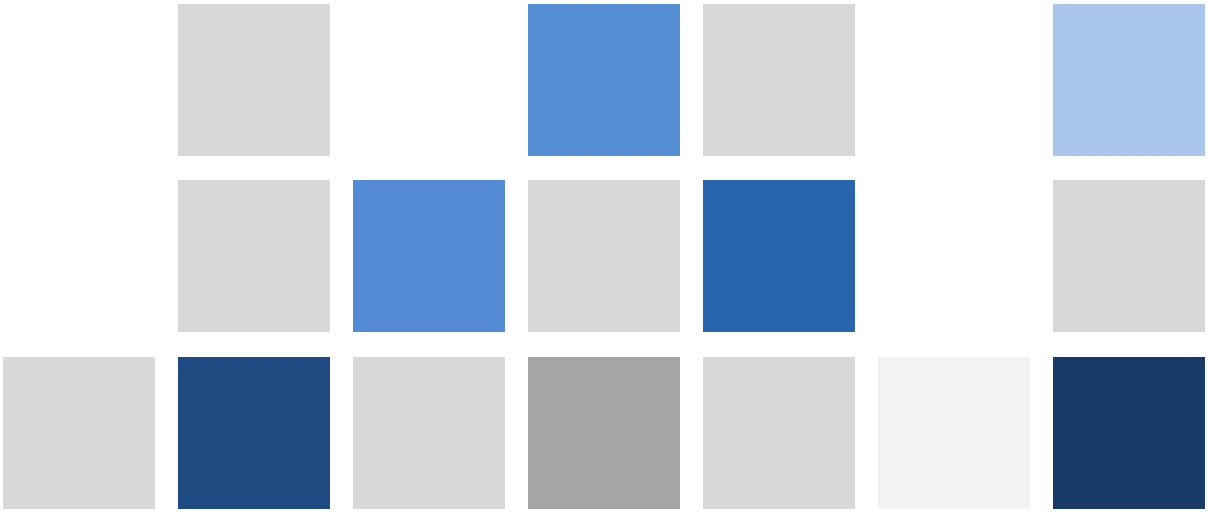
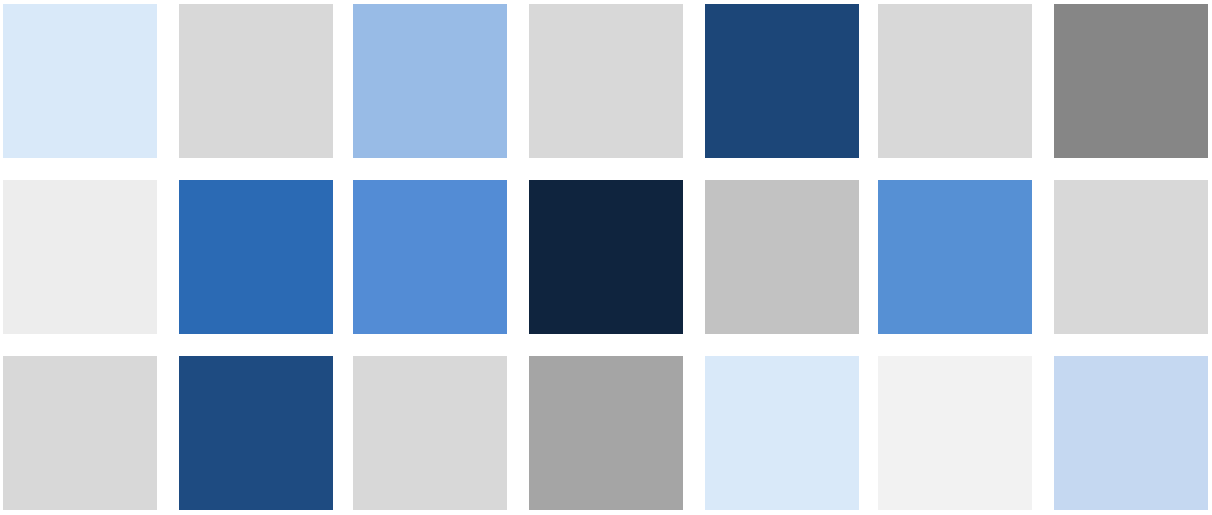


Long-term data for Europe

EURHISFIRM

D5.4: Updated technical document on the preliminary data model



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

AUTHOR(S):

Lukas Manuel Ranft (Goethe University)[1]

Pantelis Karapanagiotis (Goethe University)

Jefferson Braswell (Tahoe Blue)

Wolfgang König (Goethe University)

ABSTRACT:

The fourth report of Work Package 5 provides the latest revisions of the Common Data Model standard specifications. The different foundational elements of the Common Data Model are presented and explained. The report also summarises the results of stakeholder feedback and describes their implications on the Common Data Model. Finally, we give an outlook on the further development of the Common Data Model and its components.

APPROVED IN 2021 BY:

Jan Annaert (Universiteit Antwerpen)

Wolfgang König (Goethe Universität Frankfurt)

Angelo Riva (École d'Économie de Paris)

[1] We would like to thank the centre for Sustainable Architecture in Europe (SAFE) for their contributions to the report and support during the project. Furthermore, I would like to thank the Working Group of Identification and Standards (WGIS) for the discussions and constructive cooperation.

Table of Contents

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

List of Figures

List of Tables

Abbreviations

A	Data Admin
API	Application Programming Interface
C	Data Consumer
CDAS	Common Data Access Service
CDM	Common Data Model
CDMG	CDM Compliant Gateway
CD	Continuous Deployment
CI	Continuous Integration
D	Data Standards
DCU	Data Collection Unit
DSU	Data Submission Unit
EFII	EURHISFIRM Financial Instrument Identifier
ELEI	EURHISFIRM Legal Entity Identifier
ELF	Entity Legal Form
FAIR	Findable, Accessible, Interoperable, and Reusable
G	Data Gateway
GLEIF	Global Legal Entity Identifier Foundation

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

ISIN	International Securities Identification Number
ISO	International Organisation of Standardization
L	Legacy Database
LEDA	Legal Entity Data Artifact
NCA	National Competent Authority
NIC	Network Integration Centre
OMG	Object Management Group
RI	Research Infrastructure
S	Legacy Source
SME	Subject Matter Expert
T	Technology Standards
WGIS	Working Group of Identification and Standards
WP	Work Package

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

1 Introduction

The European data sets for historical financial and firm data are heterogeneous due to differences in the individual countries' regulations, languages, and currencies (Karapanagiotis, 2019). As a result, little progress is achieved in research using harmonised datasets with clearly defined standardised concepts. EURHISFIRM bridges this gap by proposing a flexible common standard set that brings data from different countries together. Via collaborative processes taking place in inclusive fora, such as the WGIS, the national EURHISFIRM institutes, and the consortium countries agree on flexible standards, allowing users to analyse such heterogeneous European data sources. The results of these processes are documented in an extendable common data model (CDM) with, amongst other means, uniform cross-country identification schemes for legal entities, securities, and markets.

The data sources of EURHISFIRM are, albeit historical, not yet completely digitised and processed. As the retrieval of the data has not been finalised, new requirements and modelling needs may arise in the future. Thus, the CDM and the respective common standards need to be adaptable to potential future developments. This report describes how the CDM of EURHISFIRM achieves this goal. The report uses the feedback from various stakeholders, which was collected by the confidential report D5.3, and updates the initial CDM by reflecting the criticisms raised and enhancements proposed in their answers to the questions of the survey.

Readers shall be aware that the pandemic crisis has blurred the clear-cut distinction between the preliminary CDM and its update. The pandemic outbreak resulted in a six-month delay of the second EURHISFIRM annual meeting, where the CDM had to be put on a broad discussion. Nevertheless, during this half a year of delay, based on a sequence of 25 WGIS Zoom-meetings, the initial CDM was already further developed – 'at the cost' of some obliteration of the dividing line between the preliminary CDM, intermediate update, and second update.

The remaining report is organised as follows. In section 2, we shortly review the essential elements of the initial CDM, as was their status before the stakeholder feedback. In section 3, we present the consolidated results of the stakeholder survey carried out in the scope of D5.3. We also elaborate on how this feedback relates to various elements of the prior state of the CDM. Section 4 provides initial ideas on how the EURHISFIRM project team may be evolving the CDM after the report and the end of the project. Section 5 summarises the report.

2 Preliminary Common Data Model

The design goals of EURHISFIRM give rise to standardisation requirements and solutions in four crucial parts of the infrastructure: the federated system architecture, the stages of processing of data from raw sources to harmonised end-user content, the overarching semantic equivalence among data persistence implementations, and the core data model for enhanced interoperability and easy data consumption by end-users. This section reiterates the preliminary concepts that the CDM previously introduced to meet these standardisation requirements.

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

The common standards were introduced, discussed, and agreed upon during the meetings of the Working Group of Identification and Standards (WGIS), which is EURHISFIRM's open standardisation committee consisting of experts from all work packages and external consultants, comprising both groups, operative staff members and leaders of work packages. After some tutorial sessions, the WGIS met every 14 days over 18 months to discuss and, when necessary, to revise the CDM's standards. In this way, the CDM was developed incrementally within the project's framework with the supplementary goal of enhancing communication between the project's work-packages. The committee produced extensible standards that can be quickly revised and adapted to future needs during the ongoing process.

2.1 Federated Architecture

The federated system architecture, which was already hinted in EURHISFIRM's proposal to the European Commission, was identified as the most suitable solution during the early project's meetings. Figure 1 offers an overview of the proposed architecture. The federation comprises actors who share rights and agree to common content standards and communication standards. The federation of the underlying system, with functions distributed across a National Competence Centres network, is invisible to the European data model end-user. As Figure 1 visualises, end-users access the common data via the Common Data Access Service (CDAS), but not the underlying gateways. This leads to enhanced usability for the users but requires the data to be transformed from raw sources and national standards to the common standard.

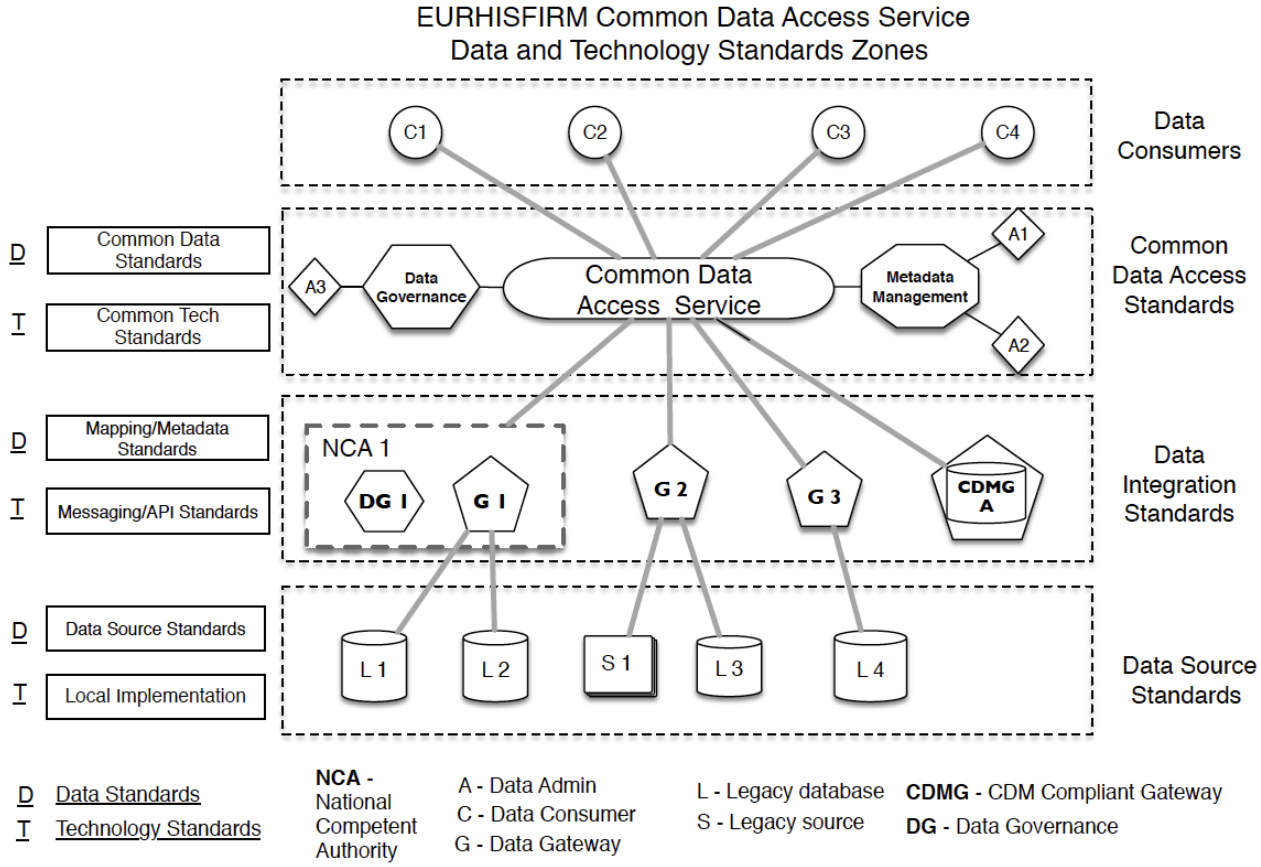


Figure 1: EURHISFIRM Common Data Access Service Data and Technology Standards Zones

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

Figure 1 was the first iteration of the federated architecture, which, after discussions and developments during the project's run, was revised and its elements were refined (compare with Figure 2). The subsequent discussion explains the various elements of Figure 1 and highlights the points of revision.

A Legacy Database (L) is a currently existing database containing relevant data in the scope of EURHISIFRM (see, for instance, the French D-FIH database, https://dfih.fr). For example, L1 and L2 in Figure 1 represent data that are digitised but non-harmonised to a format compatible with the CDM. A Legacy Source (S) is – in the EURHISFIRM system – e.g., a scan of a newspaper or a spreadsheet (potentially also converted to text-data based on OCR software). Legacy sources are also not yet harmonised for the CDM. In the initial design, the task of Data Gateways (G) was to access local data and transform them into common standards dynamically. In 2.2, the dynamic access of gateways to sources was replaced by a multi-stage process that gradually transforms the data between various concepts and formats. National Competent Authorities (NCA) are organisations that contribute data to EURHISFIRM's infrastructure. Each NCA is responsible for contributing data that are complying with the CDM standards but has its own, independent Data Governance (DG) structure and, potentially, data model. A CDM Compliant Gateway (CDMG) is an organisation with a database that complies with the CDM specifications. Consumers (C) are users that interact with the system in various forms, via e. g. a web browser or an application programming interface (API) connection. Data Admins (A) are individuals, groups, or organisations that perform metadata management or data governance functions. The Common Data Access Service (CDAS) is the service that provides a single, centralised point of entry and facilitates user requests to access data in the EURHISFIRM federated architecture. Metadata Management is the process by which a data model is maintained and enhanced, and Data Governance (DG) is the process by which data quality is ensured (Bernstein, 2003, Khatri and Brown, 2010). Both functions occur in a EURHISFIRM network of data governance and metadata management units.

Figure 1 represented the original conceptualisation of the federated system architecture. Subsequent work on the design of the federalisation refined the concept of gateways and specialised them to a sequenced multi-stage transformation process described in Figure 2. Figure 1 identifies two primary types of standards: data standards and technology standards (depicted on the left-hand side). The CDM is addressing the need for data standards. The need for technology standards[2] is being addressed by designing functional and operational system governance processes of Work Package 9 (to be published on https://eurhisfirm.eu/index.php/publications/). The layers of Figure 1 describe different types of standards (depicted on the right-hand-side). The content-standards at the top layer (CDAS) and the communication standards between this layer and the 'data-integration standards' layer – likewise the communication standards between CDAS and users – concern the CDM. The content standards at the 'data source standards' layer and the communication standards between this layer and the 'data integration standards' concern the conventions for acquiring and cataloguing source data in a form that will subsequently allow the data to be harmonised and integrated with the CDM.

[2] "Technology standards" in this context, and specifically in relation to WP9 work on system process integration, addresses the means by which information is stored (persisted), transmitted (communicated between stages), and consistently accessed (queried, navigated) by ultimate end-users.

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

The CDAS and date source standards layer standards are interoperable. As users interact with the CDAS, they may find mistakes or gaps in the data provided and propose corrections or contribute additional data. During such revisions, the end-user is enabled to drill down to the scanned sources to validate their proposed feedback and contributions. To this end, scanned sources are preserved, and mechanisms for accessing them are provided.

2.2 Data Staging

The gateway concept of Figure 1 was refined to a multi-stage process described in Report D5.2 (Karapanagiotis, 2020). This revision accomplished the objectives of transformation and integration of sources into the CDM representation.

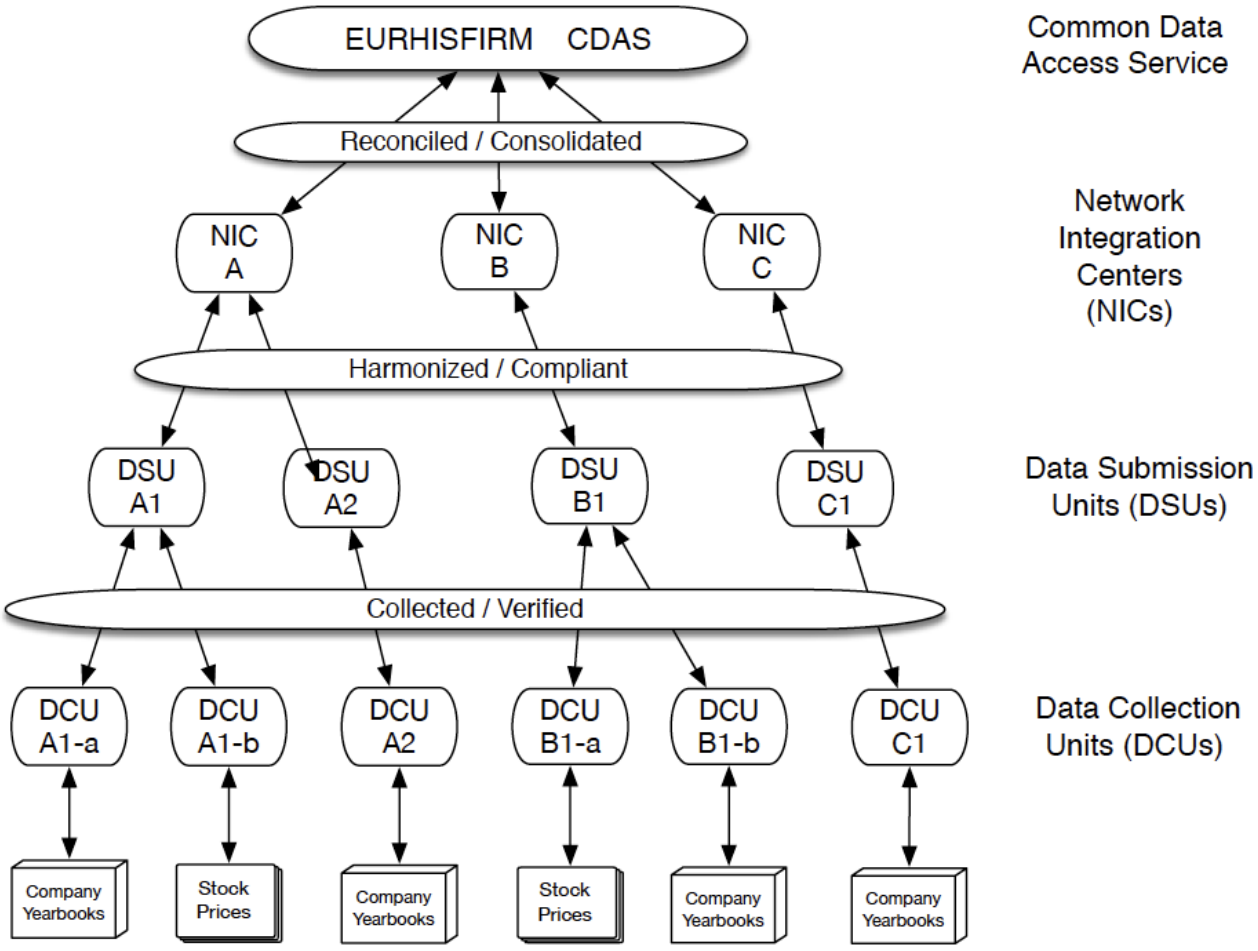


Figure 2: EURHISFIRM Data Staging

Figure 2 describes the process stages that transform the raw data (bottom layer) to CDM-compliant data (top layer). Data Collection Units (DCUs) are organisational units that are responsible for collecting data. The Data Submission Units (DSUs) have the objective of harmonising the data formats, semantics, and labels of the data collected by the DCUs. Additionally, DSUs are responsible for submitting harmonised data to the Network Integration Centres (NICs). At the DCU-level, harmonisation means transforming the raw data using the same metadata description as tags, labels, and field formats. The NICs process the data

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

submitted by DSUs to provide common identification of, for instance, legal entities and financial instruments. The NICs then integrate data and provide it to end-users via the CDAS.

