. This is a second-level content addition of the linking process.

Figure 3 illustrates how the process of linking firm entities can be organized between the national and supranational authorities. Different common model entities are represented by different geometrical shapes. Each national implementation is responsible for linking securities and firms at a national level. Linking firms and securities at a national level is more efficiently performed by regional research centers and experts with specialization in regional information. These relationships are painted orange in Figure 3. Cross country linking is performed at the common service's supranational level. The required expertise at this level is different from that of the regional level. The regional linking requires specialization on the relations between firms and securities. The supranational firm linking requires specialization on international relationships between legal entities. It also requires the standardization of international firm relations. Figure 3 draws this relationship in green. The dashed line of Figure 3 depicts the indirect linkage of securities across countries.

ii. Legal families differences (a double feedback design)

Following the DFIH/SCOB approach in harmonizing accounting information, national implementations (or consolidating hubs) are responsible for mapping historical accounting records to contemporary national ones. The mapping at a national level is performed by experts in historical accounting, possibly supported by information technology methods.

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

Both the original, as well as the resulting mapped contemporary accounts are sent to the supranational *EURHISFIRM* entity. Experts at this level are responsible for consolidating the historical accounts to contemporary, international ones. Firstly, the national mapping of financial statements' accounts can be indicative of the international standard's account to which the historical account corresponds. This benefits the harmonization of accounting data at an international level as the already matched account is used as a prior expectation when matching historical to contemporary international standard's accounts. Secondly, the resulting mapping between historical accounts and contemporary international standards is sent back to the national implementation. It is then the responsibility of the national implementation to choose whether it incorporates this information to its system. In any case, this is also beneficial for national implementations. National centers can use the mapping proposed by the CDM to check if their mapping complies with the common mapping to international standards.

In this respect, the double feedback system acts as a validation mechanism for both the national and international accounting data harmonization. The international linking procedure benefits from the national linking input which is conducted by experts specializing in national accounting systems. The national linking procedure benefits in turn by having feedback from international linking which includes also information originating from other national implementations.

The primary function of the harmonization process concerns creating correspondences between accounts of historical national financial statements with contemporary accounts. This enables researchers to access information from all countries in a uniform manner. It requires, firstly, translating the accounts of national financial statements and, secondly, associating them with the corresponding contemporary international accounts. It does not require that financial statements reported using standalone accounting systems are transformed into granular financial statements of consolidating accounting systems.

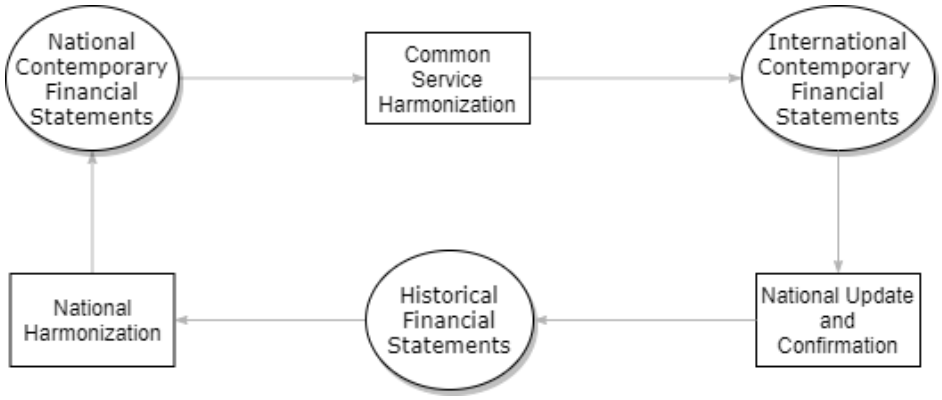

Figure 4. A double feedback design for the harmonization of accounting systems

Figure 4 conceptually describes the double feedback design. The rectangular shapes represent processes. The elliptical shapes represent data items. The historical financial statements belong to the original information layer of the common data model. Both the national and international mappings of financial statements belong to the derived layer of the overall design. The common model stores and provides both the historical and international financial statements. The national research centers may choose to use the output of the common service harmonization process also in their national implementation.

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

European Union's regulations require that the consolidated financial statements of all listed companies are compatible with the International Financial Reporting Standards (henceforth IFRS). The adoption is described in Directive 2013/34/EU. The amendment 2013/34/EU required that the Member States introduce regulation that implements this adoption. All Member States have implemented the directive during 2016. Instead of introducing a separate accounting standardization, *EURHISFIRM* can adopt, and adapt whenever appropriate, the IFRS standard.

