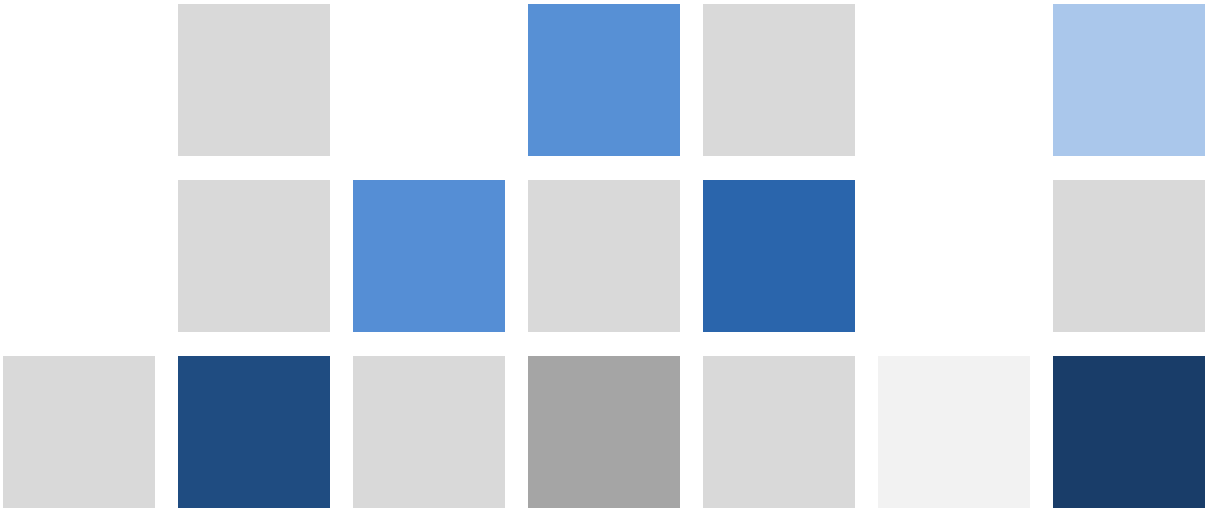
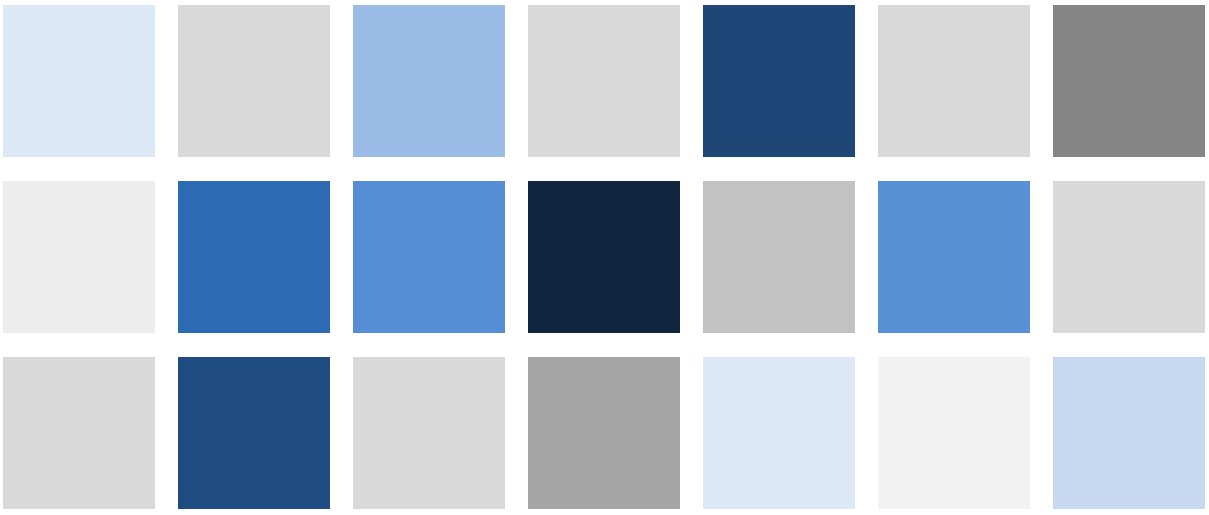


Long-term data for Europe

EURHISFIRM

D4.5: Report on EURHISFIRM documentation standard



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

AUTHOR:

Johan POUKENS (University of Antwerp)*

APPROVED IN 2019 BY:

Jan ANNAERT (University of Antwerp)

Wolfgang KÖNIG (Goethe University)

Angelo RIVA (Paris School of Economics)

* The author would like to thank Oliver Watteler (GESIS) and Coen Fierst van Wijnandsbergen (Rotterdam School of Management, Erasmus University) for their valuable comments to earlier drafts and Lana Yoo (Paris School of Economics) for revising the language of this document.