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

Besides producing CDM-compliant data, this process does not exclude users from examining intermediate, staged data. Each transformation step is transparent, and, similar to source data and CDM data, the staged data are FAIR (Tochtermann and Loebbecke, 2018). The harmonisation process between DCU and DSU is consistent with the Data Documentation Initiative (DDI) metadata definitions of variables for local data and conceptual variables for harmonised data. Local language labels for financial terms complement common terms for the same data elements during the transformation steps, but links to the original labels are preserved.

2.3 Overarching Semantic Equivalence

Different platforms and technologies are most appropriate to be used at various stages of a data collection, harmonisation, identification, and end-user common access (Karapanagiotis, 2019). In the light of such functional requirements, the different layers that exist in EURHISFIRM – as well as the data formats that would be used to transfer data between stages – request an overarching semantic approach that enables consistent transformations of data between several persistent data storage platforms (i.e., relational, object-oriented, and graph-databases).

The overarching semantic model ensures that *semantic equivalence* of financial and firm data is maintained between the different stages and organisational units. This approach allows for the interaction and integration of data on collaborating platforms in a distributed, federated architecture that can be a hybrid of technological platforms best suited for use in different system areas. Figure 3 shows how the semantic equivalence can enhance the compatibility between different technologies and sources. Historical data are collected and propagated through the system among actors with potentially different access levels using different storage technologies.

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

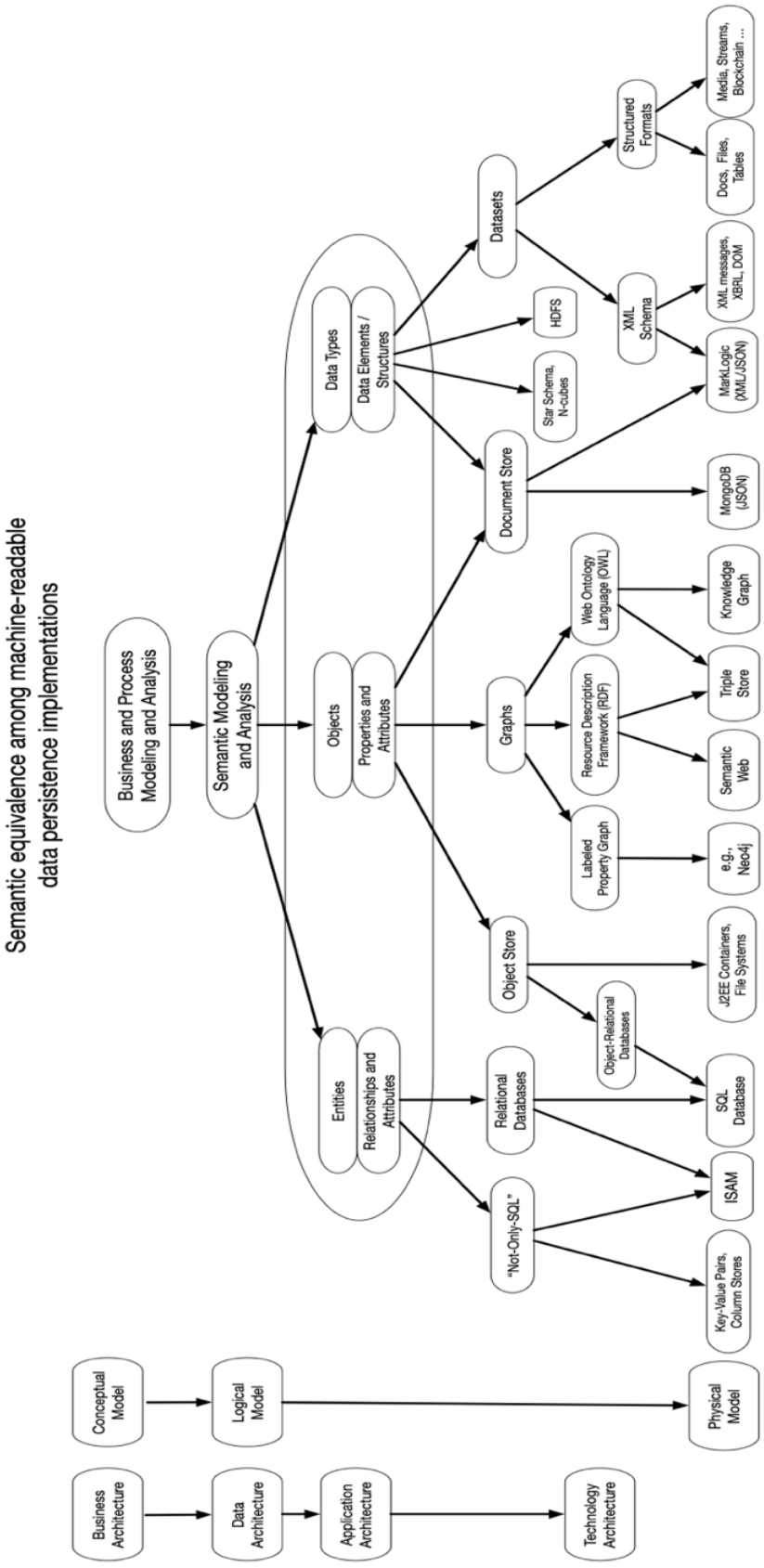


Figure 3: Overarching Semantic Equivalence

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

In the centre of Figure 3 is a common description of the variables and the common formats.

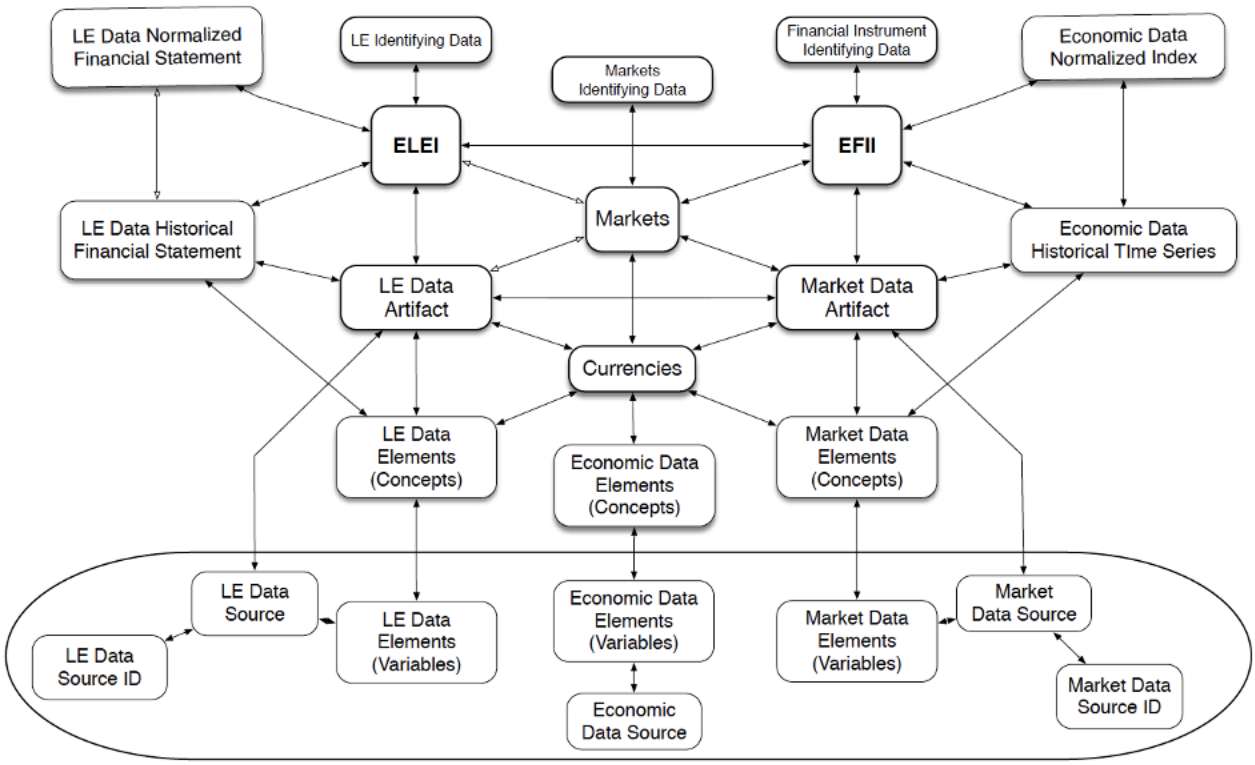
On the bottom are the different possible technologies in which source data is stored. Depending on the type, they can be grouped into more general classes such as relational data, object data, or graph data. In the federated architecture, the needs for different data descriptions exist – as defined in the widely accepted standards methodology Enterprise Architecture – on the design levels of business, data, application, and technology.

2.4 Core Data Model of Firms, Securities, and Markets

Content-wise, the core of the data model, which the future infrastructure of EURHISFIRM concerns, is the "financial realm" of corporations (as an initial legal entity type). The data model's core covers public firms' financial statement data (for instance as contained in the profit and loss statements) and, for instance, for listed companies, the financial instrument data (for example, stock data). The financial instrument data require the modelling of the stock exchange markets with, potentially multiple, trade currencies as attributes. These three modelling elements constitute the central components of Figure 4. Besides the initial classes modelled in the core data structure, subsequent incremental extensions of the data model can add additional model classes. For instance, natural persons and additional financial instruments can be added.

A common identification of the core entities enhances the common access of end-users to data. For this reason, identification standards for legal entities, financial Instruments, and markets are introduced.

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



LE	Legal Entity
ELEI	EURHISFIRM Legal Entity Identifier
EFII	EURHISFIRM Financial Instrument Identifier

Figure 4: EURHISFIRM Core Data Model

Specifications that elaborate on this semantic model's aspects have been developed using a widely accepted standards development methodology (Enterprise Architecture) by a standards development team in EURHISFRM (the WGIS). The development of these specifications evolved throughout the project by leveraging specific existing industry standards that pertain to the identification of organisational entities and financial instruments. The ELEI (EURHISFIRM Legal Entity Identifier) is derived from the reference data standards for the LEI (Legal Entity Identifier) as published by the Global Legal Entity Identifier Foundation (GLEIF). The ELEI extends the reference data of the LEI by enabling historical snapshots of changes in entity reference data; an attribute that is not present in the originating GLEIF standards. The EFII (EURHISFIRM Financial Instrument Identifier) is derived from the Financial Instrument Global Identifier (FIGI) standard of the Object Management Group (OMG). The FIGI standard provides an identification scheme for securities at three different hierarchical levels: trading venue, sovereign jurisdiction, and the issuing corporation.

The ELEI and EFII identifiers establish the foundation for organising and integrating data from various sources related to the two fundamental classes of corporations and securities. The double arrows in Figure 4 indicate how these entities relate to the data model's other data elements. The additional top layer components, the top left and top right units, are showing the fact that several types of transformations of historical data are typically performed that produce datasets with normalised financial data numerical values (e.g., monetary amounts, market prices). These transformations are performed in order to

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

consistently analyse or interpret historical data that have been normalised in the context of the timeframe of the current analysis.

Below the two core entities, there are the two corresponding artefact classes. The artefact classes offer a flexible mechanism to collect data samples in containers that record data using, in the simplest case, lists of key-value pairs. This approach can be logically extended to include self-defining data in the form of JavaScript Object Notation (JSON). In both cases, it is the ability to add the labels associated with the data to a metadata management model that defines the meaning of the labels essential to access and interpret the data in a consistent fashion. The idea of such an intermediate abstraction that separates source data from information items of the data model is analysed in (Gram *et al.*, 2020).

The encircled lower section represents entity and market data collection from sources described with appropriate DDI metadata definitions of work package 4 for entity and market data.

Current work is ongoing regarding CDM standards for collecting and identifying financial data associated with firms and financial instruments. In the following sub-chapters, we describe the current revisions of the defined standard: 7.1 Legal Entity Data Standard Version 1.09, 7.2 Financial Instrument Data Standard Version 1.05, and 7.3 Legal Entity Data Artifact Standard Version 1.05.

2.4.1 Legal Entity Data Standard

Figure 4 illustrates the core classes of the EURHISFIRM Common Data Model. Since the EURHISFIRM Research Infrastructure initiative has as a core objective the collection and integration of the financial histories of the different European states, the identification of the entities that have been the historical actors and participants in the economies of Europe is one of the central classes of things that must be uniquely identified in order to record the economic activities and facts and associate them with the entities to which they belong. The objective of the EURHISFIRM Legal Entity Identifier (ELEI) data standard is to define the type of information that is required to identify a historically active economic organisation uniquely (be that a private company, public corporation, or possibly a government-sponsored entity) such that a unique identifier can be assigned to the entity.

2.4.2 Financial Instrument Data Standard

Another fundamental aspect of a historical financial collection of economic data is the core class of financial instruments that are publicly traded in market venues (exchanges), and whose price histories reflect the market participants view of the prospects of the economic activities of firms who offer these financial instruments to the public as investments in the potential opportunities of future growth and economic return of the firm. Since there are many different types of such market instruments or investment vehicles, standards for the unique identification of these financial instruments are also needed that takes into account the changes that can take place in a given security over time. The EURHISFIRM Financial Instrument Identifier (EFII) data standards were created for that purpose.

2.4.3 Legal Entity Data Artifact Standard

Given that the core identifiers for legal entities and the market instruments that legal entities can issue have been defined, it is necessary to have a means by which historical facts and information can be collected, identified, and associated with the relevant core classes to which they are related. For example,

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

a flexible means to record a variety of financial facts about an entity which have been obtained from historical yearbooks (e.g., income statement, balance sheet items, outstanding share capital) is needed. Likewise, it is important to be able to capture a variety of market data items about financial market securities (different types of prices, trading volumes, etc.). In most of these cases, especially historically, there is no single structure that applies to all of the varied facts or elements that can be found in various sources and periods. The purpose of the Legal Entity Data Artifact Standard is to provide a flexible means by which these data can be captured and associated with the core identifiers of legal entities and their respective financial securities.

3 Implications of Stakeholder Survey

The aim of report D5.3 was to present preliminary design decisions resulting, for instance, from the important existing national implementations and project discussions with various stakeholders inside and outside the project. The results are used to – when necessary – revise the preliminary CDM design.

Although the users' feedback is centred on the EURHISFIRM CDM, users ask and must receive additional information on the eco-system in which the CDM is embedded to comprehend all critical factors influencing their desires. Thus, respective preliminary information on decision units and processes have also been provided.

Three key points are already proposed in EURHISFIRM's project application to the European Commission. One is the federated architecture of the system. Federated means, for example, that the data is kept stored in their countries of origin, and the national implementations would continue to exist. Through content standards and communication standards, and common decision-making processes, the data's comparability is made possible.

The raw data is semantically lifted from the existing format into a harmonised and comparable format as a second critical point. The end-user shall be able to access both the source data and the harmonised data.

The third key point is establishing a central identification scheme, most prominently for firms (legal entities), for which the ELEI concept was derived from the LEI concept.

The original plan was to present the preliminary CDM-design at the annual General Assembly in spring 2020 and collect and evaluate feedback during and after the event. Due to the COVID-19 pandemic, the event was made more compact and performed electronically in autumn 2020, resulting in half a year delay to the original schedule. A survey link was sent to all participants in the EURHISFIRM project after the General Assembly to get their feedback. Participants also had the opportunity to nominate other stakeholders outside of the project to participate in the survey.

Following the survey, in-depth interviews were conducted with various stakeholders. Some of these people had already taken part in the survey but expressed questions or scepticism about selected design pre-decisions. The persons' list was expanded to include knowledgeable stakeholders outside the project.

The following two tables show the survey respondents' self-classification according to their stakeholder group and the country of origin.

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

Groups of Stakeholders	Specification	#	%
Consumer	A user of European historical financial and firm data	17	48.57
Producer	A generator of European historical financial and firm data	10	28.57
Infrastructure Service Provider	Data centre manager or service provider for European historical financial and firm data	7	20.00
Other	(Respondent defined him-/herself as 'consumer AND producer')	1	2.86
TOTAL		35	100.00

Table 1: Assignment of potential stakeholders (survey participants) in groups

The predefined user classes are nicely covered.

Country of Origin	Austria	Belgium	Europe	Finland	France	Germany	Netherlands	Poland	Portugal	Spain	Sweden	Switzerland	United Kingdom	
#	1	4	1	1	5	4	4	3	3	3	1	2	3	35

Table 2: Countries of origin of stakeholders (survey participants)

Also, concerning the countries of origin, we have excellent coverage. One response is collected from a European Institution representative, so this response is labelled as "Europe".

In addition to the survey, we conducted eleven interviews with selected stakeholders representing various financial institutions, such as banks, ministries, and professors in the historical financial and firm area. When selecting the interviewees, we ensured that the stakeholder groups already mentioned in the survey - consumer, producer, and service provider - were represented. This job diversity allows us to broaden a cross-section of responses as possible and ensure that relevant requirements are included.

In the following chapter, we present important consolidated results of D5.3. First, we look at the components where respondents at least overwhelmingly confirmed the proposed design. Second, we discuss the components where respondents were not aligned with the proposal and, if available, evaluate design alternatives. For each component, we start with the explanation of the question to reason why asking this question is crucial for the design of the CDM. We then display the question asked in italic font style. Lastly, we summarise the answers of the respondents and discuss the impact on the CDM design.

3.1 Confirmed Components

From the data gathered by the questionnaire and interviews, we focus on the components where updates would directly impact the CDM. We are talking about a confirmed component, as no update of the CDM is to be made. These components are EURHISFIRM's federated architecture, the identification of entities and the harmonisation of data.

3.1.1 Federated System

A federated database consists of a semi-independent distributed database structure and provides local data autonomy (Hammer and McLeod, 1979) in compliance with centrally agreed-upon standards in content and communication structures, allowing, for instance for a common data access. This local

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

independence makes it possible to link different databases and to also transfer source data independently into a local database structure. However, some critical database design, maintenance, and administration tasks need to be agreed upon and centrally governed, for instance, the common identification schema for legal entities, although redundancy could occur due to different user data structures in each local context (Hammer and McLeod, 1979). In conclusion, in a federated database, the data is stored in local databases, whereas an overarching database administrator takes care of the compliance to the common data and communication structures and enables the connection between the different federated databases. Governance processes ensure that decisions on common structures are implemented and adhered to at various federated architecture levels. We pursue the federated architecture for the EURHISFIRM database so that the individual countries with local specialists provide, for instance, the collation and administration of the data, and end-users can retrieve the information via a common access point. Considering this understanding, we formulated the following question:

EURHISFIRM aims at designing a research infrastructure to collect, merge, extract, collate, align, and share detailed historical financial and firm data for Europe. In EURHISFIRM we plan to keep the databases in the different countries. However, the user should be able to access the data through a central access point. Do you agree to distributed data centres across Europe and a common access point for the federated architecture or do you have any comments?