Consolidating the historical accounts is not possible in most cases by examining only the statements that are found in the sources. The historical financial statements are typically recorded in standalone accounting systems. Therefore, the harmonization process at an initial stage focuses on mapping standalone national statements to standalone international statements. For this initial process, the standard described in the International Accounting Standards (abbreviated IAS) 27 "Separate Financial Statements" (2011) of IFRS can be used.

The IFRS standards are used by 144 jurisdictions worldwide[12]. Many researchers that investigate the financial statements of these jurisdictions are already familiar with these standards. The adoption of them by *EURHISFIRM* can increase its appeal to such researchers.

When it comes to consolidation, particularly relevant for *EURHISFIRM* design is IFRS 10, which proposes principles for the composition of consolidated financial statements. The standard is based on the principle of control. The control-principle is used to identify whether an investor controls the financial decisions of an investee. This establishes the parent-subsidiary relationships among firms of potentially different jurisdictions. Upon identifying such a relationship, the financial statements of the parent firm consolidate the statements of the subsidiary. As the parent-subsidiary relationship is also relevant to *EURHISFIRM*'s harmonization, the adoption of concepts found in this IFRS standard creates complementarities between the cross-country firm linking and the financial statements harmonization processes.

iii. Implementation advances dimension

Even within the consortium, there are largely-populated, databases with fixed structure (France and Belgium) as well as extended data collections organized in data files. In a distributive system design, establishing communication of the CDM with all the above cases is essential.

For the communication within *EURHISFIRM* to be smooth, it is required to establish a protocol of how communication is conducted. Consider the scenario of sending information from a national level to the common data service. When it comes to the formatting of the transferred data, the responsibility of formatting according to the common data format can belong either to the national hub or the common service.

For the implementations that are based on advanced database systems, extracting the information in particular formats is relatively straightforward. Database systems offer extraction functionality to many

[12] This number is volatile. It was retrieved by https://www.ifrs.org/use-around-the-world/use-of-ifrs-standards-by-jurisdiction/#analysis on 10.10.2019. The exact number is not the main point. It is provided as evidence of the current, widespread adoption of the standards.

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

different portable data formats. For cases of data collections, reformatting from EXCEL or STATA files to the common format can be costly. On the other side, leaving the responsibility of transformations to the common service raises questions about the compatibility of the harmonized data.

To avoid jeopardizing the quality of information that *EURHISFIRM* supplies, it is more efficient to assign the formatting responsibility to the national hubs. Besides fortifying the quality of exchanged information within the consortium, this approach paves the way for *EURHISFIRM*'s future expansions by setting a set of minimum requirements. Research centers and individual researchers that want to participate in *EURHISFIRM*'s common service have to provide data that adhere to these minimum requirements.

The common service can offer some flexibility in terms of the data file formats to accommodate the needs of the current situation in the consortium. The approach of saving data in collections of datasets is commonly used by researchers and it will potentially be encountered also in cases outside of the consortium. For this reason, the common access service should be able to handle input in file formats that are typically used in research, such as EXCEL or STATA data-files.

In terms of content, however, the CDM cannot offer any flexibility. One can participate in the common service only if the minimum requirements of the data standards are met. The data standards of *EURHISFIRM* are commonly decided by the consortium's participating members.

VI. Transformation methods and implementation process

This section contains the discussion of two implementation-oriented topics. The first topic catalogs the transformation methods that are currently used, and methods that can be used in the future, to transform the national data into collections that are compatible with the common data model. The second topic proposes a gradual implementation process for *EURHISFIRM*'s common service.

The standardization of common data model elements is not completed at the moment that this report is written. The transformation methods, therefore, focus on elements that are essential in the first iteration of *EURHISFIRM*'s design.

The discussion of the implementation process provides a comprehensive approach that *EURHISFIRM* can follow in the forthcoming implementation stages. The process is based on the standard that is set by (Riva et al., 2017). It also borrows design principles that are commonly used by private companies that provide software-based services.

i. Transformation methods

The transformation methods are split into four paragraphs. The paragraphs concern identification, financial instruments, financial statements, and corporate governance. The transformation approach is not uniform in all cases. The standardization of common data model elements is performed by the WGIS. Ongoing work of the group focuses on the standardization of financial instrument information. When it comes to identification, the WGIS focuses on the identification of companies. The progress achieved by the WGIS is documented in (WGIS, 2020).