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 purpose of Task 4.5 was selecting data documentation or metadata standards for EURHISFIRM. Metadata is defined by the International Organisation for Standardisation (ISO) as "data that defines and describes other data" or, simply, “data about data”. A metadata standard, then, is "an agreed list of common metadata items and the standardisation of terminology and definitions for these items" (Organisation for Economic Co-operation and Development, 2007, pp. 73, 76). This report covers only the EURHISFIRM standards for the documentation, i.e. description of the provenance, characteristics, structure, and contents of datasets and printed sources. Other Working Packages may propose additional standards for other purposes. In addition to metadata standards, we also propose software applications for documenting datasets and printed sources according to the proposed metadata standards.

In Report D4.1, we already reviewed several generic metadata standards and metadata standards for the social sciences, as well as software tools for producing, editing, storing and retrieving metadata. The purpose of Task 4.1 was selecting a metadata standard and an appropriate software tool for the selected standard. The selected metadata standard (Data Documentation Initiative Lifecycle) and software (Colectica Designer) were then used in Task 4.4 (Data and Sources Documentation Production and Quality Assessment) for producing the homogenous data documentation of the printed sources and datasets identified in Task 4.2 (Data and Sources Inventory). The present report builds on the findings of these previous tasks, but the focus is shifted towards the appropriate data documentation standards for future use in the EURHISFIRM federated research infrastructure itself (as opposed to their application in the design phase).

The structure of this report is as follows: First, the Data Documentation Initiative standards are briefly re-introduced.[1] Then, different functions of metadata in EURHISFIRM and the appropriate standards for these tasks will be discussed. Finally, two possible software applications which can (partially) perform these functions in the appropriate metadata standard are introduced.

2 Data Documentation Initiative overview

The Data Documentation Initiative (DDI) is a set of standards for documenting the microdata produced by surveys and other observational methods in the social, behavioural and economic sciences. It was developed by the Inter University Consortium for Political and Social Research (ICPSR) and is currently maintained by the DDI Alliance.[2] The DDI family currently contains two specifications, DDI Codebook (currently version 2.5) and DDI Lifecycle (currently version 3.2), as well as controlled vocabularies for use in various metadata elements.

- **DDI Codebook** (further referred to as DDI 2.5) can be used to document a single data collection. The current version contains 351 elements which are organised into five sections (Figure 1). The study, data files and variable descriptions respectively document a dataset, the files in a dataset

[1] A more complete introduction to DDI, as well as to other metadata standards mentioned further in the this report can be found in EURHISFIRM D4.1.
[2] https://www.ddialliance.org

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

and the variables in a data file. These are preceded by a document description which contains metadata about data documentation and followed by the other materials section which references books, articles or other works containing information related to the dataset.