The stakeholders' feedback is evident and strong that only a federated system can fit their requirements for EURHISIFIM. This also follows the European principle only to centralise what cannot be done well at the decentral level. The numerous idiosyncrasies in company law, accounting regulation, and financial market regulation between countries with different legal systems make it almost impossible to have a one-size-fits-all data model. A centralised system imposing a common model in the first stages of data collecting would most likely result in aversion. Instead, we must design a stepwise harmonisation process ‘bottom-up’.

3.1.2 Identification of Entities

We proposed to store the data of firms in EURHISFIRM using a further developed Legal Entity Identifier – the ELEI – as the primary key. The LEI is the most advanced standard to identify companies worldwide and assigns a truly unique identifier to every legal entity, including elaborated deduplication procedures and employing a kind of physical inspection in case of ambiguity (Global Legal Entity Identifier Foundation, 2020). Using a unique identifier is necessary for four reasons (Chan and Milne, 2019): First, there could occur different spellings for the same company. Second, national subsidiaries may have different names and should be assigned to the same parent company. Third, while extracting companies' additional data from sources, a firm could be assigned to different corporate identities when introducing them into the EURHISFIRM database. Moreover, fourth, errors could occur due to manual data translations. Moreover, (Bottega and Powell, 2012) state that standardised legal identifier usage results in a cost- and application-efficient implementation.

(Chan and Milne, 2019) conducted semi-structured interviews to analyse the advantages of the LEI. They highlighted its importance for analysts and economic research since other international agreements on standardised identifiers like the ISIN do not to a hundred per cent uniquely work for any legal entity. Besides, especially when dealing with historical data, using a standardised identifier is essential because

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

then the former mergers or acquisitions can be considered effectively (Bottega and Powell, 2012). As the LEI concept only reflects contemporary legal entities, we further developed the LEI to cater to historical legal entities – called the EURHISFIRM LEI (ELEI) – as the standard identifier for EURHISFIRM. Therefore, we ask:

In practice, different databases often use different identifiers. Consumers of the data can manually merge these, or the cross-tabulation can be performed by a service provider. The Legal Entity Identifier (LEI) is the most advanced standard to uniquely identify legal entities. It consists of an alphanumeric code linked to essential reference data that allows for a clear and unambiguous identification of companies participating in financial transactions. The improvement of the LEI to cater for historical legal entities is called "ELEI" (EURHISFIRM Legal Entity Identifier). In EURHISFIRM, we have agreed that all legal entities will have a EURHISFIRM Legal Entity Identifier (ELEI). Do you think that this approach to introduce the ELEI as an identifier for European historical firms is a good solution for this purpose? If not, what might you propose instead?

All stakeholders post the opinion that common unique identification is a crucial factor for the system's success. It is stated multiple times that assigning such identifiers is a resource-intensive task, but it is required to allow overarching analysis and to deduplicate data. The LEI is the ideal foundation as it is an established standard for contemporary data in Europe and beyond.

3.1.3 Data Harmonisation

Conceptually, we want to offer a common data access service (CDAS) to the users that inquire about data from EURHISFIRM components. To this end, we strive for gateways to transform user needs to data availability. We face a semantic distance between the users' expression of queries to the federated database and the different conceptual and physical data representations. We employ a sequence of data staging steps to accomplish the transformation and integration of sources into the CDM representation. The lowest stage comprises raw data from historical sources. We developed a hierarchy of stages to enrich raw data through harmonisation to EURHISFIRM compliant data. Thus, we asked:

There is a sequence of stages involved in the enrichment of raw data, as that data moves through the processes of verification, standardisation, harmonisation, and integration to be presented to the end-user in a common-data-model form of access. Would you agree that this approach to collecting and processing source data in stages and making it available to end-users in a common form of access is appropriate? If so, which organisations do you know that could help and support EURHISFIRM actors in the different phases of this process?

The overwhelming majority of stakeholders also find a harmonised version of the data necessary, with common identification being perceived as more important than other information's comparability. Nevertheless, the raw data remains essential, for instance, to allow a reader to verify the harmonisation's accuracy. Most respondents find a harmonised version essential to carry out quick analyses and not have to have the necessary time and expertise to harmonise the data by themselves.

However, one EURHISFIRM member regards each harmonisation of data as a fundamental error of manipulating data suggesting a 'clean' world that is not given. As our data architecture allows access to

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

raw and to harmonised data, the end-user can freely decide what he/she wants to access. So, no change of the CDM is regarded necessary.

3.2 Challenged Components

We classify components as challenged when a stakeholder's response may imply a change to the CDM. This does not mean that a single user's request requires the CDM to be updated, but requests need to be discussed in the WGIS group. This group decides about revisioning the CDM and incorporates the stakeholders' feedback. The challenged components are identifying financial instruments, external sources of data and user privileges in a collaboration platform.

3.2.1 Identification of Financial Instruments

Also, financial instruments, for instance, securities, need to be uniquely identified. The requirements for the identification of financial instruments are similar to those of firms. Considering different treatments of, for instance, the firm's shares when listed on multiple stock exchanges, the data needs to be merged. Moreover, users need to access and gain information about the market values of securities in different jurisdictions. In reflection of these necessities, we select the identification standard of the non-profit Object Management Group's (OMG.org) Financial Instrument Global Identifier (FIGI).

We have established a scheme for identifying financial instruments, the EURHISFIRM Financial Instrument Identifier (EFII). It comprises a three-level hierarchy for the structure of this identifier, following the Object Management Group's (OMG) Financial Instrument Global Identifier (FIGI), which provides for the identification of financial instruments at the hierarchical levels composed by the identification of the trading venue, the identification of the sovereign jurisdiction, and the identification of the issuing firm. This supports the identification of a financial instrument at the point of collection and the subsequent harmonisation of the information collected about the same financial instrument at other trading venues. Do you think that this approach to introduce the EFII as an identifier for European historical financial instruments is a good solution for this purpose? If not, what might you propose instead?

First, it is essential to state that all stakeholders agree that having common identification of financial instruments is essential. The debated part of this question is the underlying FIGI as the base standard for the derived EFII. Multiple stakeholders state – but this is still a minority vote – that it should be considered to take the International Securities Identification Number (ISIN) as the foundation for the EURHISFIRM EFII. Therefore, the topic was again brought up in one of the biweekly WGIS meetings with the subject matter experts from the different work packages. When the WGIS chose FIGI the ISIN was already considered but rejected. With its three hierarchical levels, the FIGI does fit better the needs and requirements for EURHISFIRM to work with historical financial instruments. As a result, the EFII standard specification does not need to be revised regarding the feedback.

3.2.2 External Sources of Data

We asked for the needs of interoperability and integration of EURHISFIRM in and with other databases providing historical financial or firm data. There are differently structured databases like Thomson Reuters Eikon, Bloomberg, S&P's Capital IQ, CRSP, COMPUSTAT, the global LEI system, and EUROFIDAI that are used in different scientific disciplines for answering research questions. However, it is essential to ensure

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

interoperability between these databases. Interoperability means, on top of the capability to export data, that data can, for example, be integrated with other databases, which is achieved, for instance, by having the same data type definitions and identification scheme for data. However, this degree of interoperability between extensive repositories of financial data, each with its own set of definitional standards, is an undertaking that will require a large amount of work to develop these equivalent mappings to external systems. A more immediate way to refer to data in external repositories is to add links to external data that associate identified instances of core objects and classes in the local system (i.e., EURHISFIRM).

What other standards should be considered for integrating the European historical financial and firm data with other data sources to ensure interoperability with other research infrastructures? What other research infrastructures would you like to integrate with EURHISFIRM possibly?

The majority of the stakeholders highlight that with the standards designed by the project team, the needs and requirements are majorly covered – also against the background that the FAIR principles have to be fulfilled. This coverage is confirmed by the fact that there were not many additional suggestions from the stakeholders.

Some stakeholders – still a minority – highlight that the topic of controlled vocabularies (such as Getty Thesaurus of Geographic Names) should be considered for a standard definition. This topic has been discussed in the WGIS and is added to the roadmap of topics considered for the further development of the CDM.

3.2.3 User Privileges in a Collaboration Platform

The EURHISFIRM database enriches with the involvement of many different stakeholders. The success of EURHISFIRM depends on the data's contribution, which we – to a large extent – can expect to be the stakeholders' common goal. This type of enhancing a database is called a collaborative network (Camarinha-Matos and Afsarmanesh, 2005). Collaborative consolidation, in this case, is defined as "a process of combining, integrating or transforming something into something else that is complete, effective, coherent or elaborated" (Junior and Pereira, 2020).

The collaboration based on the federated architecture increases the database's usability as multiple subject matter experts can assist each other (Hvannberg, Law and Halldorsdottir, 2019). We, therefore, asked the stakeholders about their ideas on which would be a good way to work together.

In your opinion, what would be a good way to collaboratively participate in the collection, enrichment, and publication of data of historical European financial and firm origin?

In general, all stakeholders consider it essential that EURHISFIRM allows the collaboration of various user groups and is – at least in parts – accessible to the public. The point of discussion is who should be allowed to make modifications to attribute values stored in EURHISFIRM or even to the CDM. The opinions vary from an unrestricted Wikipedia style where everyone should be able to make a change up to a restricted model where changes only fall into the area of responsibility of the various centres of competencies or the EURHISFIRM central organisation. Most stakeholders support the possibility that users of the solution should be able to request changes. To request a change, opinions vary by the evidence level that needs to

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

be provided to support such a change request. As this topic falls into the governance area, it is not directly impacting the CDM development. The issue is also put on the roadmap for future consideration in WGIS.

3.3 Resulting Changes to the Common Data Model

As mentioned in the introduction, the information gathering to update the CDM was delayed by the corona virus outbreak. Nevertheless, the development of the CDM continued between spring and autumn 2020. The developed updates can be obtained in the change summary at the beginning of each standard document in the appendices. During this timeframe the ELEI and EFII received updates and the LEDA standard was introduced in an initial version. The need for this additional specification was raised by the various stakeholders represented in the EURHISFIRM project team.

The questionnaire and interview data results did not result in significant updates to the specifications, but provided an important confirmation of the work done by the project team.

4 Future Development of the Common Data Model

The future development of the CDM can be viewed on two levels. One level is the process level, which describes the mechanism by which the CDM and its components will be developed. This process is described in report D5.5 (Ranft et al., 2021).

To shortly illustrate the governance of these changes to the CDM, we describe a procedure based on the model of GS1, an organisation that has been implementing a process for generating and updating worldwide standards for logistics data management for many decades (GS1, 2019). This process of self-binding cooperative decisions on standards (a concept of committed consensus) has already been similarly applied by the Working Group of Identification and Standards (WGIS) within the project. For this purpose, after some ramp-up activities of all working packages, we agreed on an initial list of to be standardised objects. Subject matter experts of each work package were then delegated to the group, which developed and agreed on the standards together. These standards were designed to align with the FAIR principles to provide a research infrastructure that facilitates findable, accessible, interoperable, and reusable data (Wilkinson *et al.*, 2016).

The second level describes the direction in which the CDM is to be developed. This is derived primarily from feedback from the various stakeholder groups. The WGIS team is currently discussing a list of topics that need to be discussed in the future. In addition to developing standards of controlled vocabularies, stakeholders provide feedback on what the EURHISFIRM project team should focus on after the end of the project. One idea is to choose a common industry that existed in most European countries. Collecting data from the same industry shows the nuances where data differs between the different countries. Another approach is to choose the biggest stock listed companies of each European country as the data availability is probably the best. This data should then be used to build a prototype. Another suggestion was to closely collaborate with state institutions such as national archives or even district courts as they have much data on firms in their archives. The last suggestion was to look for other projects (such as GAIA-X) that are trying to build a data infrastructure in the European finance and firm realm. Synergies between the projects might be beneficial for both parties.

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

5 Summary

The report provides the most current state of the EURHISFIRM CDM development at a late point in the project (INFRADEV phase 1). With identification being a central point, these standards set the design foundation for further development and subsequent implementation of the system.

We show that the standards fit many of our identified stakeholder groups' needs and requirements and provide potential options for the future development of EURHISFIRM.

The report does not provide a finalised set of standards as the development is an ongoing, incremental process which is further described in WP5's report D5.5 (Ranft et al., 2021). So even after a system implementation, the CDM needs to be revised in – at least at the beginning – small steps to continuously fulfil the needs and requirements of EURHISFIRM's stakeholders. In such a revolving update cycle, we expect the change rate for standardised items to drop quickly over time.

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

6 References

Bernstein, P.A. (2003) Published. 'Applying Model Management to Classical Meta Data Problems'. *CIDR,* 2003. Citeseer, pp.209-220.

Bottega, J.A. and Powell, L.F. (2012) 'Creating a linchpin for financial data: Toward a universal legal entity identifier'. *Journal of Economics and Business,* 64 (1), pp. 105-115.

Camarinha-Matos, L.M. and Afsarmanesh, H. (2005) 'Collaborative networks: a new scientific discipline'. *Journal of intelligent manufacturing,* 16 (4-5), pp. 439-452.

Chan, K.K. and Milne, A. (2019) 'The global legal entity identifier system: How can it deliver?'. *Journal of Risk and Financial Management,* 12 (1), pp. 39.

Global Legal Entity Identifier Foundation (2020) *About GLEIF: History of the Global LEI System.* Available at: https://www.gleif.org/en/about/history (Accessed: 08.12.2020).

Gram, D. *et al.* (2020) 'An extensible model for historical financial data with an application to German company and stock market data'.

GS1 (2019) 'Global Standards Management Process (GSMP) Q1 2019 Management Report'. Available at: http://www.gs1.org/gsmp/reports.

Hammer, M. and McLeod, D. (1979) 'On Database Management System Architecture'.

Hvannberg, E.T., Law, E. and Halldorsdottir, G. (2019) 'Argumentation models for usability problem analysis in individual and collaborative settings'. *International Journal of Human–Computer Interaction,* 35 (3), pp. 256-273.

Junior, D.S. and Pereira, R. (2020) Published. 'Consolidation in Collaborative Design: An Exploratory Case Study'. *Anais do V Workshop sobre Aspectos Sociais, Humanos e Econômicos de Software,* 2020. SBC, pp.11-20.

Karapanagiotis, P. (2019) 'EURHISFIRM D5.1: Technical document on national data models'. Available at: https://doi.org/10.5281/zenodo.3467926.

Karapanagiotis, P. (2020) 'EURHISFIRM D5.2: Technical Document on Preliminary Common Data Model'. Available at: https://doi.org/10.5281/zenodo.3686930.

Khatri, V. and Brown, C.V. (2010) 'Designing data governance'. *Communications of the ACM,* 53 (1), pp. 148-152.

Ranft, L.B., Jefferson; König, Wolfgang (2021) 'EURHISFIRM D5.5: Report on process for extendable data models'. *Zenodo*.

Tochtermann, K. and Loebbecke, C. (2018) Published. 'Towards a FAIR Internet of Data, Services and Things for Practicing Open Science'. *Thirty-ninth International Conference on Information Systems,* 2018 San Francisco.

Wilkinson, M.D. *et al.* (2016) 'The FAIR Guiding Principles for scientific data management and stewardship'. *Scientific Data,* 3 (1), pp. 160018.

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

7 Appendix

7.1. Legal Entity Data Standard 1.09

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

EURHISFIRM Common Data Model

Legal Entity Data Standard 1.0

Version of 28 November 2020

Date	Revision	Description
2020-01-15	ver 1.01	First Draft to be circulated (formatting fixed)
2020-02-06	ver 1.02	Added partitioning scheme for ELEI in Appendix 11; made `Line1` of the street address in the **`Address`** type optional; qualified the term "entity" with "legal entity" throughout; revised the legal entity record transition diagram appendix
2020-02-19	ver 1.03	Added `EntityCreationDate` start date for legal entity to complement the `EntityExpirationDate` Added `RegistrationUpToDate` and `RegistrationAsOfDate` to bracket the start and stop dates that this reference data is valid Added `RegistrationSourceReference` in order to provide a reference to the definitive data source from which this data was collected Added several revisions to element definitions
2020-02-20	ver 1.04	Added ELEI Root record (needed to anchor the unique ELEI in order to support multiple (historical) ELEI Data records
2020-03-04	ver 1.05	Revised the ELEI root record to include `EarliestRegistrationDate, LatestRegistrationDate. EntityStatus` Added `EntityEvents` to ELEI data record Revised `CountryCode` to be `ECountryCode` and `RegionCode` to be `ERegionCode` in order to include historical countries and regions

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

2020-03-18	ver 1.06	Consolidated comments and accepted revisions Revised `LegalForm` to be `ELegalForm`, `RegistrationAuthority` to be `ERegistrationAuthority` and `RegistrationAuthorityEnum` to be `ERegistrationAuthorityEnum` in order to be able to accommodate historical forms
2020-04-02	ver 1.07	Converted the ERegistrationAuthority elements into a data type Added OtherRegistrationAuthority element to allow for historical government legal entity authorities that are not yet included in the RegistrationAuthority external code list to be recorded; Elaborated on the `RegistrationSource` element in the Registration Record to make it a source metadata reference structure containing `RegistrationSourceLocatorKey`, `RegistrationSourceLocatorAnchor`, `DCUID`, and `RegistrationSourceEntityID` elements. Added a `DCUID` data element in order to identify Data Collection Units as well as Data Submission Units; Added preliminary elements for the Header structure;
2020-04-14	ver 1.08	Renamed `RegistrationSourceLocator` to `RegistrationSource`. Added `RegistrationSourceCitation` to `RegistrationSource` Data Type
2020-11-26	ver 1.09	Changed references to "legal entity" to "entity" in order to allow the inclusion of public sector organizations (sovereign governments, nationalized cororations, etc)

Abstract

First, the semantic content of these attributes must be fully specified. Second, some additional elements, such an indication of the status of the information, are necessary for effective use of the data. Third, the form the information takes at any given local point of source data capture must be such that it can be made to conform to a common standard, which must also be specified. This document proposes the standards necessary in these areas to support the EURHISFIRM Common Data Model.

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

Status of this document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. The latest status of this document series will be maintained on SeaFile.

This draft is a **Working Draft** which can be circulated to any interested parties for review and comment. It is a draft document and may be updated, replaced or made obsolete by other documents at any time. It is inappropriate to use Working Drafts as reference material or to cite them as other than "work in progress." This is work in progress and does not imply endorsement by the EURHISFIRM ExCo.

Comments on this document should be sent to [TBD: insert mailing list or URL].

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

Table of Contents

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

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

1 Introduction

Legal entities are clearly one of the core classes of objects in a historical financial Research Infrastructure such as EURHISFIRM, on which the majority of other EURHISFIRM financial data depend. As such, legal entities should be uniquely and unambiguously identified in EURHISFIRM, and this identification is accomplished based on the identifying attributes associated with each entity.

First, the semantic content of those attributes must be fully specified. Second, some additional elements, such an indication of the status of the information, are necessary for effective use of the data. Third, the form the information takes at any given local point of source data capture must be such that it can be made to conform to a common standard, which must also be specified. This document proposes the standards necessary in these areas to support the reference data attributes of legal entities in the EURHISFIRM Common Data Model

This document proposes initial standards for EURHISFIRM entity reference data. It is important that this reference data should uniquely identify the legal entities that are harvested from contributing sources and then collected and assimilated into the common EURHISFIRM platform.