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

This section discusses also how elements of the national implementations that go beyond the scope of financial instruments and firm identification can be mapped to common data model elements. It ties the discussion of section V, on harmonization, and of section IV, on modeling evaluations, with the transformation process. Using international accounting standards facilitates the mapping of national financial statements to the common data model. Adopting a flexible, schemaless design for some elements of the common model facilitates the mapping of corporate governance information to the common model.

Identification

In terms of the top-level concepts of (Karapanagiotis, 2019), the standardization of these elements introduces some sub-classes of the economic entity metadata class. In particular, the attributes of legal entities of non-official nature are defined. The standardization approach groups the attributes of the national sources that were identified by Work Package 4 to common model variables. At the moment of composing this report, the WGIS focuses on the standardization of identifying attributes of legal entities. Subsequent work will expand the focus to other types of entities.

The discussion is split into two steps. The first step concerns the standardization of identifying attributes from an atemporal way. The second step considers the additional attributes that are required to identify the entities through longer periods of time.

Financial instruments

The standardization of financial instruments follows the same bottom-up approach that is used in the standardization of firm-identification elements. The concepts of the sources that were identified by Work Package 4 are grouped and mapped to the common model's concepts. The standardization of financial instruments is scheduled to take place during the WGIS meetings in May and June.

Financial statements

The transformation of national financial statements to the common model's statements is closely related to the harmonization processes of *EURHISFIRM*. In particular, the harmonization of financial statements is discussed in section V.ii. The report proposes a double-feedback mechanism, involving both national and international expertise. The double-feedback mechanism is essentially a methodology for transforming national data to international data.

With respect to the adoption of IFRS standards, there are two cases to be considered. In cases in which the mapped national financial statements are recorded in a way that complies with IFRS standards, the transformation is straightforward in terms of content; it potentially reduces to a simple translation of the accounts. In cases in which the mapped national financial statements are not stored in an IFRS complying system, the transformation of statements is performed by experts with a background in accounting at a common model level. The national implementation can then adopt the established mapping and use it in its design. If it does so, all the subsequent data transmissions regarding the already mapped accounts from a national to the common level fall in the complying IFRS case.

Corporate governance

The identification of data items, in this case, is difficult. Moreover, not all countries have corporate governance elements in their collections. The common data model can promote the collected data in cases

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

that this is feasible and relevant. SCOB and DFIH provide a relational schema for organizing this information. However, this schema is convoluted and it would be inefficient to use for organizing data of non-extensive collections.

The initial service can provide corporate governance information in the form of a collection of documents potentially organized using a NoSQL framework. As the extraction of historical corporate data advances in the consortium's countries and person's identification and linking advances in the common model level, the implementation brings more structured, relational elements into the design. These elements can be based on SCOB and DFIH implementations.

The common format for the communication of corporate governance should be standardized in order to accommodate the common model's automatic consumption of input data. The name, the occupation, the role of a person in the relevant organizational entities, the date, and the title of the source that the information is located is expected to be included in a comprehensive standardization of the content. In particular, for the standardization of occupations, the Historical International Standard of Classification of Occupations (hereafter HISCO) can be taken into considerations[13]. HISCO provides a categorization of 19th and 20th centuries' occupations that are commonly used in economic history research.

ii. Implementation process

Iterative content releases

The implementation process of the CDM does not have to be a one-off project. A more organic approach that gradually releases parts of the common model to the community is a viable alternative. Instead of waiting until a benchmark implementation is ready, *EURHISFIRM* similar to many start-up firms partitions the functionality of its services and releases them asynchronously to the public. The community then can evaluate the performance of the services in terms of approachability and content. Such feedback can be also valuable for subsequent development steps of *EURHISFIRM*. The focus of the common model implementation is based on the research needs of the users.

Drawing from the discussion on the appropriateness of database technologies, the following implementation approach can be used. At the initial implementation iteration, a small portion of the data model is released to the public. For instance, this should contain a sample of core *EURHISFIRM* functionality such as basic firm information, accompanying (not necessarily granular) financial statements and corresponding securities' price data. This part of the model can be implemented using relational methods. Both SCOB and DFIH have already relational implementations that cover this part of the data model.