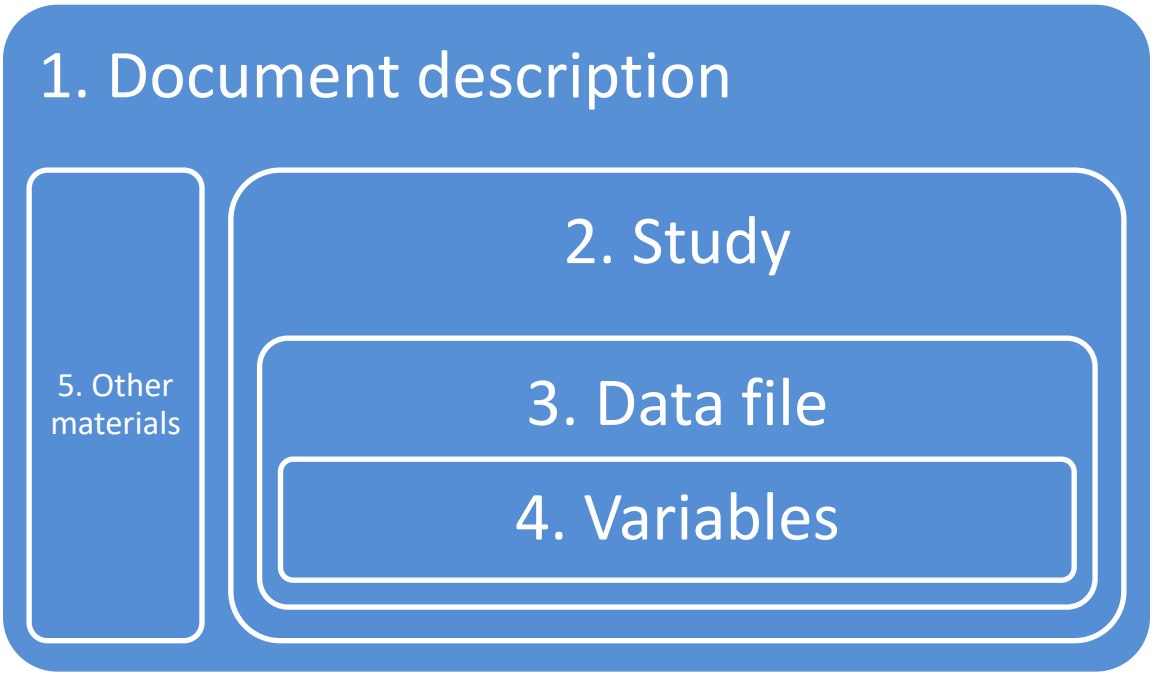


Figure 1: Sections in a DDI Codebook data description

- **DDI-Lifecycle** (further referred to as DDI 3.2) supports the entire research data lifecycle (from planning to archiving). The current version of DDI 3.2 contains 1,154 elements which are arranged according to modules. Modules roughly relate to stages in the research data lifecycle (e.g. data collection or archiving) and publishing packages (e.g. studies or series). The studyUnit publishing package describes a single study and most closely corresponds to DDI 2.5.

DDI was chosen in D4.1 as the metadata standard for the Data and Sources Documentation (D4.4) over other standards such as Dublin Core, DataCite and da|ra. DDI was not only selected for its own merits, but also because of its broad acceptance in the social sciences community.

- The high number of elements in DDI is illustrative for the level of detail with which it is possible to document datasets down to the lowest level of the individual variables in data files. In D4.1, we preferred DDI 3.2 over DDI 2.5 because it was, amongst other things, better suitable for the documentation of printed serial sources and the harmonization of variables. Only DDI 3.2 has metadata elements for the documentation of groups and conceptual variables.
- DDI is the standard of choice for the Consortium of European Social Sciences Data Archives (CESSDA). Since the EURHISFIRM standard of choice must meet CESSDA's requirements, DDI is also the natural choice for EURHISFIRM. Furthermore, DDI is very much alive. The DDI Alliance, for instance, provides regular updates of DDI and CESSDA (as well as other users of DDI) are continuously developing applications that support the use of DDI.

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

DDI is reconfirmed here as the documentation standard of choice for the EURHISFIRM research infrastructure, albeit with some differences because of the different functions of data documentation in the post-design phase. These functions will be discussed in the next section.

3 Functions of metadata in EURHISFIRM

Metadata are paramount for meeting the FAIR Data Principles (FAIR being an acronym of Findability, Accessibility, Interoperability and Reusability). The FAIR Data Principles were formulated to support the reuse of research data from a data user's (i.e. researcher's) standpoint (Wilkinson et al., 2016). Complete and standardised metadata helps researchers to discover and understand data collected by others independently, but can also serve data producers like EURHISFIRM in the process of adding new data to existing data.

3.1 Data upload

Long-term data on companies will be added to the federated EURHISFIRM infrastructure in at least two ways. Firstly, high-quality scans of historical printed sources can be transformed into structured datasets by the software which is developed by Working Package 7. Secondly, datasets created by researchers and teams who have manually or otherwise collected data will be matched and integrated or connected to existing data by the processes and procedures which are developed by Working Package 6. Both approaches need metadata for the correct identification and interpretation of resources that are being contributed by researchers.

- The first approach (data extraction from digitised images of printed serial sources) principally needs simple descriptive metadata about the source (for instance, title, creator, publisher, publication date, temporal and spatial coverage and an abstract and keywords describing the contents). If scans are also provided, additional metadata about the data files need to be provided (for instance, file names and file formats).
- The second approach (adding data from existing datasets) requires not only the aforementioned metadata about the dataset and the data files in general, but also about the data elements or variables contained in the data files (for instance, variable names, labels and data types, e.g. numeric, text, date or codes). Moreover, the documentation of the dataset also needs to provide a description of the collection and the sources of the data. These details on the provenance of the data are necessary for assessing the quality of the data.