A EURHISFIRM Entity Identifier (ELEI) code that resolves to this entity-identifying reference data is also introduced.

- A partitioning scheme for the structure of this identifier is defined that allows the independent and concurrent minting and assignment of ELEI codes to entity reference data that is produced by multiple Data Submission Units (DSUs) in the federated EURHISFIRM Research Infrastructure network.

The standard set by this document is expected, among other things, to reduce the risk of duplicates stemming from differences in formats and conventions of locally sourced data provided by regional contributors, to ensure data quality in the EURHISFIRM system, and to enable the subsequent detection and resolution of multiple identifiers for the same entity to the extent that they occur. The standard is expected to be used as a format for reference data consolidated from all sources in order to be promoted to the level of common EURHISFIRM data published for end-user access.

The contents of this document are as follows:

- Section 2 defines terminology and typographical conventions.
- Section 3 specifies the abstract content of EURHISFIRM entity reference data conforming to this standard, including a detailed description of each element of reference data associated with legal entities in EURHISFIRM. Allowable values for data elements that are code lists will also be (subsequently) specified. The partitioning scheme for the structure of the EURHISFIRM Legal Enty Identifier (ELEI) code will also be subsequently described.
- Section 4 specifies data validation constraints and other considerations intended to lead to high-quality data content.
- Section 5 specifies a concrete realization of the data definitions above in XML syntax, by means of XML schema (XSD 1.0).

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

- Section 6 specifies how this file format may be changed in the future, providing for versioning, forward- and backward-compatibility, etc.
- Section 7 provides examples to illustrate the file format.

2 Terminology and Typographical Conventions

Within this specification, the terms SHALL, SHALL NOT, SHOULD, SHOULD NOT, MAY, NEED NOT, CAN, and CANNOT are to be interpreted as specified in Annex G of the ISO/IEC Directives, Part 2, 2001, 4th edition [ISODir2]. When used in this way, these terms will always be shown in ALL CAPS; when these words appear in ordinary typeface they are intended to have their ordinary English meaning.

All sections of this document, with the exception of Section 1 are normative, except where explicitly noted as non-normative.

The following typographical conventions are used throughout the document:

- ALL CAPS type is used for the special terms from [ISODir2] enumerated above.
- `Monospace` type is used to denote programming language, UML, and XML identifiers, as well as for the text of XML documents.

➢ Placeholders for open issues and/or changes that need to be made to this document prior to its reaching the final stage of approved Proposed Standard are prefixed by a rightward-facing arrowhead, as this paragraph is.

3 Abstract Data Content

This section specifies the abstract data content of a a data file conforming to this standard.

A data file conforming to this standard SHALL consist of:

- An optional ELEI File Header, as specified in Section 3.1.
- An ELEI Root Record as specified in Section 3.2.
- Zero or more ELEI Data Records, as specified in Section 3.3.

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

3.1 ELEI File Header

- TBD Define elements to go into a "header" area for an ELEI file. The purpose of the header is to provide context about the file and its contained ELEI data records. However, the header will not contain anything necessary to interpret the meaning of any ELEI record; e.g., things like default values for ELEI data records will not be in the header (such things would mean that the meaning of an ELEI record could change if taken away from the header). Examples of things that might be useful to include in the header:

Element Name	Type	Card	Description
ContentDate	DateTime	1,1	The date and time of generation of the data
Originator		0,1	The identifier of the creator of the content of this file
FileContent		1,1	A code describing the content of this data file.
ProcessStage		0,1	A code indicating the stage of this file in the EURHISFIRM workflow
RecordCount		1,1	The number of data records in the file. Can be a positive whole (integer) number, or zero (0).

3.2 ELEI Root Record

The ELEI Root Record anchors a single ELEI. This root record is needed in order to provide a unique ELEI anchor to support multiple (historical) ELEI Data Records. Each ELEI Root record in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
ELEI	ELEI	1	The 20-character ELEI of the entity described by this ELEI Data Record.

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

Element Name	Type	Card	Description
SuccessorELEI	ELEI	0,1	The ELEI of the ELEI registration that supersedes or subsumes this ELEI registration for the same entity. If RegistrationStatus is DUPLICATE, then SuccessorELEI is the ELEI of the surviving ELEI Registration. If RegistrationStatus is MERGED, then SuccessorELEI is the ELEI Registration of the new/acquiring entity. Otherwise, SuccessorELEI is omitted. When a successor ELEI is assigned to an ELEI registration, the ELEI registration will no longer be updated (since another ELEI registration has superseded the registration that has just been assigned a successor). The ELEIRecordLastUpdate that is recorded with the update assigning a SuccessorELEI will be the last update performed on the superseded registration record. As a consequence, other fields of the superseded registration record (address, entity status, etc) may no longer reflect the actual state of the entity.
ELEICreationDate	DateTime	1	Date/time the ELEI root record was initially created in the system
EntityStatus	EntityStatusEnum	1	The status of the entity. This is not to be confused with the status of the registration; see RegistrationStatus. If this ELEI record contains a non-empty SuccessorELEI field, EntityStatus is the last status of the entity before the successor ELEI record superseded this one, which is not necessarily the current status of the entity.
EntityFormationDate	DateTime	0,1	Date/time the Entity was formed, if known
EntityExpirationDate	DateTime	0,1	The date that the entity ceased to operate, whether due to dissolution, merger or acquisition. Omitted if the entity has not ceased to operate, or if this ELEI record contains a non-empty SuccessorELEI field.

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

Element Name	Type	Card	Description
EntityExpirationReason	EntityExpirationReasonEnum	0,1	he reason that a entity ceased to operate. This element SHALL be present if EntityExpirationDate is present, and omitted otherwise.
EarliestRegistrationDate	DateTime	1	Date/time of the Registration record with the earliest historical date
LatestRegistrationDate	DateTime	1	Date/time of the Registration record with the most recent historical date

3.3 ELEI Data Record

An ELEI Data Record describes a single ELEI. Each ELEI Data record in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
ELEI	ELEI	1	The 20-character ELEI of the entity described by this ELEI Data Record.
Entity	Entity (Section 3.3.1)	1	Attributes describing the entity itself
Registration	Registration (Section 0)	1	Attributes describing the registration of this ELEI.
Extension	Extension (Section 0)	0,1	An optional element for including data beyond the standard data elements in an ELEI data file. This may include data specific to an DSU, data specific to a publisher of ELEI data, and so on.

3.3.1 Entity Section of ELEI Data Record

The Entity section of an ELEI Data Record in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
EntityName	Name	1	The preferred name of the Entity. If an Entity is in a jurisdiction with more than one Legal Name (e.g., in different languages), this is the Primary Legal Name (see otherEntityNames for other names).

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

Element Name	Type	Card	Description
LegalName	Name	0,1	The Legal Name of the Entity. If an Entity is in a jurisdiction with more than one Legal Name (e.g., in different languages), this is the Primary Legal Name (see otherEntityNames for other names).
OtherEntityNames	Other EntityName	0..n	An optional list of other Name instances for the Entity.
EntityLocation	Address	0,1	The address of the Entity
LegalAddress	Address	1	The address of the Entity as recorded in the registration of the Entity in its legal jurisdiction
HeadquartersAddress	Address	0,1	The address of the headquarters of the Entity
OtherAddresses	OtherAddress	0..n	An optional list of other Address instances for the Entity. This may be used to provide alternative language forms of legal address or headquarters address..
ERegistrationAuthority	ERegistration AuthorityType	0,1	A code that identifies the business register, or other registration authority that supplied the value of ERegistrationAuthorityEntityID.

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

Element Name	Type	Card	Description
ERegistrationAuthority EntityID	String	0,1	The identifier of the entity as maintained by a business registry in the jurisdiction of legal registration, OR If the entity is one that is not recorded in a business registry (e.g., one of the varieties of funds registered instead with financial regulators), the identifier of the entity in the appropriate registration authority. The RegistrationAuthorityEntityID element SHALL be included if RegistrationAuthority is included, and SHALL be omitted if RegistrationAuthority is omitted.
LegalJurisdiction	ERegionCode	0,1	The jurisdiction of legal formation and registration of the entity (and on which the legalForm data element is also dependent).
EntityCategory	EntityCategoryTypeEnum	0,1	Indicates the general category of the type of entity identified by this ELEI data record
LegalForm	ELegalFormType	0,1	The legal form of the entity, from an external code list that incorporates the ISO Entity Legal Form (ELF) code list mai

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

Element Name	Type	Card	Description
EntityStatus	EntityStatusEnum	1	The status of the entity. This is not to be confused with the status of the registration; see RegistrationStatus. If this ELEI record contains a non-empty SuccessorELEI field, EntityStatus is the last status of the entity before the successor ELEI record superseded this one, which is not necessarily the current status of the entity.
EntityEvents	EntityEvent	0,n	Corporate events that occurred during this historical period

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

3.3.2 Registration Section of ELEI Data Record

The `Registration` section of an ELEI Data Record in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`RegistrationCreationDate`	`DateTime`	1	Date/time this ELEI record was initially created in the system
`RegistrationUpdateDate`	`DateTime`	1	Date/time that this historical ELEI record was most recently updated in the system.
`RegistrationSource`	`RegistrationSourceType`	0,n	A locator of the source of the entity reference data in this registration record (mechanism TBD, typically company yearbooks)
`RegistrationStatus`	`RegistrationStatusEnum`	1	Status of the ELEI registration. This is not to be confused with the status of the entity itself; see `EntityStatus`.
`RegistrationUpToDate`	`DateTime`	0,1	Should it exist and be known, the historical date up until which time the ELEI identifying reference data for this entity is valid. (Not the same thing as when the data was changed in the system). If present, a subsequent historical record for this ELEI may exist with the revised information.

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

Element Name	Type	Card	Description
RegistrationAsOfDate	DateTime	0,1	The date that this version of the reference data is known to be valid. This allows a record of historical changes to the entity identifying data to be recorded. This date would typically be derived from either: 1. The publication date of the source reference, or 2. Information contained in the source reference
ResponsibleDSU	DSUID	1	The Identifier of the Data Submission Unit (DSU) that produced and manages this ELEI registration.
ValidationSources	ValidationSourcesEnum	0,1	The current validation status of this ELEI record, or omitted if the validation status is not known or not revealed.

3.3.3 Extension Section of ELEI Data record

The Extension section of an ELEI record may be used to include additional data not defined in this standard. For example, an DSU may use Extension to publish additional data elements it collects as part of registration.

- TBD: include the details of how this works. Basically, the idea is to use an XSD schema wildcard with namespace ##other, permitting the inclusion of XML elements from other XML namespaces.

3.4 Data Types

This section specifies the data types referenced by the tables in Section 3.2, in alphabetical order.

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

3.4.1 Address Data Type

A value of type `Address` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Cardinality	Description
`lang`	`LanguageCode`	0,1	The language in which all of the string-valued components of this address are expressed.
`Line1`	`String`	0,1	The first line of the street address
`Line2`	`String`	0,1	The second line of the street address
`Line3`	`String`	0,1	The third line of the street address. This element SHALL be omitted if `addressLine2` is omitted.
`Line4`	`String`	0,1	The fourth line of the street address. This element SHALL be omitted if `addressLine3` is omitted.
`City`	`String`	1	The name of the city
`Region`	`ERegionCode`	0,1	The "EURHISFIRM" region code for a region (state, province, county, parish, etc) that extends the ISO 3166-2 region code with historical additions
`Country`	`ECountryCode`	1	The "EURHISFIRM" country code that extends the 2-character ISO 3166-1 country code with historical additions [e.g., https://en.wikipedia.org/wiki/Lists_of_sovereign_states_by_year]
`PostalCode`	`String`	0,1	The postal code of this address as specified by the local postal service.

- TBD: should postal code be optional? [YES]

3.4.2 DateTime Data Type

A value of type `DateTime` in a conforming to this standard SHALL be a point in time expressed as a string conforming to ISO 8601 having the following format:

- TBD: Provision for different calendars ? (i.e., Gregorian, Julian, etc.)

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

YYYY–*MM*–*DD*T*hh*:*mm*:*ss.sss*TZ

where the components of the above string are as follows:

- *YYYY* is the year
- *MM* is the month (01 = January, …, 12 = December)
- *DD* is the day of the month (01 = first day of the month)
- T is the single character 'T'
- *hh* is the hour (00 – 23)
- *mm* is the minute
- *ss.sss* is the second and milliseconds. From one to three digits may be used for milliseconds, or omitted entirely along with the decimal point.
- *TZ* is the time zone specifier, which can be either:
  - Z the single character 'Z', denoting Coordinated Universal Time (UTC); or
  - +*hh*:*mm* denoting a positive offset from UTC; or
  - –*hh*:*mm* denoting a negative offset from UTC

In the XML representation specified in Section 5, the XSD type xs:dateTime is used; however, whereas xs:dateTime permits the time zone specifier to be omitted, DateTime values in files conforming to this standard SHALL always include a time zone specifier.

Explanation (non-normative): milliseconds are hardly necessary for ELEI reference data, and likewise it might seem simpler to allow only "Z" as a time zone specifier; however, XML processing tools support the full syntax given above and it is not always possible to restrict such tools to avoid milliseconds or force the use of "Z" as the time zone specifier. The restriction that the time zone specifier must be present is equivalent to using XSD type xs:timestamp; however this was introduced in XSD 1.1 and not supported by the majority of XML processing tools which still only implement XSD 1.0.

3.4.3 DCUID Data Type

A value of type DCUID in a file conforming to this standard SHALL be a [TBD]-character Data Collection Unit Identifier conforming to [TBD].

3.4.4 DSUID Data Type

A value of type DSUID in a file conforming to this standard SHALL be a [TBD]-character Data Submission Unit Identifier conforming to [TBD].

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

3.4.5 ECountryCode Data Type

A value of type `CountryCode` in a file conforming to this standard SHALL be a 2-character country code conforming to ISO ISO 3166-1 alpha-2 [ISO3166]. Note that ISO 3166-1 alpha-2 codes are all uppercase.

- TBD: CountryCode external code list must extend ISO country codes with start/stop dates and historical changes

3.4.6 ELegalForm Data Type

A value of type `LegalForm` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`LegalForm`	`ELegalFormCode`	1,1	The legal form of the entity
`OtherLegalForm`	`String`	0,1	Interim free-text legal form information in the process of transition to an ELF standard code

3.4.7 ELegalFormCode Data Type

`LegalFormCode` is a code that is a normative enumeration that subsumes the ISO Entity Legal Form (ELF) code list maintained by GLEIF, but which incorporates additional codes for historical legal forms that are no longer in use.

A value of type `LegalForm` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`ELegalFormCode`	`ELegalFormEnum`	1,1	The legal form of the entity

- TBD: Research on additions and extensions to ELF code list is needed, including start-stop date periods and historical jurisdictions (referencing extensions to the Country Code external code list)

3.4.8 ERegionCode Data Type

A value of type `RegionCode` in a file conforming to this standard SHALL be a code conforming to ISO 3166-2. Note that ISO 3166-2 codes are all uppercase.

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

- TBD: Like CountryCode code list, RegionCode external code list needs historical extensions and additions as well.

3.4.9 ERegistrationAuthorityType Data Type

A value of type `ERegistrationAuthority` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`ERegistrationAuthorityID`	`ERegistrationAuthorityEnum`	0,1	An identifier for the entity **registry** of the entity in the jurisdiction of legal registration, or in the appropriate registration authority.
`EOtherRegistrationAuthority`	`String`	0,1	A legacy / historical reference code of a registration authority which is not yet entered in the ERegistration Authorities List (RAL), or the designation of an interim register until such time as an entry from RAL can be delivered .
`ERegistrationAuthorityEntityID`	`String`	0,1	An identifier for the **entity** at the registry in the jurisdiction of legal registration, or in the appropriate registration authority.

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

3.4.10 ELEI Data Type

A value of type `ELEI` in a file conforming to this standard SHALL be a 20-character Entity Identifier conforming to [ISO17422]. (See Appendix 11: ELEI Code Partitioning Scheme for details on the partitioning of the ELEI code.)

3.4.11 EntityCategory Data Type

A value of type `EntityCategory` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`EntityCategory`	`EntityCategoryTypeEnum`	0,1	Indicates the general category of the type of entity identified by this ELEI data record

3.4.12 LanguageCode Data Type

A value of type `LanguageCode` in a file conforming to this standard SHALL be a 2-character language code conforming to [ISO639-1]. Note that ISO 639-1 language codes are all lowercase.

- TBD: Alternatively, we could use IETF language tags (RFC 4646) instead. IETF language tags can distinguish between variations of the same language in different countries. For example, the IETF language tags `fr-CA` and `fr-FR` denote Canadian French and French as spoken in France, respectively; in ISO 639-1 these would both be simply `fr`.

3.4.13 Name Data Type

A `Name` is a string expressed in a natural language, including a code indicating which natural language is used.

A value of type `Name` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`lang`	`LanguageCode`	0,1	The language of `name`
`Name`	`String`	1	The name itself.

3.4.14 OtherAddress Data Type

A value of type `OtherAddress` in a file conforming to this standard SHALL include data elements as specified below. Each `Address` element includes an optional language code, permitting `OtherAddress` to be repeated as many times as necessary to express the same address type in multiple languages. The purpose of the `AddressTypeEnum` code list is to

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

accommodate legal address and headquarters address in different languages, not to add other address types (which could conceivably be added in the future).

Element Name	Type	Card	Description
type	AddressTypeEnum	1	The type of address represented by this OtherAddress instance.
Address	Address	1	The address

3.4.15 OtherEntityName Data Type

A value of type OtherEntityName in a file conforming to this standard SHALL include data elements as specified below. Each Name element includes an optional language code, permitting OtherEntityName to be repeated as many times as necessary to express the same name type in multiple languages. When type is PREFERRED_ROMANIZED_LEGAL or AUTO_ROMANIZED_LEGAL, the language code specifies the language of the name prior to Romanization.

Element Name	Type	Card	Description
type	EntityNameTypeEnum	1	The type of name represented by this OtherEntityName instance. The EntityNameType observes language, since 'Name' type has a language attribute.
Name	Name	1	The name. If type is PREFERRED_ROMANIZED_LEGAL or AUTO_ROMANIZED_LEGAL, then this value SHALL only include characters from the character set specified in Section 9.

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

3.4.16 RegistrationSource Data Type

RegistrationSourceCitation	RegistrationSourceCitationType	0,1	A locator of the source of the entity reference data in this registration record (mechanism TBD, typically company yearbooks)
RegistrationSourceLocatorKey	RegistrationSourceLocatorKeyType	0,1	A locator of the source of the entity reference data in this registration record (mechanism TBD, typically company yearbooks)
RegistrationSourceLocatorAnchor	RegistrationSourceLocatorAnchorType	0,1	Additional information to allow pinpointing where in the RegistrationSource can the entity data be found
RegistrationSourceDCU	DCUID	0,1	The identifier of the Data Collection Unit that sourced the registration data
RegistrationSourceEntityID	String	0,1	An identifier of the entity that may have been locally used or assigned by the DCU at collection time

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

3.4.17 RegistrationSourceCitation Data Type

An associative locator (link or key) that resolves to a DDI metadata block regarding the source of the entity reference data in this registration record (mechanism TBD, typically company yearbooks -)

DDI 3.2 uses the `CitationType` Element for referencing publications (e.g. yearbooks). It contains 11 elements which are described in detail in the *DDI Lifecycle XML Schema*. The table below lists the suggested cardinality for each element and some notes on special uses. The name of each element is a link to its description in the DDI XML Schema.