Data on corporate governance are harder to identify and to link. Delaying the overall implementation for such a feature can introduce a bottleneck in the development of the project and it should be avoided. In the consortium, however, the SCOB implementation has already extensive corporate governance features into its model. The CDM does not have to abstain from the provision of such a feature. Since common service elements such as harmonization, identification, and linking would be difficult to implement, the

[13] The standardization of occupations can be found in (Miles, Leeuwen, & Maas, 2002). I want to thank Johan Poukens for bringing HISCO into my attention.

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

data heterogeneity both within and across countries suggests that this part of the CDM can be implemented using NoSQL technologies.

Transforming Unique Selling Points to services

EURHISFIRM has two fundamental unique selling points. Firstly, it provides long-term, historical, harmonized European, firm-level data. This service has not been implemented by any institution up to today. Secondly, it promotes the common European cultural heritage by digitizing, storing, and promoting historical archives[14]. A target driven implementation of the common model accommodates the successful provision of the aforementioned services.

As the implementation advances, the main two services can be further divided into sub-services. Many research infrastructures, for instance, EUROFIDAI and CRSP, follow such an approach. The provision of research services can be divided into packages and provided under different terms to the users.

The harmonization, consolidation, and linking services require the allocation of resources both in terms of personnel and infrastructure. *EURHISFIRM* can consider providing these services under a subscription to support its financial sustainability. Part of the revenue can potentially be directed to participating national hubs, depending on their contribution to the harmonization, consolidation and linking processes.

Data staging

The technical implementation of the data transformation process is convoluted and, in many cases, does not evolve in a linear fashion. Not all sources available to *EURHISFIRM* reach the common data model stage. Minimum requirements with respect to the data quality[15], as well as verifiability of the content of the source, are central to *EURHISFIRM*.

EURHISFIRM data are staged depending on its origin, properties, compliance with the minimum requirements, and compatibility with the common model. The WGIS considers a transformation process that distinguishes five stages. The specification of the attributes of the stages is ongoing work of Work Package 9. This section relates these stages with the implementation process of the common data model.

Stage 0. Unverified sources: This stage contains data that has yet not been verified and their compliance with the minimum requirements has not been checked. For example, uploaded data from individual researchers that are not affiliated with *EURHISFIRM* initially belong at this stage.

Stage 1. Verified sources: This stage contains verified data. The data can be organized in datasets that contain derived data from individual researchers. The verification should at least ensure that the content of the work is appropriate for *EURHISFIRM*. Datasets from verified community users that want to contribute to *EURHISFIRM* can be potentially staged at this level. The *EURHISFIRM* standardization committee can use this pool of community transformations to expand the standardization of the model towards the direction of community-driven innovations.

[14] Of course, this service is subject to compliance with national and European laws.
[15] See (Poukens, 2019a, Section 5) for a disqusion on the data quality of the sources.

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

Stage 2. Complaint sources: This stage contains data that are both verified (in the sense of stage 1) and compliant with the standard of the common data model. Data stemming from the Optical Character Recognition (henceforth OCR) process of the consortium can be staged at this level.

Stage 3. Nationally consolidated data: This stage contains data that are provided by national research centers that are *EURHISFIRM* members. Membership requires that both the verification of stage 1 and the minimum requirements of stage 2 are fulfilled. For instance, the national implementations of DFIH and SCOB belong at this stage.

Stage 4. EU Consolidated data: This stage is the final data stage that information can reach. Data at this stage are linked, consolidated, and harmonized at a European level. Users that want to access the services of *EURHISFIRM* gain access to the informational content of this stage.

Standardization and organization

How should the future *EURHISFIRM* system be operationally organized? The first step to answer this question is to specify the involved governing bodies and the associated responsibilities. Subsequent work of Work Package 10 should elaborate more on the financial aspect and personnel needs of the organizational structure.

Figure 5 gives an overview of system design for *EURHISFIRM* that adheres to the principle of least intrusiveness. There are two conceptual system levels. At a national level, research centers act independently and are engaged in harvesting and linking information at a national level, as well as in research projects of national interest. All the research centers are members of *EURHISFIRM*'s consortium. At an international level, *EURHISFIRM* provides a common access interface for the information of the consortium's members. The provision of harmonization, identification and linking services can be institutionalized. Researchers outside of the consortium can promote their work and contribute back through their research to *EURHISFIRM*.

There are two crucial elements of the overall system that do not belong in any of the aforementioned conceptual levels. The first one is the OCR system. The second one concerns the provision of scanned archives. In the case of the OCR system, the depiction of Figure 5 takes into account a modularity consideration. In the case of the scanned archives, it takes an efficiency consideration.

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

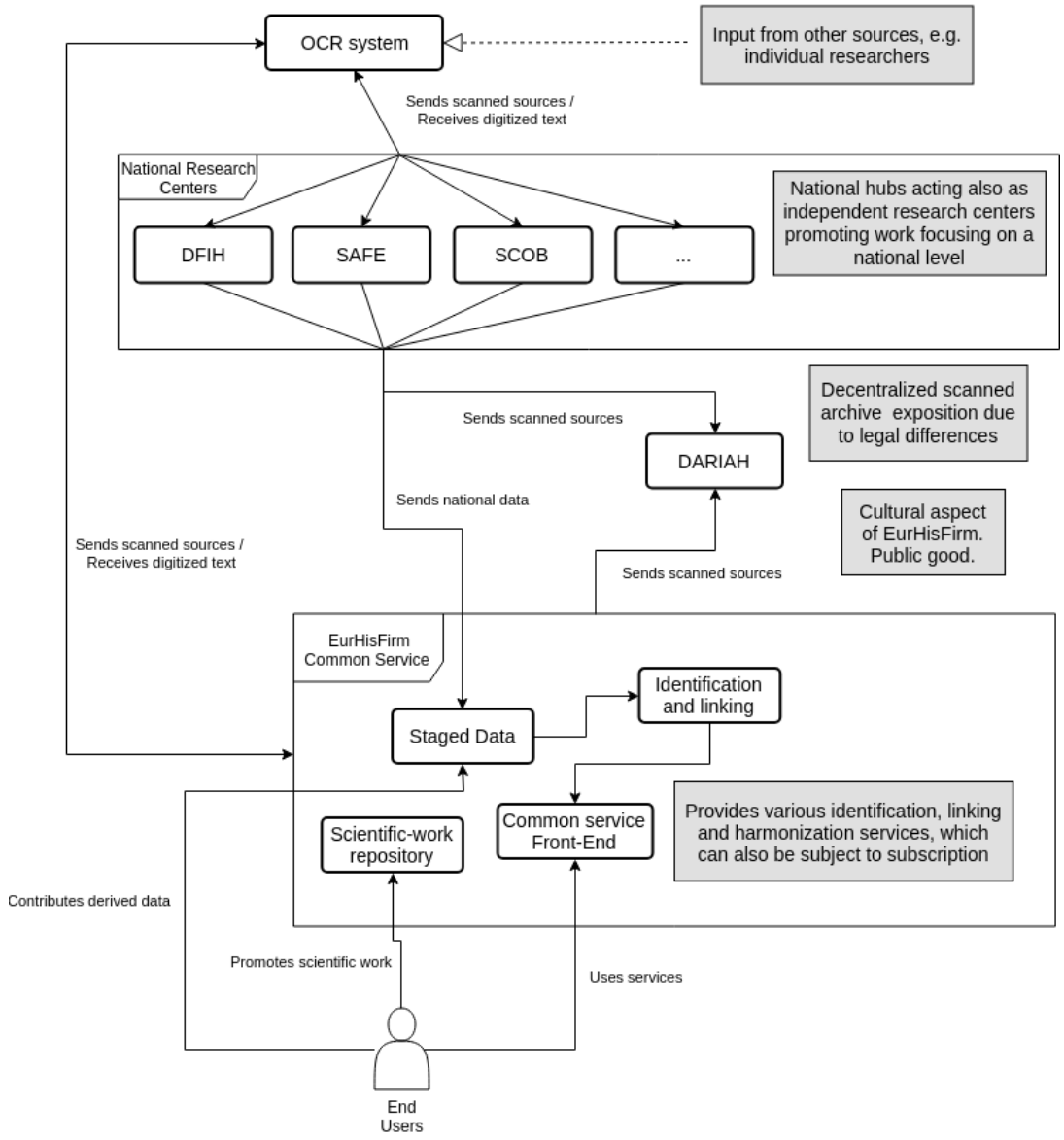

Figure 5. Overview of EURHISFIRM system's structure.