Element Name	Type	Card	Description
Title	DDI Title	1,1	
SubTitle	DDI SubTitle	0,1	
AlternateTitle	DDI AlternateTitle	0,n	
Creator	DDI Creator	0,n	
Publisher	DDI Publisher	0,1	
Contributor	DDI Contributor	0,n	
PublicationDate	DDI PublicationDate	0,1	For monographs: SimpleDate For serials (e.g. yearbooks): StartDate and EndDate (the date/year when the first and the last volume of a serial were published – the specific volume and page numbers of the year from which information about the entity was taken can then be recorded in the RegistrationSourceLocatorAnchor element)
Language	DDI Language	0,n	
InternationalIdentifier	DDI InternationalIdentifier	0,n	
Copyright	DDI Copyright	0,1	
dc:isPartof	DDI dc:isPartof	0,1	Use for instance to record the title of the newspaper in which a stock exchange price list is published

The Citation as we would want to use it in the Entity Reference Data would be considered as a DataSource in DDI 3.2 (Origin element).

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

3.4.18 RegistrationSourceLocatorAnchorType Data Type

The `RegistrationSourceLocatorAnchorType` provides additional information to allow pinpointing where in the RegistrationSource referenced by the `RegistrationSourceLocatorKey` can the entity data be found

3.4.19 RegistrationSourceLocatorKeyType Data Type

An associative locator (link or key) that resolves to a DDI metadata block regarding the source of the entity reference data in this registration record (mechanism TBD, typically company yearbooks -)

3.5 Enumerated Code Lists

This section specifies the enumerated code list data types (all having the suffix `Enum`) referenced by the tables in Sections 3.2 and 0, in alphabetical order.

3.5.1 AddressTypeEnum Code List

The `AddressTypeEnum` value in an `OtherAddress` instance specifies how the alternative address relates to the entity.

A value of type `AddressTypeEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
`LEGAL_ADDRESS`	Registered address of the entity in the legal jurisdiction
`HEADQUARTERS_ADDRESS`	Address of the headquarters of the entity

3.5.2 ELegalFormEnum Code List

The value of `ELegalFormEnum` SHALL include codes published in the ISO Entity Legal Form (ELF) code list maintained by GLEIF, but with additions needed for historical variations

- TBD: Research on additions and extensions to ELF code list is needed, including start-stop date periods and historical jurisdictions (referencing extensions to the Country Code external code list)

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

3.5.3 ERegistrationAuthorityEnum Code List

The `RegistrationAuthorityEnum` value in an `Entity` instance specifies what business register provided the value of `RegistrationAuthorityEntityID` for the entity.

The value of `ERegistrationAuthorityEnum` SHALL be a code published in a list of registration authority codes.

- TBD: the external code list for Registration Authorities needs to be developed, but will borrow from, incorporate or reference the Registration Authorities Code List published by the Global LEI Foundation.
- TBD: this may become a metadata reference to the source document (e.g., yearbook) that provided the information about the Entity

3.5.4 EntityCategoryTypeEnum

The `EntityCategoryTypeEnum` value in an ELEI record specifies the general classification category of the entity.

- TBD: the table below is just an example with some possible categories, and is incomplete, and needs to be elaborated.

A value of type `EntityCategoryTypeEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

- TBD: Alternative to "Corporation" ? (Note: these enumerated codes are currently used by the GLOBAL LEI System – changing them creates differences with the current entity identification system) [http://dx.doi.org/10.1080/00076791.2013.837893]

Code	Definition
PUBLIC_CORPORATION	Public shareholder entity
PRIVATE_CORPORATION	Privately owned limited liability corporation
LLC	Privately owned limited liability company
SOLE_PROPRIETOR	Private business owned and operated by a single individual
LLP	Privately owned limited liability partnership
TRUST	Legal trust
FUND	Alternative Investment Scheme
GSE	Government sponsored entity
NGO	Non-Government Organization
GOV	Public Sector, Government Organization

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

3.5.5 EntityExpirationReasonEnum Code List

The `EntityExpirationReasonEnum` value in an ELEI record specifies the reason that the entity expired.

A value of type `EntityExpirationReasonEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
DISSOLVED	The entity ceased to exist, and was legally shuttered.
CORPORATE_ACTION	The entity was acquired or merged with another entity
OTHER	The reason for expiry is not one of the above

3.5.6 EntityNameTypeEnum Code List

The `EntityNameTypeEnum` value in an `EntityName` specifies how the name relates to the entity.

A value of type `EntityNameTypeEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
OTHER_LEGAL	Registered name of the entity in an alternate language in the legal jurisdiction in which the entity is registered
PREFERRED_ROMANIZED_LEGAL	Romanized form, preferred by the entity
AUTO_ROMANIZED_LEGAL	Romanized form, auto-transliterated

3.5.7 EntityStatusEnum Code List

The `EntityStatusEnum` value in an ELEI record indicates the status of the entity itself. This is not to be confused with the status of the ELEI registration, which is specified by `RegistrationStatusEnum` (Section 3.4.8). See also Section 10, which illustrates how the `EntityStatusEnum` value changes over the lifecycle of an ELEI registration.

A value of type `EntityStatusEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
ACTIVE	As of the last report or update, the entity reported is legally registered and operating.

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

Code	Definition
INACTIVE	It has been determined that the entity that was assigned the ELEI is no longer legally registered and/or operating, whether as a result of: 1. Business closure 2. Acquisition by or merger with another (or new) entity 3. Determination of illegitimacy [perhaps not required for historical entities]
NOT_SPECIFIED	The ELEI record is in a state in which it does not provide information about whether the entity is legally registered and operating.

3.5.8 RegistrationStatusEnum Code List

The RegistrationStatusEnum value in an ELEI record indicates the status of the registration of the entity with an DSU. This is not to be confused with the status of the ELEI itself, which is specified by EntityStatusEnum (Section 3.5.7). See also Section 10, which illustrates how the EntityStatusEnum value changes over the lifecycle of an ELEI registration.

A value of type RegistrationStatusEnum in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
PENDING	An application for an ELEI that has been submitted and which is being processed and validated. *NOTE: ELEI registrations in the PENDING state are not intended for public release, but could be used internally between DSUs.*
ISSUED	An ELEI Registration that has been validated and issued, and which identifies ggal entity that was an operating entity as of the last update.
DUPLICATE	An ELEI Registration that has been determined to be a duplicate registration of the same entity as another ELEI Registration; the DUPLICATE status is assigned to the non-surviving registration (i.e., the ELEI that should no longer be used). Only one of the potential multiple identifiers will survive; for all other duplicate registrations: 1. The RegistrationStatus is set to DUPLICATE, 2. The ELEI of the surviving ELEI Registration is set in the SuccessorELEI data element of (each) duplicate ELEI registration; 3. The LastUpdateDate is set to reflect the date of this update, and 4. No further updates of the DUPLICATE registration record will occur.

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

Code	Definition
MERGED	An ELEI registration for a entity that has been merged into another entity, such that this entity no longer exists as an operating entity. If • After being issued an ELEI, the entity is acquired by, or merged with, another entity; • Per agreements among the parties to the transaction, the ELEI of the acquired or merged entity will not be used to identify the surviving entity (or if a new entity is created that is issued a new ELEI) Then 1. The ELEIRegistrationStatus is set to "MERGED", 2. The ELEI of the surviving/new entity is set in the successorELEI data element of (each) ELEI registration that is no longer to be used; 3. The ELEIRecordLastUpdate is set to reflect the date of this update, and 4. No further updates of the MERGED registration record(s) will occur.
RETIRED	An ELEI registration for a entity that has ceased operation, without having been merged into another entity. If • The responsible DSU determines by public sources that the entity has been dissolved or ceased to operate (and the DSU seeks to confirm this status through all available channels) Then 1. The ELEIRegistrationStatus is set to "RETIRED"; 2. The ELEIRecordLastUpdate is set to reflect the date of this update; 3. The EntityExpirationDate is also set to the date of this update; 4. The EntityExpirationReason is set; 5. No further updates of the RETIRED registration record will occur.

3.5.9 ValidationSourcesEnum Code List

A value of type ValidationSourcesEnum in a file conforming to this standard SHALL be one of the code strings specified in the following table:

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

Code	Definition
PENDING	The validation of the reference data for the entity has not yet occurred.
SUBMITTED	Based on the validation procedures in use by the DSU responsible for the record, the information associated with this record has significant reliance on the information that a submitter provided due to the unavailability of corroborating information.
PARTIALLY_CORROBORATED	Based on the validation procedures in use by the DSU responsible for the record, the information supplied for the entity can be partially corroborated by available sources, while some of the record is dependent upon information that the submitter collected, either due to conflicts with authoritative information, or due to data unavailability.
FULLY_CORROBORATED	Based on the validation procedures in use by the DSU responsible for the record, there is sufficient information contained in authoritative public sources to corroborate the information regarding the entity provided in the record.

.

4 Constraints and Data Validation

All values of type `String` specified in Section 3 SHALL be 500 or fewer characters in length.

- TBD. Specify here any additional constraints not implicit in the data types or explicit in the definition of each reference data element

5 XML Syntax

This section specifies the XML schema for an ELEI data file conforming to this standard.

5.1 XML Design Rules

- The XSD schema conforms to [XSD1,XSD2]
- The XML namespace is `http://www.EURHISFIRM.eu/schema/ELEIdata/1`
- All interior elements and attributes are namespace-qualified (element/attribute form = qualified)
- Element names are upper camel case
- Attribute name are lower camel case
- XSD type names are upper camel case
- Enumeration code list values are all caps with underscores

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

- Elements are used in preference to attributes *except* for language and type qualifiers
- For a data element specified in Section 3 as having unbounded cardinality, the XML includes a single container element whose sub-elements are one or more instances of the data element whose cardinality is unbounded. The name of the container element is formed as the plural of the name of the contained elements.

➢ TBD: specify the approach to enumeration types, recognizing the need for change management. See https://www.ibm.com/developerworks/library/x-extenum/ for various possible approaches.

➢ TBD: finalize the XML namespace

➢ TBD: take a final decision on whether to use element/attribute form = qualified or unqualified

6 Change Management

➢ TBD. Use the <vnext> strategy for forward/backward compatibility, and/or include a version number in the XML. Needs to be explained

7 Examples (non-normative)

➢ TBD.

8 References

[ISO646] ISO, "Information technology -- ISO 7-bit coded character set for information interchange," ISO/IEC 646:1991.

[ISO3166-1] ISO, "Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes," ISO 3166-1:2013.

[ISO3166-2] ISO, "Codes for the representation of names of countries and their subdivisions – Part 2: Country subdivision code," ISO 3166-2:2013.

[ISO8601] ISO, "Data elements and interchange formats – Information interchange – Representation of dates and times," ISO 8601:2004.

[ISO10646] ISO, "Information technology -- Universal Coded Character Set (UCS)," ISO 10646:2012.

[ISO17422] ISO, "Financial Services - Entity Identifier (LEI)," ISO/DIS 17442:2012.

[ISODir2] ISO, "Rules for the structure and drafting of International Standards (ISO/IEC Directives, Part 2, 2001, 4th edition)," July 2002.

[XSD1] H. Thompson, D. Beech, M. Maloney, N. Mendelsohn, "XML Schema Part 1: Structures," W3C Recommendation, May 2001, http://www.w3.org/TR/xmlschema-1/.

[XSD2] P. Biron, A. Malhotra, "XML Schema Part 2: Datatypes," W3C Recommendation, May 2001, http://www.w3.org/TR/xmlschema-2/.

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

9 Appendix: Character Codes Allowed in Romanized Names

When a `Name` instance is of type `PREFERRED_ROMANIZED` or `AUTO_ROMANIZED`, the value of the `name` field SHALL consist only of non-control characters drawn from the "invariant subset" of ISO 646. These characters are enumerated below. The "Hex Value" column indicates the code point value (expressed in hexadecimal) for each character in both ISO 646 and ISO 10646.

Graphic Symbol	Name	Hex Value	Graphic Symbol	Name	Hex Value
!	Exclamation Mark	21	M	Capital Letter M	4D
"	Quotation Mark	22	N	Capital Letter N	4E
%	Percent Sign	25	O	Capital Letter O	4F
&	Ampersand	26	P	Capital Letter P	50
'	Apostrophe	27	Q	Capital Letter Q	51
(	Left Parenthesis	28	R	Capital Letter R	52
)	Right Parenthesis	29	S	Capital Letter S	53
*	Asterisk	2A	T	Capital Letter T	54
+	Plus sign	2B	U	Capital Letter U	55
,	Comma	2C	V	Capital Letter V	56
-	Hyphen/ Minus	2D	W	Capital Letter W	57
.	Full Stop	2E	X	Capital Letter X	58
/	Solidus	2F	Y	Capital Letter Y	59
0	Digit Zero	30	Z	Capital Letter Z	5A
1	Digit One	31	_	Low Line	5F
2	Digit Two	32	a	Small Letter a	61
3	Digit Three	33	b	Small Letter b	62
4	Digit Four	34	c	Small Letter c	63
5	Digit Five	35	d	Small Letter d	64
6	Digit Six	36	e	Small Letter e	65
7	Digit Seven	37	f	Small Letter f	66
8	Digit Eight	38	g	Small Letter g	67
9	Digit Nine	39	h	Small Letter h	68
:	Colon	3A	i	Small Letter i	69
;	Semicolon	3B	j	Small Letter j	6A

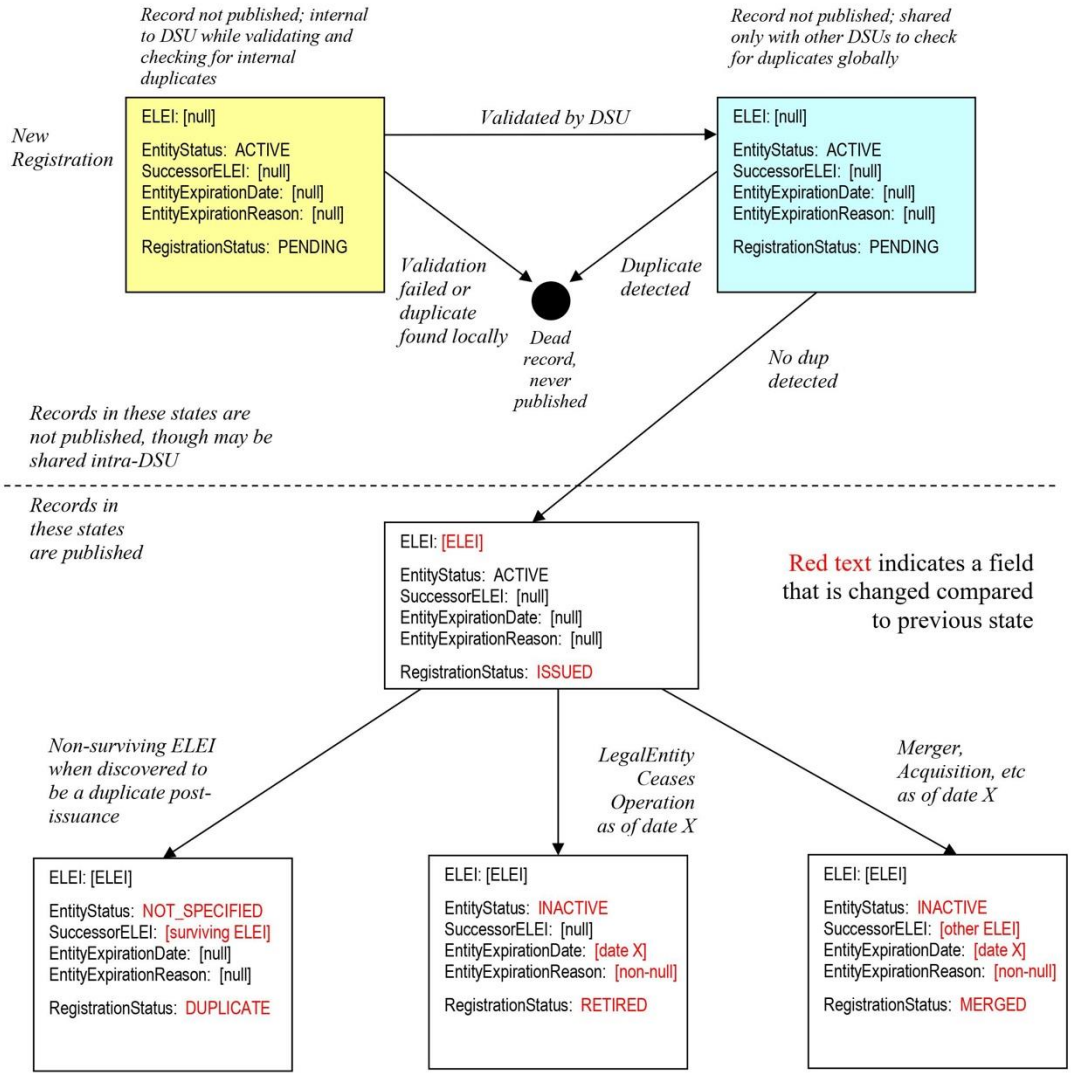
This project has received funding from
the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

Graphic Symbol	Name	Hex Value	Graphic Symbol	Name	Hex Value
<	Less-than Sign	3C	k	Small Letter k	6B
=	Equals Sign	3D	l	Small Letter l	6C
>	Greater-than Sign	3E	m	Small Letter m	6D
?	Question Mark	3F	n	Small Letter n	6E
A	Capital Letter A	41	o	Small Letter o	6F
B	Capital Letter B	42	p	Small Letter p	70
C	Capital Letter C	43	q	Small Letter q	71
D	Capital Letter D	44	r	Small Letter r	72
E	Capital Letter E	45	s	Small Letter s	73
F	Capital Letter F	46	t	Small Letter t	74
G	Capital Letter G	47	u	Small Letter u	75
H	Capital Letter H	48	v	Small Letter v	76
I	Capital Letter I	49	w	Small Letter w	77
J	Capital Letter J	4A	x	Small Letter x	78
K	Capital Letter K	4B	y	Small Letter y	79
L	Capital Letter L	4C	z	Small Letter z	7A
	Space	20			

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

10 Appendix: ELEI Record Transition Diagram

The following diagram illustrates the relationship between status fields of the ELEI record.



In rare circumstances, a record may pass directly from `PENDING` to `RETIRED` or `MERGED`; e.g., retroactive reporting of a failed or merged legal g. In that case, the `EntityStatus` would be `INACTIVE` while the record is in the `PENDING` state.

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

11 Appendix: ELEI Code Partitioning Scheme

The following diagram illustrates the partitioning scheme of the ELEI code. The ELEI code is an ISO 17442-compliant, opaque identifier that resolves to the EURHISFIRM entity reference data that uniquely identifies a entity in the EURHISFIRM network.

The ELEI is minted by a Data Submission Unit (DSU) as part of the process that would add the ELEI -- and the unique-identifying reference data -- for a entity to the EURHISFIRM network under the EURHISFIRM Common Data Model (CDM) data standard.

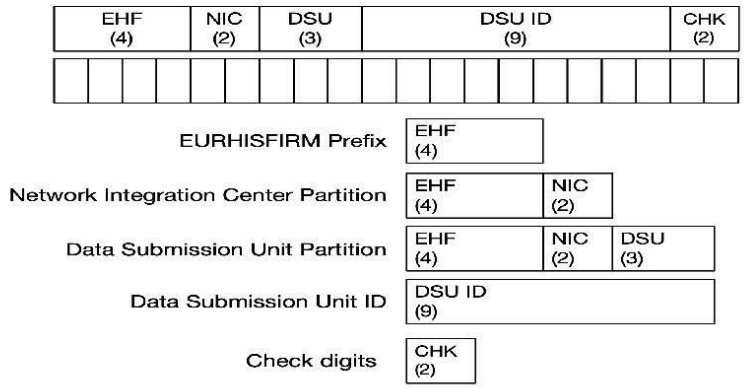
The DSU should, if possible, perform a check with the consolidated database of existing EURHISFIRM ELEI identifier reference data to verify that a entity has not already been identified and been assigned an ELEI before submitting a newly minted ELEI to a Network Integration Center (NIC). Otherwise, the Network Integration Center would need to perform this check for the existence of an ELEI that had already been assigned to the entity in question in order to maintain the uniqueness of the ELEI assigned to the entity.