The OCR system can be accessed directly from national centers, as well as from the supranational common structure. It is placed as an independent entity outside the common service. An alternative approach would be to have it a distinct institutional body of the common service. The first approach is more flexible in terms of allowing the OCR system to provide services with a different scope than that of *EURHISFIRM*.

The provision of scanned archives is on its own a demanding task. The cultural aspect of *EURHISFIRM* requires the development of an infrastructure that can support the storage and provision to the public of thousands of scanned files. Moreover, the provision of cultural services typically goes beyond the mere provision of images. It involves the collection and supply of metadata information about the exhibited objects. In both of these directions, *EURHISFIRM* can benefit be the expertise developed in DARIAH.

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

The distributive design of *EURHISFIRM* intensifies the standardization needs. The common standard ensures that the information provided by different members of the consortium, either directly or through the common service, are compatible. It reduces the operational costs of communication and allows experts from different national research centers to conduct research with comparable input data. Last but not least, the common standard acts as a minimum requirement level that the data models of current and future participating members of the consortium should meet.

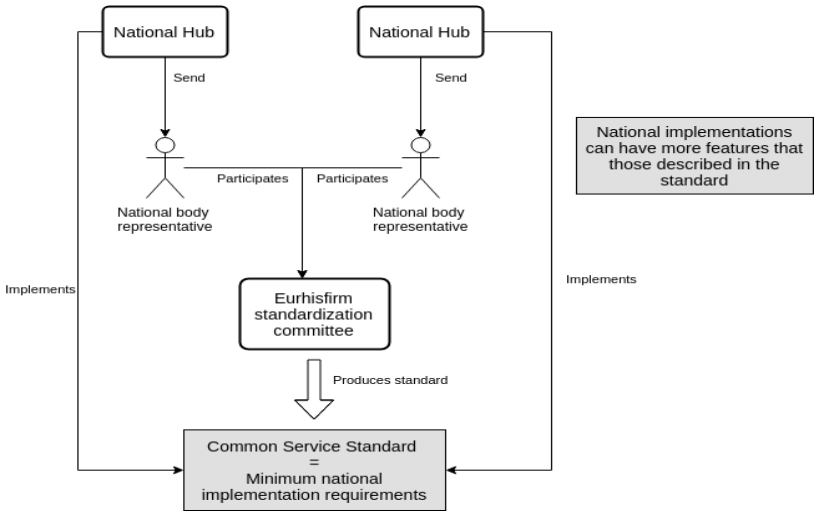

Figure 6. Developing common standards in a distributive organizational structure.

Figure 6 shows how the principle of least intrusiveness can be combined with a distributive organizational structure. A National hub consists of one or more national research centers. To ensure that the standardization process is representative, each hub sends a national representative. The representatives formulate the standardization committee of *EURHISFIRM*. The committee decides on the development of *EURHISFIRM*'s standards. The common standards set the minimum national implementation requirements and are implemented by the national centers. The supranational part of *EURHISFIRM* may support the national implementations. Besides adhering to the minimum requirements, the national centers can independently incorporate additional features into their models.

The initial organizational steps for the implementation of the common service are

1. The consortium agrees on a minimalistic part of the model that is to be initially implemented. In any case, this involves basic firm information. The consortium agrees on the context of the model. For instance, only firms from the class of economic entities can be considered at the first implementation step. The consortium agrees on which information should be initially included in the model in a less fixed-structure manner. For example, corporate governance elements can be included here. The choices of this step should be made by targeting research in particular research areas.

2. The implementation of the first iteration is launched. The finished product is released to the community. The consortium supports research programs of the previously targeted research areas and promotes data usage from the community. Based on the interaction with the users, feedback on missing elements of the model that the researchers find important is obtained.

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

3. The consortium decides how to expand the model and its content. Whenever relevant, the extensions can be performed either by incorporating parts of the non-relational design into the relational one. The user's feedback is a valuable device for orienting future extensions. The feasibility of archive digitization is also an important steering factor.

In order for the above recursive implementation-plan to be feasible, the consortium decides how the communication between the supranational common data layer and the national research centers occurs. In particular, once the elements that are included in the first implementation-iteration are agreed upon, the consortium has to specify the communication protocol.