In cases where the NIC may not be able to do this, or in which a duplicative entry existed but was not found, remediation and cleanup of multiple ELEIs for the same entity could be detected and rectified by background data quality processes that periodically sweep the ELEI reference data or by challenges submitted from the EURHISFIRM community.

Note that the partitioning scheme does not add intelligence or actually identify any DSU or NIC, as multiple prefixes could be issued to a single DSU under one of multiple prefixes that may be issued to a NIC. The partitioning scheme is only for the purpose of guaranteeing the ability to concurrently create unique ELEI codes in a federated and decentralized entity data "supply chain" in the EURHISFIRM network. (The EHF is the prefix that would be assigned to EURHISFIRM by the Global LEI Foundation in order to be globally unique in the Global LEI System)

EURHISFIRM Legal Entity Identifier (ELEI)
(ISO 17742 Compliant)

Code Partitioning Scheme



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

7.2. Financial Instrument Identification Data Standard 1.05

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

EURHISFIRM Common Data Model Standard 1.0

Financial Instrument Identification Data

Version of 22 July 2020

Date	Revision	Description
2020-05-30	ver 1.01	First Draft to be circulated (formatting fixed)
2020-06-20	ver 1.02	Multiple revisions and edits in response to comments
2020-07-07	ver 1.03	Market Sector enumeration data type added. DataMaturityStage enumeration type added Additional revisions and updates
2020-07-09	ver 1.04	Revisions to DataMaturityStage enumerations
2020-07-20	ver 1.05	Acceptance of changes

Abstract

First, the semantic content of these attributes must be fully specified. Second, some additional elements, such an indication of the status of the information, are necessary for effective use of the data. Third, the form the information takes at any given local point of source data capture must be such that it can be made to conform to a common standard, which must also be specified. This document proposes the standards necessary in these areas to support the EURHISFIRM Common Data Model.

Status of this document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. The latest status of this document series will be maintained on SeaFile.

This draft is a **Working Draft** which can be circulated to any interested parties for review and comment. It is a draft document and may be updated, replaced or made obsolete by other documents at any time. It is inappropriate to use Working Drafts as reference material or to cite

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

them as other than "work in progress." This is work in progress and does not imply endorsement by the EURHISFIRM ExCo.

Comments on this document should be sent to [TBD: insert mailing list or URL].

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

Table of Contents

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

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

1 Introduction

Financial Instruments, and in particular market securities that have been publicly traded on exchanges, are clearly one of the core classes of objects in a historical financial Research Infrastructure such as EURHISFIRM, on which a substantial amount of other EURHISFIRM financial data depend. Such financial instruments should be uniquely and unambiguously identified in EURHISFIRM, and this identification is accomplished based on the identifying attributes associated with each financial instrument.

First, the semantic content of those attributes must be fully specified. Second, some additional elements, such an indication of the status of the information, are necessary for effective use of the data. Third, the form the information takes at any given local point of source data capture must be such that it can be made to conform to a common standard, which must also be specified. This document proposes the standards necessary in these areas to support the reference data attributes of financial instruments in the EURHISFIRM Common Data Model

This document proposes initial standards for EURHISFIRM financial instrument reference data. It is important that this reference data should uniquely identify the financial instruments that are harvested from contributing sources and then collected and assimilated into the common EURHISFIRM platform.

A EURHISFIRM Financial Instrument Identifier (EFII) code that resolves to this financial instrument-identifying reference data is also introduced.

- A classification hierarchy for the structure of this identifier, modeled after the Object Management Group Financial Instrument Global Identifier (FIGI) is defined that provides for the identification of publicly traded market instruments at three levels: a unique identifier of a financial instrument traded at a given exchange, an identifier of a financial instrument traded within a country (national jurisdiction), and an identifier for a financial instrument issued at a share class level[1] by the institution that created the instrument. This allows for the independent and concurrent minting and assignment of EFII codes to financial instrument reference data that is collected by multiple Data Collection Units (DCUs) working at the exchange level, and the subsequent global identification of the financial instrument in the federated EURHISFIRM Research Infrastructure network.

The standard set by this document is expected, among other things, to reduce the risk of duplicates stemming from differences in formats and conventions of locally sourced data provided by regional contributors (i.e., Data Collection Units, or DCUs), to ensure data quality in the EURHISFIRM system, and to enable the subsequent detection and resolution of multiple identifiers for the same financial instrument traded on different national exchanges. The standard is expected to be used as a format for reference data consolidated from all sources in order to be promoted to the level of common EURHISFIRM data published for end-user access.

The contents of this document are as follows:

[1] "Share class" should not be interpreted as limited only to securities with "shares" – it is instead intended to distinguish among different unique securities offerings made by issuers. This is just the terminology that FIGI uses, but we could provide clarifying language or terms that would not be misinterpreted.

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

- Section 2 defines terminology and typographical conventions.
- Section 3 specifies the abstract content of EURHISFIRM financial instrument reference data conforming to this standard, including a detailed description of each element of reference data associated with financial instruments in EURHISFIRM. Allowable values for data elements that are code lists will also be (subsequently) specified. The partitioning scheme for the structure of the EURHISFIRM Financial Instrument Identifier (EFII) code will also be subsequently described.

2 Terminology and Typographical Conventions

Within this specification, the terms SHALL, SHALL NOT, SHOULD, SHOULD NOT, MAY, NEED NOT, CAN, and CANNOT are to be interpreted as specified in Annex G of the ISO/IEC Directives, Part 2, 2001, 4th edition [ISODir2]. When used in this way, these terms will always be shown in ALL CAPS; when these words appear in ordinary typeface they are intended to have their ordinary English meaning.

All sections of this document, with the exception of Section 1 are normative, except where explicitly noted as non-normative.

The following typographical conventions are used throughout the document:

- ALL CAPS type is used for the special terms from [ISODir2] enumerated above.
- `Monospace` type is used to denote programming language, UML, and XML identifiers, as well as for the text of XML documents.

➢ Placeholders for open issues and/or changes that need to be made to this document prior to its reaching the final stage of approved Proposed Standard are prefixed by a rightward-facing arrowhead, as this paragraph is.

The specifications of data types and elements include a column for "cardinality" (usually abbreviated 'CARD') that describes if the element is either mandatory or optional, and whether the number of occurrences of the element can be more than one, or just one.

3 Abstract Data Content

This section specifies the abstract data content of a data file conforming to this standard.

A data file conforming to this standard SHALL consist of:

- An optional EFII File Header, as specified in Section 3.1.
- An EFII Root Record as specified in Section 3.2.
- Zero or more EFII Data Records, as specified in Section 3.3.

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

3.1 EFII File Header

- TBD Define elements to go into a "header" area for an EFII file. The purpose of the header is to provide context about the file and its contained EFII data records. However, the header will not contain anything necessary to interpret the meaning of any EFII record; e.g., things like default values for EFII data records will not be in the header (such things would mean that the meaning of an EFII record could change if taken away from the header). Examples of things that might be useful to include in the header:

Element Name	Type	Card	Description
ContentDate	DateTime	1,1	The date and time of generation of the data
Originator		0,1	The identifier of the creator of the content of this file
FileContent		1,1	A code describing the content of this data file.
ProcessStage		0,1	A code indicating the stage of this file in the EURHISFIRM workflow
RecordCount		1,1	The number of data records in the file. Can be a positive whole (integer) number, or zero (0).

3.2 EFII Root Record

The EFII Root Record anchors a single EFII. This root record is needed in order to provide a unique EFII anchor to support multiple (historical) EFII Data Records should there be changes in aspects of the security or its reference data (e.g., name change, stock split, etc) that require keeping a record. Each EFII Root record in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
EFII	EFII	1	The 12-character EFII of the financial instrument described by this EFII Data Record.
EFIICreationDate	DateTime	1	Date/time the EFII root record was initially created in the system
FinancialInstrumentStatus	FinancialInstrumentStatusEnum	1	The status of the financial instrument. This is not to be confused with the status of the registration; see RegistrationStatus.

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

Element Name	Type	Card	Description
FinancialInstrumentIssueDate	DateTime	0,1	Date/time the Financial Instrument was issued, if known
FinancialInstrumentExpirationDate	DateTime	0,1	The date that the financial instrument ceased to exist, whether due to dissolution, merger or acquisition of the issuing firm, or the maturity, redemption or other action in the terms of the security. Omitted if the financial instrument has not ceased to exist, or if this EFII record contains a non-empty SuccessorEFII field.
FinancialInstrumentExpirationReason	FinancialInstrumentExpirationReasonEnum	0,1	The reason that a financial instrument ceased to exist. This element SHALL be present if FinancialInstrumentExpirationDate is present, and omitted otherwise.
EarliestRegistrationDate	DateTime	1	Date/time of the Registration record with the earliest historical date
LatestRegistrationDate	DateTime	1	Date/time of the Registration record with the most recent historical date

3.3 EFII Data Record

An EFII Data Record describes a single EFII. Each EFII Data record in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
EFII	EFII	1	The 12-character EFII of the financial instrument described by this EFII Data Record.
FinancialInstrument	FinancialInstrument (Section 3.3.1)	1	Attributes describing the financial instrument itself
Registration	Registration (Section 0)	1	Attributes describing the registration of this EFII.
Extension	Extension (Section 0)	0,1	An optional element for including data beyond the standard data elements in an EFII data file. This may include data specific to an DSU, data specific to a publisher of EFII data, and so on.

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

3.3.1 Financial Instrument Section of EFII Data Record

The `Instrument` section of an EFII Data Record in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`InstrumentName`	`Name`	1,1	The Registered Name of the Instrument, exactly as it appears on the official list. If an Instrument is in a jurisdiction with more than one Registered Name (e.g., in different languages), this is the Primary Registered Name (see `OtherInstrumentName` for other names).
`OtherInstrumentName`	`Name`	0,n	Other registered names of the financial instrument, whether in different languages in the jurisdiction of the legal issuer, or in different languages used in exchanges in foreign jurisdictions. `OtherInstrumentNameType` enumerated attributes will qualify the type of each `OtherInstrumentName`
`Ticker`	`Name`		Ticker is a specific identifier for a financial instrument that reflects common usage. Tickers are not, however, unique to specific exchanges or specific pricing sources. Tickers may change in conjunction with Corporate Actions.
`ExchangeCode`	`ExchangeCodeEnum`		Code for the trading venue or environment on which the instrument trades. If an exchange is specified, the code will be for the specified exchange. When not specified, the code will be according to the user default exchange, which can be the composite or primary exchange.
`SecurityType`	`SecurityTypeEnum`		Classification (enumerated type) of the instrument type within its market sector.

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

Element Name	Type	Card	Description
MarketSector	MarketSectorEnum		Market Sector refers to the classification of the asset type of the instrument. (For example: Commodity, Equity, Municipals ,Preferred, Money Market, Government, Corporate, Index , Currency, Mortgage)
LocalEFII	EFII		Twelve character, alphanumeric identifier. The first 2 characters are upper-case consonants (including "Y"), the third character is the upper-case "G", characters 4 -11 are any upper-case consonant (including "Y") or integer between 0 and 9, and the last character is a check-digit. An identifier is assigned to instruments of all asset classes, is unique to an individual instrument and once issued will not change for an instrument. For equity instruments an identifier is issued per instrument per trading venue (e.g., stock exchange or other marketplace with listed securities).
CompositeEFII			Twelve character, alphanumeric identifier. The first 2 characters are upper-case consonants (including "Y"), the third character is the upper-case "G", characters 4 -11 are any upper-case consonant (including "Y") or integer between 0 and 9, and the last character is a check-digit. The Composite level of assignment is provided in cases where there are multiple trading venues for the instrument within a single country or market. The Composite Eurhisfirm Financial Instrument Identifier (EFII) enables users to link multiple EFIIs at the trading venue-level within the same country or market in order to obtain an aggregated view for that instrument within that country or market.

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

Element Name	Type	Card	Description
ShareClassEFII			Twelve character, alpha-numeric identifier. The first 2 characters are upper-case consonants (including "Y"), the third character is the upper-case "G", characters 4 -11 are any upper-case consonant (including "Y") or integer between 0 and 9, and the last character is a check-digit. A Share Class level Eurhisfirm Financial Instrument Identifier is assigned to an instrument that is traded in more than one country. This enables users to link multiple Composite EFIIs for the same instrument in order to obtain an aggregated view for that instrument across all countries globally.
SecurityDescription			A description of the security
SecurityShortDescription			Alternate Short Description for a given security comprised of the ticker, coupon and maturity year (YY). For strips/scripts it returns the ticker, coupon, and maturity (M/YY).
OtherInstrumentNames	OtherInstrumentName	0..n	An optional list of other Name instances for the Instrument.
LegalJurisdiction	ERegionCode	0,1	The jurisdiction of legal issuance and registration of the financial instrument (and on which the legalForm data element is also dependent).
InstrumentCategory	InstrumentCategoryTypeEnum	0,1	Indicates the general category of the type of financial instrument identified by this EFII data record
FinancialInstrumentStatus	FinancialInstrumentStatusEnum	1	The status of the financial instrument. This is not to be confused with the status of the registration; see RegistrationStatus.
FinancialInstrumentEvents	FinancialInstrumentEvent	0,n	Corporate events that occurred during this historical period

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

3.3.2 Registration Section of EFII Data Record

The `Registration` section of an EFII Data Record in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`RegistrationCreationDate`	`DateTime`	1	Date/time this EFII record was initially created in the system
`RegistrationUpdateDate`	`DateTime`	1	Date/time that this historical EFII record was most recently updated in the system.
`RegistrationSource`	`RegistrationSourceType`	0,n	A locator of the source of the financial instrument reference data in this registration record (mechanism TBD, typically company yearbooks, official price lists, and exchange registrations)
`RegistrationStatus`	`RegistrationStatusEnum`	1	Status of the EFII registration. This is not to be confused with the status of the financial instrument itself; see `FinancialInstrumentStatus`.
`RegistrationUpToDate`	`DateTime`	0,1	Should it exist and be known, the historical date up until which time the EFII identifying reference data for this financial instrument is valid. (Not the same thing as when the data was changed in the system). If present, a subsequent historical record for this EFII may exist with the revised information.

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

Element Name	Type	Card	Description
RegistrationAsOfDate	DateTime	0,1	The date that this version of the reference data is known to be valid. This allows a record of historical changes to the financial instrument identifying data to be recorded. This date would typically be derived from either: 1. The publication date of the source reference, or 2. Information contained in the source reference
ResponsibleDSU	DSUID	1	The Identifier of the Data Submission Unit (DSU) that produced and manages this EFII registration.
DataMaturityStage	DataMaturityStageEnum	0,1	The current data maturity stage of this EFII record, or omitted if the data maturity stage is not known.

3.3.3 Extension Section of EFII Data record

The Extension section of an EFII record may be used to include additional data not defined in this standard. For example, a DSU may use Extension to publish additional data elements it collects as part of registration.

- TBD: include the details of how this works. Basically, the idea is to use an XSD schema wildcard with namespace ##other, permitting the inclusion of XML elements from other XML namespaces.

3.4 Data Types

This section specifies the data types referenced by the tables in Section 3.2, in alphabetical order.

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

3.4.1 DateTime Data Type

A value of type `DateTime` in a conforming to this standard SHALL be a point in time expressed as a string conforming to ISO 8601 having the following format:

- TBD: Provision for different calendars ? (i.e., Gregorian, Julian, etc.)

YYYY`-`*MM*`-`*DD*`T`*hh*`:`*mm*`:`*ss.sss*`TZ`

where the components of the above string are as follows:

- *YYYY* is the year
- *MM* is the month (01 = January, …, 12 = December)
- *DD* is the day of the month (01 = first day of the month)
- `T` is the single character 'T'
- *hh* is the hour (00 – 23)
- *mm* is the minute
- *ss.sss* is the second and milliseconds. From one to three digits may be used for milliseconds, or omitted entirely along with the decimal point.
- *TZ* is the time zone specifier, which can be either:
  - `Z` the single character 'Z', denoting Coordinated Universal Time (UTC); or
  - *+hh:mm* denoting a positive offset from UTC; or
  - *-hh:mm* denoting a negative offset from UTC

In the XML representation specified in Section **Error! Reference source not found.**, the XSD type `xs:dateTime` is used; however, whereas `xs:dateTime` permits the time zone specifier to be omitted, `DateTime` values in files conforming to this standard SHALL always include a time zone specifier.

Explanation (non-normative): milliseconds are hardly necessary for EFII reference data, and likewise it might seem simpler to allow only "Z" as a time zone specifier; however, XML processing tools support the full syntax given above and it is not always possible to restrict such tools to avoid milliseconds or force the use of "Z" as the time zone specifier. The restriction that the time zone specifier must be present is equivalent to using XSD type `xs:timestamp`; however this was introduced in XSD 1.1 and not supported by the majority of XML processing tools which still only implement XSD 1.0.

3.4.2 DCUID Data Type

A value of type `DCUID` in a file conforming to this standard SHALL be a [TBD]-character Data Collection Unit Identifier conforming to [TBD].

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

3.4.3 DSUID Data Type

A value of type `DSUID` in a file conforming to this standard SHALL be a [TBD]-character Data Submission Unit Identifier conforming to [TBD].

3.4.4 ECountryCode Data Type

A value of type `CountryCode` in a file conforming to this standard SHALL be a 2-character country code conforming to ISO 3166-1 alpha-2 [ISO3166]. Note that ISO 3166-1 alpha-2 codes are all uppercase.

- TBD: CountryCode external code list must extend ISO country codes with start/stop dates and historical changes

3.4.5 ERegionCode Data Type

A value of type `RegionCode` in a file conforming to this standard SHALL be a code conforming to ISO 3166-2. Note that ISO 3166-2 codes are all uppercase.

- TBD: Like CountryCode code list, RegionCode external code list needs historical extensions and additions as well.

3.4.6 ERegistrationAuthorityType Data Type

A value of type `ERegistrationAuthority` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`ERegistrationAuthorityID`	`ERegistrationAuthorityEnum`	0,1	An identifier for the financial instrument **registry** of the financial instrument in the jurisdiction of legal registration, or in the appropriate registration authority.

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

Element Name	Type	Card	Description
EOtherRegistrationAuthority	String	0,1	A legacy / historical reference code of a registration authority which is not yet entered in the ERegistration Authorities List (RAL), or the designation of an interim register until such time as an entry from RAL can be delivered .

3.4.7 EFII Data Type

A value of type EFII in a file conforming to this standard SHALL be a 12-character Financial Instrument Identifier conforming to [OMG FIGI].

3.4.8 FinancialInstrumentCategory Data Type

A value of type FinancialInstrumentCategory in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
FinancialInstrumentCategory	FinancialInstrumentCategoryTypeEnum	0,1	Indicates the general category of the type of financial instrument identified by this EFII data record

3.4.9 LanguageCode Data Type

A value of type LanguageCode in a file conforming to this standard SHALL be a 2-character language code conforming to [ISO639-1]. Note that ISO 639-1 language codes are all lowercase.

- TBD: Alternatively, we could use IETF language tags (RFC 4646) instead. IETF language tags can distinguish between variations of the same language in different countries. For example, the IETF language tags fr-CA and fr-FR denote Canadian French and French as spoken in France, respectively; in ISO 639-1 these would both be simply fr.

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

3.4.10 Name Data Type

A `Name` is a string expressed in a natural language, including a code indicating which natural language is used.

A value of type `Name` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`lang`	`LanguageCode`	0,1	The language of `name`
`Name`	`String`	1	The name itself.

3.4.11 OtherFinancialInstrumentName Data Type

A value of type `OtherFinancialInstrumentName` in a file conforming to this standard SHALL include data elements as specified below. Each `Name` element includes an optional language code, permitting `OtherFinancialInstrumentName` to be repeated as many times as necessary to express the same name type in multiple languages. When `type` is `PREFERRED_ROMANIZED_LEGAL` or `AUTO_ROMANIZED_LEGAL`, the language code specifies the language of the name prior to Romanization.

Element Name	Type	Card	Description
`type`	`FinancialInstrument NameTypeEnum`	1	The type of name represented by this `OtherFinancialInstrumentName` instance. The FinancialInstrumentNameType observes language, since 'Name' type has a language attribute.
`Name`	`Name`	1	The name. If `type` is `PREFERRED_ROMANIZED_LEGAL` or `AUTO_ROMANIZED_LEGAL`, then this value SHALL only include characters from the character set specified in Section 7.

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

3.4.12 RegistrationSource Data Type

RegistrationSourceCitation	RegistrationSourceCitationType	0,1	A locator of the source of the financial instrument reference data in this registration record (mechanism TBD, typically company yearbooks, official price lists, or exchange registrations)
RegistrationSourceLocatorKey	RegistrationSourceLocatorKeyType	0,1	A locator of the source of the financial instrument reference data in this registration record (mechanism TBD, typically company yearbooks)
RegistrationSourceLocatorAnchor	RegistrationSourceLocatorAnchorType	0,1	Additional information to allow pinpointing where in the RegistrationSource of the financial instrument data can be found
RegistrationSourceDCU	DCUID	0,1	The identifier of the Data Collection Unit that sourced the registration data
RegistrationSourceInstrumentID	String	0,1	An identifier of the financial instrument that may have been locally used or assigned by the DCU at collection time

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

3.4.13 RegistrationSourceCitation Data Type

An associative locator (link or key) that resolves to a DDI metadata block regarding the source of the financial instrument reference data in this registration record (mechanism TBD, typically company yearbooks -)

DDI 3.2 uses the `CitationType` Element for referencing publications (e.g. yearbooks). It contains 11 elements which are described in detail in the *DDI Lifecycle XML Schema*. The table below lists the suggested cardinality for each element and some notes on special uses. The name of each element is a link to its description in the DDI XML Schema.

Element Name	Type	Card	Description
Title	DDI Title	1,1	
SubTitle	DDI SubTitle	0,1	
AlternateTitle	DDI AlternateTitle	0,n	
Creator	DDI Creator	0,n	
Publisher	DDI Publisher	0,1	
Contributor	DDI Contributor	0,n	
PublicationDate	DDI PublicationDate	0,1	For monographs: SimpleDate For serials (e.g. yearbooks): StartDate and EndDate (the date/year when the first and the last volume of a serial were published – the specific volume and page numbers of the year from which information about the financial instrument was taken can then be recorded in the RegistrationSourceLocatorAnchor element)
Language	DDI Language	0,n	
InternationalIdentifier	DDI InternationalIdentifier	0,n	
Copyright	DDI Copyright	0,1	
dc:isPartof	DDI dc:isPartof	0,1	Use for instance to record the title of the newspaper in which a stock exchange price list is published

The Citation as we would want to use it in the Financial Instrument Reference Data would be considered as a DataSource in DDI 3.2 (Origin element).

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

3.4.14 RegistrationSourceLocatorAnchorType Data Type

The `RegistrationSourceLocatorAnchorType` provides additional information to allow pinpointing where in the RegistrationSource referenced by the `RegistrationSourceLocatorKey` can the financial instrument data be found

3.4.15 RegistrationSourceLocatorKeyType Data Type

An associative locator (link or key) that resolves to a DDI metadata block regarding the source of the financial instrument reference data in this registration record (mechanism TBD, typically company yearbooks -)

3.5 Enumerated Code Lists

This section specifies the enumerated code list data types (all having the suffix `Enum`) referenced by the tables in Sections 3.2 and 0, in alphabetical order.

3.5.1 FinancialInstrumentCategoryTypeEnum

The `FinancialInstrumentCategoryTypeEnum` value in an EFII record specifies the general classification category of the financial instrument.

- TBD: the table below is just an example with some possible categories, and is incomplete, and needs to be elaborated.

A value of type `FinancialInstrumentCategoryTypeEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
	To Be Determined

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

3.5.2 FinancialInstrumentExpirationReasonEnum Code List

The `FinancialInstrumentExpirationReasonEnum` value in an EFII record specifies the reason that the financial instrument expired.

A value of type `FinancialInstrumentExpirationReasonEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
DISSOLVED	The financial instrument ceased to exist because the issuing entity was dissolved, and was legally shuttered.
CORPORATE_ACTION	The financial instrument was merged with another financial instrument due to an acquisition or merger of the issuing entity.
MATURITY	The terms of the financial instrument provide for its expiration and settlement
OTHER	The reason for expiry is not one of the above

3.5.3 FinancialInstrumentNameTypeEnum Code List

The `FinancialInstrumentNameTypeEnum` value in an `FinancialInstrumentName` specifies how the name relates to the financial instrument.

A value of type `FinancialInstrumentNameTypeEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
REGISTERED_NAME	Primary Registered Name of the security as designated by the issuer
OTHER_REGISTERED_NAME	Registered name of the financial instrument in an alternate language in the legal jurisdiction in which the financial instrument is registered
FOREIGN_REGISTERED_NAME	Primary Registered Name of the financial instrument on a trading venue in a sovereign jurisdiction that is not the jurisdiction where the financial instrument is registered
OTHER_FOREIGN_REGISTERED_NAME	Alternate Registered Name of the financial instrument on a trading venue in a sovereign jurisdiction that is not the jurisdiction where the financial instrument is registered

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

Code	Definition
AUTO_ROMANIZED_LEGAL	Romanized form, auto-transliterated

3.5.4 FinancialInstrumentStatusEnum Code List

The `FinancialInstrumentStatusEnum` value in an EFII record indicates the status of the financial instrument itself. This is not to be confused with the status of the EFII registration, which is specified by `RegistrationStatusEnum` (Section 3.4.8). See also Section **Error! Reference source not found.**, which illustrates how the `FinancialInstrumentStatusEnum` value changes over the lifecycle of an EFII registration.

A value of type `FinancialInstrumentStatusEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
ACTIVE	As of the last report or update, the financial instrument reported is legally registered and operating.
INACTIVE	It has been determined that the financial instrument that was assigned the EFII is no longer legally registered and/or operating, whether as a result of: 1. Business closure 2. Acquisition by or merger of the issuing institution with another (or new) financial instrument 3. Contractual term or maturity reached (e.g., options or bonds) 4. Determination of illegitimacy [perhaps not required for historical entities]
NOT_SPECIFIED	The EFII record is in a state in which it does not provide information about whether the financial instrument is legally registered and operating.

3.5.5 MarketSectorEnum Code List

Code signifying the market sector of the security.

Code	Definition
COMMODITY	
EQUITY	

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

Code	Definition
MUNICIPALS	
PREFERRED	
MONEY_MARKET	
GOVERNMENT	
CORPORATE	
INDEX	
CURRENCY	
MORTGAGE	

3.5.6 RegistrationStatusEnum Code List

The `RegistrationStatusEnum` value in an EFII record indicates the status of the registration of the financial instrument with an DSU. This is not to be confused with the status of the EFII itself, which is specified by `FinancialInstrumentStatusEnum` (Section 3.5.4). See also Section **Error! Reference source not found.**, which illustrates how the `FinancialInstrumentStatusEnum` value changes over the lifecycle of an EFII registration.

A value of type `RegistrationStatusEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
PENDING	An application for an EFII that has been submitted and which is being processed and validated. *NOTE: EFII registrations in the `PENDING` state are not intended for public release, but could be used internally between DSUs.*
ISSUED	An EFII Registration that has been validated and issued, and which identifies the financial instrument as of the last update.

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

Code	Definition
DUPLICATE	An EFII Registration that has been determined to be a duplicate registration of the same financial instrument as another EFII Registration; the DUPLICATE status is assigned to the non-surviving registration (i.e., the EFII that should no longer be used). Only one of the potential multiple identifiers will survive; for all other duplicate registrations: 1. The RegistrationStatus is set to DUPLICATE, 2. The EFII of the surviving EFII Registration is set in the SuccessorEFII data element of (each) duplicate EFII registration; 3. The LastUpdateDate is set to reflect the date of this update, and 4. No further updates of the DUPLICATE registration record will occur.
MERGED	An EFII registration for a financial instrument that has been merged into another financial instrument, such that this financial instrument no longer exists as an active financial instrument. If • After being issued an EFII, the financial instrument is merged with the financial instrument of another entity; • Per agreements among the parties to the transaction, the EFII of the acquired or merged financial instrument will not be used to identify the surviving financial instrument (or if a new financial instrument is created that is issued a new EFII) Then 1. The EFIIRegistrationStatus is set to "MERGED", 2. The EFIIRecordLastUpdate is set to reflect the date of this update, and 3. No further updates of the MERGED registration record(s) will occur.

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

Code	Definition
RETIRED	An EFII registration for a financial instrument that has ceased operation, without having been merged into another financial instrument. If • The responsible DSU determines by public sources that the financial instrument has ceased to exist (and the DSU seeks to confirm this status through all available channels) Then 1. The EFIIRegistrationStatus is set to "RETIRED"; 2. The EFIIRecordLastUpdate is set to reflect the date of this update; 3. The FinancialInstrumentExpirationDate is also set to the date of this update; 4. The FinancialInstrumentExpirationReason is set; 5. No further updates of the RETIRED registration record will occur.

3.5.7 DataMaturityStageEnum Code List

A value of type DataMaturityStageEnum in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
UNVERIFIED	(e.g., "Raw") Data has been acquired without undergoing any EURHISFIRM data quality measures
VERIFIED	(e.g., "Collected") Source-level metadata attribution and data collection measures have been performed. This maturity stage is what the output of a Data Collection Unit (DCU) would achieve.
COMPLIANT	(e.g., "Harmonized") Source-level data elements have been harmonized to EURHISFIRM field-level Common Data Model conventions. In DDI terms: source variables have been mapped to conceptual (i.e., semantic) standards. This maturity stage is typically associated with the output of a Data Submission Unit (DSU)
CONSOLIDATED	(e.g., "Reconciled")Data previously collected and harmonized is compared with other EURHISFIRM data (whether previously published or submitted by other DSUs) in order to identify multiple references to the same unique object (legal entity, issued security, etc) and resolve duplicate references (identifiers) to the same object.

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

Code	Definition
CONSOLIDATED	(e.g., "Promoted"). After reconciliation (deduplication / unique identification), data is promoted to the maturity stage of published Common Data Model EURHISFIRM data
EU REVISED	After being promoted, the data was revised or edited

.

4 Change Management

- TBD. Use the <vnext> strategy for forward/backward compatibility, and/or include a version number in the XML. Needs to be explained

5 Examples (non-normative)

- TBD.

6 References

[ISO646] ISO, "Information technology -- ISO 7-bit coded character set for information interchange," ISO/IEC 646:1991.

[ISO3166-1] ISO, "Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes," ISO 3166-1:2013.

[ISO3166-2] ISO, "Codes for the representation of names of countries and their subdivisions – Part 2: Country subdivision code," ISO 3166-2:2013.

[ISO8601] ISO, "Data elements and interchange formats – Information interchange – Representation of dates and times," ISO 8601:2004.

[ISO10646] ISO, "Information technology -- Universal Coded Character Set (UCS)," ISO 10646:2012.

[ISO17422] ISO, "Financial Services - Legal Entity Identifier (LEI)," ISO/DIS 17442:2012.

[ISODir2] ISO, "Rules for the structure and drafting of International Standards (ISO/IEC Directives, Part 2, 2001, 4th edition)," July 2002.

[XSD1] H. Thompson, D. Beech, M. Maloney, N. Mendelsohn, "XML Schema Part 1: Structures," W3C Recommendation, May 2001, http://www.w3.org/TR/xmlschema-1/.

[XSD2] P. Biron, A. Malhotra, "XML Schema Part 2: Datatypes," W3C Recommendation, May 2001, http://www.w3.org/TR/xmlschema-2/.

7 Appendix: Character Codes Allowed in Romanized Names

When a Name instance is of type PREFERRED_ROMANIZED or AUTO_ROMANIZED, the value of the name field SHALL consist only of non-control characters drawn from the "invariant

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

subset" of ISO 646. These characters are enumerated below. The "Hex Value" column indicates the code point value (expressed in hexadecimal) for each character in both ISO 646 and ISO 10646.

Graphic Symbol	Name	Hex Value	Graphic Symbol	Name	Hex Value
!	Exclamation Mark	21	M	Capital Letter M	4D
"	Quotation Mark	22	N	Capital Letter N	4E
%	Percent Sign	25	O	Capital Letter O	4F
&	Ampersand	26	P	Capital Letter P	50
'	Apostrophe	27	Q	Capital Letter Q	51
(	Left Parenthesis	28	R	Capital Letter R	52
)	Right Parenthesis	29	S	Capital Letter S	53
*	Asterisk	2A	T	Capital Letter T	54
+	Plus sign	2B	U	Capital Letter U	55
,	Comma	2C	V	Capital Letter V	56
-	Hyphen/ Minus	2D	W	Capital Letter W	57
.	Full Stop	2E	X	Capital Letter X	58
/	Solidus	2F	Y	Capital Letter Y	59
0	Digit Zero	30	Z	Capital Letter Z	5A
1	Digit One	31	_	Low Line	5F
2	Digit Two	32	a	Small Letter a	61
3	Digit Three	33	b	Small Letter b	62
4	Digit Four	34	c	Small Letter c	63
5	Digit Five	35	d	Small Letter d	64
6	Digit Six	36	e	Small Letter e	65
7	Digit Seven	37	f	Small Letter f	66
8	Digit Eight	38	g	Small Letter g	67
9	Digit Nine	39	h	Small Letter h	68
:	Colon	3A	i	Small Letter i	69
;	Semicolon	3B	j	Small Letter j	6A
<	Less-than Sign	3C	k	Small Letter k	6B
=	Equals Sign	3D	l	Small Letter l	6C

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

Graphic Symbol	Name	Hex Value	Graphic Symbol	Name	Hex Value
>	Greater-than Sign	3E	m	Small Letter m	6D
?	Question Mark	3F	n	Small Letter n	6E
A	Capital Letter A	41	o	Small Letter o	6F
B	Capital Letter B	42	p	Small Letter p	70
C	Capital Letter C	43	q	Small Letter q	71
D	Capital Letter D	44	r	Small Letter r	72
E	Capital Letter E	45	s	Small Letter s	73
F	Capital Letter F	46	t	Small Letter t	74
G	Capital Letter G	47	u	Small Letter u	75
H	Capital Letter H	48	v	Small Letter v	76
I	Capital Letter I	49	w	Small Letter w	77
J	Capital Letter J	4A	x	Small Letter x	78
K	Capital Letter K	4B	y	Small Letter y	79
L	Capital Letter L	4C	z	Small Letter z	7A
	Space	20			

8

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

7.3. Legal Entity Data Artifact 1.05

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

EURHISFIRM Common Data Model Standard 1.0

Legal Entity Data Artifact (LEDA)

Version of 20 November 2020

Date	Revision	Description
2020-07-23	ver 1.01	First Draft to be circulated
2020-09-23	ver 1.02	Revisions to Header and Financial Data Elements
2020-10-29	ver 1.03	Major revisions
2020-11-01	ver 1.04	Additional formatting, revisions to EntityDataElement
2020-11-18	ver 1.05	Numerous revisions

Abstract

First, the semantic content of these attributes must be fully specified. Second, some additional elements, such an indication of the status of the information, are necessary for effective use of the data. Third, the form the information takes at any given local point of source data capture must be such that it can be made to conform to a common standard, which must also be specified. This document proposes the standards necessary in these areas to support the EURHISFIRM Common Data Model.

Status of this document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. The latest status of this document series will be maintained on SeaFile.

This draft is a **Working Draft** which can be circulated to any interested parties for review and comment. It is a draft document and may be updated, replaced or made obsolete by other documents at any time. It is inappropriate to use Working Drafts as reference material or to cite them as other than "work in progress." This is work in progress and does not imply endorsement by the EURHISFIRM ExCo.

Comments on this document should be sent to [TBD: insert mailing list or URL].

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

Table of Contents

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

1 Introduction

The economic activity of businesses and firms is, and has been, expressed in financial statements that periodically report on the state and status of the business operations and financial "health" of organizations. Such financial statements and reports (e.g., balance sheets and income statements) have, for centuries, used classifications and categories that conform to the accepted accounting standards in the jurisdiction and historical era in which the reports have been prepared.

Although there has definitely been an evolution and ongoing revision of a number of primary accounting standards in the modern era (e.g., GAAP, IFRS, and "Local GAAP"), the fundamentals of balance sheet and income statement reporting have, for the most part, and at a high level, been reasonably stable and consistent over the historical time frame that is of interest to the scope of EURHISFIRM.

However, although balance sheets and income statements (and cash flow statements) have generally well-accepted concepts at the highest level of summary, or consolidation, of financial data, this high-level understanding of such basic concepts as balance sheet assets, liabilities and equity -- or income statement concepts of income, expenses, earnings (profit) -- is too general to be considered any kind of standard definition of the structure of financial statements. Even so, the structure of most financial statements reflects, in varying degrees and levels of detail, the basic underlying principles of financial accounting that are common to most enterprises.

At the highest level of summary, there is little if any information about the type of economic activities, lines of business, or the products and services that form the basis and rationale of a particular firm, especially when a firm has multiple lines of business or products that are offered to the marketplace. Such information is almost always contained in lower levels of the financial statements in which taxonomies and categories of classification and categorization of types of business activities (business lines, business units, products and services) are themselves summarized and itemized.

The standardization of detailed financial reporting taxonomies (such as those established in recent times by the eXtensible Business Reporting Language, or XBRL) for the wide variety of companies in the industries and sectors of the economies of the different countries in Europe is outside the initial scope of the high-level standards for financial statements in the EURHISFIRM Common Data Model.

However, the economic data of businesses and firms contained in their historical financial records constitute a collection of the core elements of data in a historical financial Research Infrastructure such as EURHISFIRM, and as such is essential to the justification of the rationale to create the EURHISFIRM research infrastructure. Financial statement information (i.e., the labels associated with historical financial facts) should be semantically defined in EURHISFIRM, but – owing to the incomplete nature of many historical records -- it is not (initially) possible, nor necessarily advantageous, to standardize a normative structure of all line items and financial information that will or could be collected and recorded in the EURHISFIRM research infrastructure.

Hence, the further elaboration of the standardization of financial statement taxonomies will be an ongoing task that must follow the establishment of an initial framework and means to capture data contained in financial statements by first defining the types of data elements and facts that

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

are present in historical records that contain financial information that may or may not be found in the form of complete financial statements.

In addition to information contained in financial statements or other published financial data pertaining to legal entities, many other types of data (e.g., market prices, manufacturing output, macro economic data, etc. can also be captured in this format.

- .

2 Terminology and Typographical Conventions

Within this specification, the terms SHALL, SHALL NOT, SHOULD, SHOULD NOT, MAY, NEED NOT, CAN, and CANNOT are to be interpreted as specified in Annex G of the ISO/IEC Directives, Part 2, 2001, 4th edition [ISODir2]. When used in this way, these terms will always be shown in ALL CAPS; when these words appear in ordinary typeface they are intended to have their ordinary English meaning.