The standardization of communication has at least two basic elements. Firstly, how communicated data are formatted. In particular, if there is some form of automatization of the interactions between the national and the CDM layer, the file formats should be agreed upon so the necessary software infrastructure is developed in the receiving counterparts. A typical communication network file format is JSON. Secondly, the content of the communicated data should be precisely standardized. In order for any harmonization at a higher level to be meaningful, the national data that are sent should measure exactly the same thing. If closing prices is among the agreed initial common set of variables, the specification should describe what happens, for instance, in cases that a price is missing from the time series. In cases that one implementation decides to impute the price and another one decides to leave it empty, the resulting overall harmonization would be misleading.

VII. Conclusion

The second report of WP5 concludes the examination of the back-end design elements of the common model. The current report extended the analysis of previous work of Work Packages 4 and 5. Based on the insights of the previous reports and the requirements of the proposal, this report gives a series of approaches that can be employed in the implementation of the common data model.

The report selectively reviewed the formatting of elements that are commonly found in the data collections of consortiums' countries. The review does not only examines the domain of the formats but also how parts of composite elements relate to each other.

The report examined the functional and informational requirements of an overarching identification system. It also considered requirements with respect to the standardization and governance of the *EURHISFIRM*'s system. It concluded with a proposal that consolidates the principle of least intrusiveness of the distributive system design with the need to develop common standards.

In technical terms, the report evaluated when it is appropriate to use relational and non-relational technologies. It concluded that different parts of the system can be implemented using different technological solutions and be joined through a technology-independent identification design. The report also evaluated the trade-off between extending current implementations and developing new ones, and the application of metadata standards. It also examined the requirements for linking historical data with a variety of contemporary databases and proposed a backward stepping linking approach.

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

Moreover, the report added details on how the harmonization across national data can be achieved. Specifically, these details concern the cross-country firm linking, the accounting system harmonization, and the convergence of national implementations. A double feedback design that involves experts both at a national and an international level was proposed.

Finally, the report discussed methods and procedures that transform national data sources into elements that are compatible with the common data model. It concluded with the introduction of an overview of the organization and the implementation process that can be used in *EURHISFIRM* system's development.

References

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

Bloomberg. (2014). *Getting Started Guide*.

Campbell, J. Y., Giglio, S., Polk, C., & Turley, R. (2018). An intertemporal CAPM with stochastic volatility. *Journal of Financial Economics*, *128*(2), 207–233. https://doi.org/10.1016/j.jfineco.2018.02.011

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

Dimson, E., Marsh, P., & Staunton, M. (2012). Equity Premia Around the World. *SSRN Electronic Journal*, *2011*(July), 1–19. https://doi.org/10.2139/ssrn.1940165

EUROFIDAI. (2015). *Data Description Guide: Daily Database*.

Karapanagiotis, P. (2019). *Technical Document on National Data Models*. https://doi.org/10.5281/zenodo.3467926

Kleppmann, M. (2017). *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. Retrieved from https://www.oreilly.com/library/view/designing-data-intensive-applications/9781491903063/%0Ahttp://shop.oreilly.com/product/0636920032175.do

LSPD. (2019). *London Share Price: lspm201812 & lspd201812 Reference Manual* (Vol. 0).

Mehra, R., & Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, *15*(2), 145–161. https://doi.org/10.1016/0304-3932(85)90061-3

Miles, A., Leeuwen, M. van, & Maas, I. (2002). *HISCO. Historical International Standard Classification of Occupations*. Belgium: Leuven University Press.

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. (2018). *Information system and documentation standards*. https://doi.org/10.5281/ZENODO.3246455

Poukens, J. (2019a). *Report on data and sources documentation and quality assessment*. https://doi.org/10.5281/ZENODO.3246465

Poukens, J. (2019b). *Report on EURHISFIRM documentation standard*.

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

Poukens, J. (2019c). *Report on the semantics of data and sources*. https://doi.org/10.5281/ZENODO.3246463

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

Rodriguez-Lujan, I., & Huerta, R. (2016). An Algorithm for Matching Heterogeneous Financial Databases: A Case Study for COMPUSTAT/CRSP and I/B/E/S Databases. *Applied Economics and Finance*, *3*(1), 161–172. https://doi.org/10.2139/ssrn.2456793

Sutter, H. (2005). The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software. *Dr. Dobb's Journal*, *30*(3), 202–210. Retrieved from http://www.gotw.ca/publications/concurrency-ddj.htm

Thomson Reuters. (2018). *Starter's guide – Eikon*.

WGIS. (2020). *EURHISFIRM Common Data Model* (Work in progress. Version 1.02 (06.02.2020)).

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