All sections of this document, with the exception of Section 1 are normative, except where explicitly noted as non-normative.

The following typographical conventions are used throughout the document:

- ALL CAPS type is used for the special terms from [ISODir2] enumerated above.
- `Monospace` type is used to denote programming language, UML, and XML identifiers, as well as for the text of XML documents.
- ➢ Placeholders for open issues and/or changes that need to be made to this document prior to its reaching the final stage of approved Proposed Standard are prefixed by a rightward-facing arrowhead, as this paragraph is.

The specifications of data types and elements include a column for "cardinality" (usually abbreviated 'CARD') that describes if the element is either mandatory or optional, and whether the number of occurrences of the element can be more than one, or just one.

3 Abstract Data Content

This section specifies the abstract data content of a data file conforming to this standard.

A data file conforming to this standard SHALL consist of:

- An optional LEDA File Header, as specified in Section 3.1.
- An LEDA Root Record as specified in Section **Error! Reference source not found.**.
- Zero or more LEDA EntityArtifacts, as specified in Section 3.3.

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

3.1 LEDA EntityArtifact Record

The LEDA EntityArtifact Record establishes a context and container for the collection of a number of Artifact Items associated with a given legal entity identified by the ELEI, as of a particular historical date, and potentially covering a particular interval of time.

Element Name	Type	Card	Description
ELEI	ELEI	1	The ELEI of the organization
PLEI	PLEI	0,1	"Previous Identifier" of the organization, generally in the form of IdentifierDomain.Identifier, e.g. "SCOB.94342"
EntityArtifactItems	EntityArtifactItems	1	Artifact Items collected in this EntityArtifact
Extension	Extension	0,1	An optional element for including data beyond the standard artifact Items in a LEDA data file. This may include data specific to an DSU, data specific to a publisher of LEDA data, and so on.
EntityArtifactCreationDate	DateTime	1	Date/time this LEDA record was initially created in the system
EntityArtifactUpdateDate	DateTime	1	Date/time that this historical LEDA record was most recently updated in the system.
EntityArtifactSource	EntityArtifactSourceType	1,n	A locator of the source of the data Artifact reference data in this EntityArtifact record (mechanism TBD, typically company yearbooks and exchange registrations)
EntityArtifactStatus	EntityArtifactStatusEnum	1	Status of the LEDA EntityArtifact.

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

Element Name	Type	Card	Description
EntityArtifactUpToDate	DateTime	0,1	Should it exist and be known, the historical date up until which time the LEDA data for this data Artifact is valid. (Not the same thing as when the data was changed in the system). If present, a subsequent historical record for this LEDA may exist with the revised information.
EntityArtifactAsOfDate	DateTime	0,1	The date that this version of the data is known to be valid. This date would typically be derived from either: 1. The publication date of the source reference, or 2. Information contained in the source reference
OriginalDCU	DCUID	1	The Identifier of the Data Collection Unit (DCU) that originally gathered or created the data in this EntityArtifact.
ResponsibleDSU	DSUID	1	The Identifier of the Data Submission Unit (DSU) that produced and manages this LEDA EntityArtifact.
DataMaturityStage	DataMaturityStageEnum	0,1	The current data maturity stage of this LEDA record, or omitted if the data maturity stage is not known.

3.1.1 EntityArtifactItems of the LEDA EntityArtifact

The EntityArtifactItems section of an LEDA EntityArtifact in a file conforming to this standard SHALL include one or more entity Artifact Items as specified below.

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

Element Name	Type	Card	Description
EntityArtifactItem	EntityArtifactItemType	1,n	One or more Artifact Items that are collected in this EntityArtifact

3.1.2 Extension Section of LEDA EntityArtifact

The Extension section of an LEDA record may be used to include additional data not defined in this standard. For example, a DSU may use Extension to publish additional Artifact Items it collects as part of EntityArtifact.

- TBD: include the details of how this works. Basically, the idea is to use an XSD schema wildcard with namespace ##other, permitting the inclusion of XML elements from other XML namespaces.

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

3.2 Data Types

This section specifies the data types referenced by the tables in Section **Error! Reference source not found.**, in alphabetical order.

3.2.1 DateTime Data Type

A value of type `DateTime` in a conforming to this standard SHALL be a point in time expressed as a string conforming to ISO 8601 having the following format:

- TBD: Provision for different calendars ? (i.e., Gregorian, Julian, etc.)

YYYY–*MM*–*DD*T*hh*:*mm*:*ss.sss*TZ

where the components of the above string are as follows:

- *YYYY* is the year
- *MM* is the month (01 = January, …, 12 = December)
- *DD* is the day of the month (01 = first day of the month)
- T is the single character 'T'
- *hh* is the hour (00 – 23)
- *mm* is the minute
- *ss.sss* is the second and milliseconds. From one to three digits may be used for milliseconds, or omitted entirely along with the decimal point.
- *TZ* is the time zone specifier, which can be either:
  - Z the single character 'Z', denoting Coordinated Universal Time (UTC); or
  - *+hh:mm* denoting a positive offset from UTC; or
  - *-hh:mm* denoting a negative offset from UTC

In the XML representation specified in Section **Error! Reference source not found.**, the XSD type `xs:dateTime` is used; however, whereas `xs:dateTime` permits the time zone specifier to be omitted, `DateTime` values in files conforming to this standard SHALL always include a time zone specifier.

Explanation (non-normative): milliseconds are hardly necessary for LEDA reference data, and likewise it might seem simpler to allow only "Z" as a time zone specifier; however, XML processing tools support the full syntax given above and it is not always possible to restrict such tools to avoid milliseconds or force the use of "Z" as the time zone specifier. The restriction that

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

the time zone specifier must be present is equivalent to using XSD type `xs:timestamp`*; however this was introduced in XSD 1.1 and not supported by the majority of XML processing tools which still only implement XSD 1.0.*

3.2.2 DCUID Data Type

A value of type `DCUID` in a file conforming to this standard SHALL be a [TBD]-character Data Collection Unit Identifier conforming to [TBD].

3.2.3 DSUID Data Type

A value of type `DSUID` in a file conforming to this standard SHALL be a [TBD]-character Data Submission Unit Identifier conforming to [TBD].

3.2.4 EntityArtifactCategory Data Type

A value of type `EntityArtifactCategory` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`EntityArtifactCategory`	`EntityArtifactCategoryTypeEnum`	0,1	Indicates the general category of the type of data Artifact identified by this LEDA EntityArtifact

3.2.5 EntityArtifactItem Data Type

An `EntityArtifactItem` in an LEDA EntityArtifact in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
`ArtifactItemConceptName`	`ItemName`	1,1	The Concept Name of the Artifact Item.
`ArtifactItemVariableName`	`ItemName`	1,1	The Variable Name of the Artifact Item.
`OtherArtifactItemNames`	`String`	0..n	An optional list of other `Name` instances for the EntityArtifactItem, whether in different languages in the jurisdiction of the legal entity, or in different languages used in foreign jurisdictions.

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

Element Name	Type	Card	Description
EntityArtifactItemCategory	EntityArtifactItemCategoryType	0,1	Indicates the domain category of the type of this EntityArtifactItem
EntityArtifactItemAsOfDate	Date	0,1	AsOf date of the EntityArtifactItem fact (assuming it is date-dependent)
EntityArtifactItemValue	String	0,1	Artifact Item value
EntityArtifactItemType	ElementDataType	0,1	Artifact Item type
EntityArtifactItemStatus	EntityArtifactItemStatusEnum	1	The status of the EntityArtifactItem. This is not to be confused with the status of the EntityArtifact; see EntityArtifactStatus.

3.2.6 EntityArtifactItemCategory Data Type

An EntityArtifactItem in an LEDA EntityArtifact in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
ArtifactItemDomainName	Name	1,1	The Domain Name of the Data Concept.
ArtifactItemDefinition	URL	1,1	Link to semantic web definition of this element

3.2.7 EntityArtifactItemStatus Data Type

An EntityArtifactItemValue in an LEDA EntityArtifact in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
ArtifactItemDataStatus	ElementDataStatus	1,1	Status of the ArtifactItemValue (TBD)

3.2.8 EntityArtifactItemType Data Type

An EntityArtifactItemValue in an LEDA EntityArtifact in a file conforming to this standard SHALL include data elements as specified below.

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

Element Name	Type	Card	Description
ArtifactItemDataType	ElementDataType	1,1	DataType of the ArtifactItemValue
ArtifactItemDefinition	URL	1,1	Link to semantic web definition of this element

3.2.9 ItemName Data Type

A `ItemName` is a string expressed in a natural language, including a code indicating which natural language is used.

A value of type `ItemName` in a file conforming to this standard SHALL include data elements as specified below.

Element Name	Type	Card	Description
lang	LanguageCode	0,1	The language of `name`
NameDomain	String	1,1	The Domain Name of the `ItemName`
Name	String	1	The name itself.

3.2.10 LanguageCode Data Type

A value of type `LanguageCode` in a file conforming to this standard SHALL be a 2-character language code conforming to [ISO639-1]. Note that ISO 639-1 language codes are all lowercase.

- TBD: Alternatively, we could use IETF language tags (RFC 4646) instead. IETF language tags can distinguish between variations of the same language in different countries. For example, the IETF language tags `fr-CA` and `fr-FR` denote Canadian French and French as spoken in France, respectively; in ISO 639-1 these would both be simply `fr`.

3.2.11 Name Data Type

A `Name` is a string expressed in a natural language, including a code indicating which natural language is used.

A value of type `Name` in a file conforming to this standard SHALL include data elements as specified below.

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

Element Name	Type	Card	Description
lang	LanguageCode	0,1	The language of name
Name	String	1	The name itself.

3.2.12 OtherEntityArtifactItemName Data Type

A value of type OtherEntityArtifactItemName in a file conforming to this standard SHALL include data elements as specified below. Each Name element includes an optional language code, permitting OtherEntityArtifactItemName to be repeated as many times as necessary to express the same name type in multiple languages. When type is PREFERRED_ROMANIZED_LEGAL or AUTO_ROMANIZED_LEGAL, the language code specifies the language of the name prior to Romanization.

Element Name	Type	Card	Description
type	EntityArtifactItemNameTypeEnum	1	The type of name represented by this OtherEntityArtifactItemName instance. The EntityArtifactItemNameType observes language, since 'Name' type has a language attribute.
Name	Name	1	The name. If type is PREFERRED_ROMANIZED_LEGAL or AUTO_ROMANIZED_LEGAL, then this value SHALL only include characters from the character set specified in Section 7.

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

3.2.13 EntityArtifactSource Data Type

EntityArtifactSourceCitation	EntityArtifactSourceCitationType	0,1	A locator of the source of the EntityArtifactItemreference data in this EntityArtifact record (mechanism TBD, typically company yearbooks or exchange EntityArtifacts)
EntityArtifactSourceLocatorKey	EntityArtifactSourceLocatorKeyType	0,1	A locator of the source of the EntityArtifactItemreference data in this EntityArtifact record (mechanism TBD, typically company yearbooks)
EntityArtifactSourceLocatorAnchor	EntityArtifactSourceLocatorAnchorType	0,1	Additional information to allow pinpointing where in the EntityArtifactSource of the EntityArtifactItemdata can be found
EntityArtifactSourceDCU	DCUID	0,1	The identifier of the Data Collection Unit that sourced the EntityArtifact data
EntityArtifactSourceArtifactID	String	0,1	An identifier of the EntityArtifactItemthat may have been locally used or assigned by the DCU at collection time

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

3.2.14 EntityArtifactSourceCitation Data Type

An associative locator (link or key) that resolves to a DDI metadata block regarding the source of the EntityArtifactItemreference data in this EntityArtifact record (mechanism TBD, typically company yearbooks -)

DDI 3.2 uses the `CitationType` Element for referencing publications (e.g. yearbooks). It contains 11 elements which are described in detail in the *DDI Lifecycle XML Schema*. The table below lists the suggested cardinality for each element and some notes on special uses. The name of each element is a link to its description in the DDI XML Schema.

Element Name	Type	Card	Description
Title	DDI Title	1,1	
SubTitle	DDI SubTitle	0,1	
AlternateTitle	DDI AlternateTitle	0,n	
Creator	DDI Creator	0,n	
Publisher	DDI Publisher	0,1	
Contributor	DDI Contributor	0,n	
PublicationDate	DDI PublicationDate	0,1	For monographs: SimpleDate For serials (e.g. yearbooks): StartDate and EndDate (the date/year when the first and the last volume of a serial were published – the specific volume and page numbers of the year from which information about the EntityArtifactItemwas taken can then be recorded in the EntityArtifactSourceLocatorAnchor element)
Language	DDI Language	0,n	
InternationalIdentifier	DDI InternationalIdentifier	0,n	
Copyright	DDI Copyright	0,1	
dc:isPartof	DDI dc:isPartof	0,1	Use for instance to record the title of the newspaper in which a stock exchange price list is published

The Citation as we would want to use it in the EntityArtifactReference Data would be considered as a DataSource in DDI 3.2 (Origin element).

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

3.2.15 EntityArtifactSourceLocatorAnchorType Data Type

The `EntityArtifactSourceLocatorAnchorType` provides additional information to allow pinpointing where in the EntityArtifactSource referenced by the `EntityArtifactSourceLocatorKey` can the EntityArtifactdata be found

3.2.16 EntityArtifactSourceLocatorKeyType Data Type

An associative locator (link or key) that resolves to a DDI metadata block regarding the source of the EntityArtifact reference data in this EntityArtifact record (mechanism TBD, typically company yearbooks -)

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

3.3 Enumerated Code Lists

This section specifies the enumerated code list data types (all having the suffix `Enum`) referenced by the tables in Sections **Error! Reference source not found.** and 3.1, in alphabetical order.

3.3.1 EntityArtifactCategoryTypeEnum

The `EntityArtifactCategoryTypeEnum` value in an LEDA record specifies the general classification category of the EntityArtifact.

- TBD: the table below is just an example with some possible categories, and is incomplete, and needs to be elaborated.

A value of type `EntityArtifactCategoryTypeEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
	To Be Determined

3.3.2 DataMaturityStageEnum Code List

A value of type `DataMaturityStageEnum` in a file conforming to this standard SHALL be one of the code strings specified in the following table:

Code	Definition
`UNVERIFIED`	(e.g., "Raw") Data has been acquired without undergoing any EURHISFIRM data quality measures
`VERIFIED`	(e.g., "Collected") Source-level metadata attribution and data collection measures have been performed. This maturity stage is what the output of a Data Collection Unit (DCU) would achieve.
`COMPLIANT`	(e.g., "Harmonized") Source-level data elements have been harmonized to EURHISFIRM field-level Common Data Model conventions. In DDI terms: source variables have been mapped to conceptual (i.e., semantic) standards. This maturity stage is typically associated with the output of a Data Submission Unit (DSU)

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

Code	Definition
CONSOLIDATED	(e.g., "Reconciled")Data previously collected and harmonized is compared with other EURHISFIRM data (whether previously published or submitted by other DSUs) in order to identify multiple references to the same unique object (legal entity, issued security, etc) and resolve duplicate references (identifiers) to the same object.
CONSOLIDATED	(e.g., "Promoted"). After reconciliation (deduplication / unique identification), data is promoted to the maturity stage of published Common Data Model EURHISFIRM data
EU REVISED	After being promoted, the data was revised or edited

.

4 Change Management

- TBD. Use the <vnext> strategy for forward/backward compatibility, and/or include a version number in the XML. Needs to be explained

5 Examples (non-normative)

- TBD.

6 References

[ISO646] ISO, "Information technology -- ISO 7-bit coded character set for information interchange," ISO/IEC 646:1991.

[ISO3166-1] ISO, "Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes," ISO 3166-1:2013.

[ISO3166-2] ISO, "Codes for the representation of names of countries and their subdivisions – Part 2: Country subdivision code," ISO 3166-2:2013.

[ISO8601] ISO, "Data elements and interchange formats – Information interchange – Representation of dates and times," ISO 8601:2004.

[ISO10646] ISO, "Information technology -- Universal Coded Character Set (UCS)," ISO 10646:2012.

[ISO17422] ISO, "Financial Services - Legal Entity Identifier (LEI)," ISO/DIS 17442:2012.

[ISODir2] ISO, "Rules for the structure and drafting of International Standards (ISO/IEC Directives, Part 2, 2001, 4th edition)," July 2002.

[XSD1] H. Thompson, D. Beech, M. Maloney, N. Mendelsohn, "XML Schema Part 1: Structures," W3C Recommendation, May 2001, http://www.w3.org/TR/xmlschema-1/.

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

[XSD2] P. Biron, A. Malhotra, "XML Schema Part 2: Datatypes," W3C Recommendation, May 2001, http://www.w3.org/TR/xmlschema-2/.

7 Appendix: Character Codes Allowed in Romanized Names

When a `Name` instance is of type `PREFERRED_ROMANIZED` or `AUTO_ROMANIZED`, the value of the `name` field SHALL consist only of non-control characters drawn from the "invariant subset" of ISO 646. These characters are enumerated below. The "Hex Value" column indicates the code point value (expressed in hexadecimal) for each character in both ISO 646 and ISO 10646.

Graphic Symbol	Name	Hex Value	Graphic Symbol	Name	Hex Value
!	Exclamation Mark	21	M	Capital Letter M	4D
"	Quotation Mark	22	N	Capital Letter N	4E
%	Percent Sign	25	O	Capital Letter O	4F
&	Ampersand	26	P	Capital Letter P	50
'	Apostrophe	27	Q	Capital Letter Q	51
(	Left Parenthesis	28	R	Capital Letter R	52
)	Right Parenthesis	29	S	Capital Letter S	53
*	Asterisk	2A	T	Capital Letter T	54
+	Plus sign	2B	U	Capital Letter U	55
,	Comma	2C	V	Capital Letter V	56
-	Hyphen/ Minus	2D	W	Capital Letter W	57
.	Full Stop	2E	X	Capital Letter X	58
/	Solidus	2F	Y	Capital Letter Y	59
0	Digit Zero	30	Z	Capital Letter Z	5A
1	Digit One	31	_	Low Line	5F
2	Digit Two	32	a	Small Letter a	61
3	Digit Three	33	b	Small Letter b	62
4	Digit Four	34	c	Small Letter c	63
5	Digit Five	35	d	Small Letter d	64
6	Digit Six	36	e	Small Letter e	65
7	Digit Seven	37	f	Small Letter f	66
8	Digit Eight	38	g	Small Letter g	67

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

Graphic Symbol	Name	Hex Value	Graphic Symbol	Name	Hex Value
9	Digit Nine	39	h	Small Letter h	68
:	Colon	3A	i	Small Letter i	69
;	Semicolon	3B	j	Small Letter j	6A
<	Less-than Sign	3C	k	Small Letter k	6B
=	Equals Sign	3D	l	Small Letter l	6C
>	Greater-than Sign	3E	m	Small Letter m	6D
?	Question Mark	3F	n	Small Letter n	6E
A	Capital Letter A	41	o	Small Letter o	6F
B	Capital Letter B	42	p	Small Letter p	70
C	Capital Letter C	43	q	Small Letter q	71
D	Capital Letter D	44	r	Small Letter r	72
E	Capital Letter E	45	s	Small Letter s	73
F	Capital Letter F	46	t	Small Letter t	74
G	Capital Letter G	47	u	Small Letter u	75
H	Capital Letter H	48	v	Small Letter v	76
I	Capital Letter I	49	w	Small Letter w	77
J	Capital Letter J	4A	x	Small Letter x	78
K	Capital Letter K	4B	y	Small Letter y	79
L	Capital Letter L	4C	z	Small Letter z	7A
	Space	20			

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