3.2 Data harmonisation and alignment

We can expect that most of the datasets that will be contributed by researchers will be designed with specific, narrow aims and are not standardised or interoperable. After upload, new data needs to be harmonised before it can be added to the pool of existing data. This means that variables from different datasets which describe a particular characteristic of a company, security or person in a similar way need to be matched to a common denominator. Common denominators can be data elements from the EURHISFIRM Common Data Model or may originate from external vocabulary sets such as the Financial Industry Business Ontology (or FIBO, see Figure 2 for an example). The metadata standards need to accommodate for storing these links between variables.

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

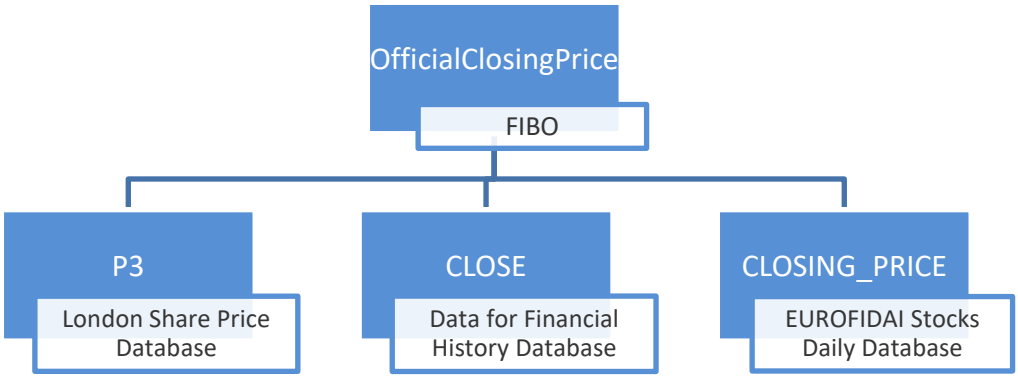


Figure 2: Harmonisation of variables from different datasets

Furthermore, a new resource can be an iteration of an existing resource. A newly uploaded scan of a yearbook, for instance, could be a missing volume from a series that is already in the infrastructure. Or, in case of more recent data, an external data provider might annually contribute data about the previous year. Alternatively, multiple sources or datasets can also form a single time series together (Figure 3). The metadata standards also need to accommodate for storing links between serial sources or datasets.

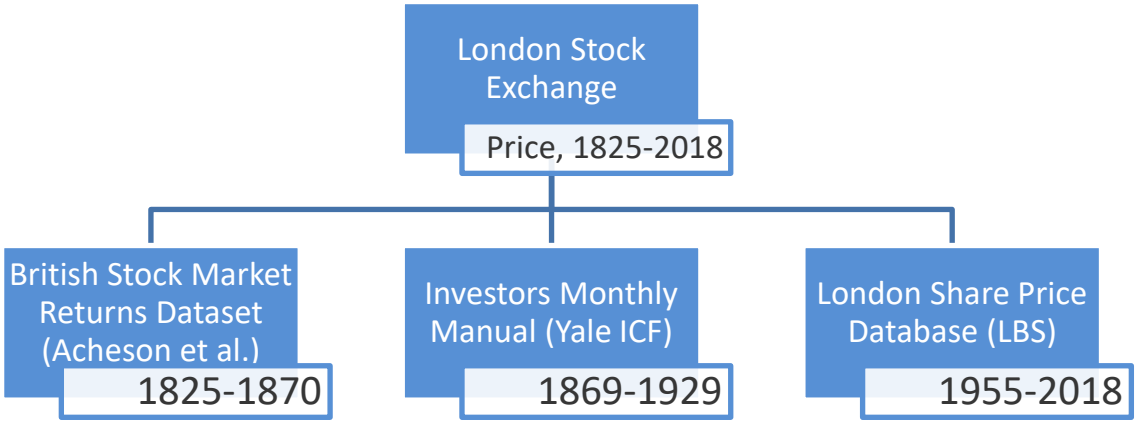


Figure 3: Alignment of multiple datasets into time series

In the long-run, it should become an aim of EURHISFIRM to empower the research community to contribute complete and well-structured data by providing guidelines, templates and other tools for the

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

preparation of various types of data by primary researchers. This should result in more common characteristics shared among datasets contributed by different researchers and will greatly improve the harmonisation and alignment process.

3.3 Discoverability

Data published on the EURHISFIRM platform needs to be findable for researchers. Findability or discoverability of course implies the availability of descriptive metadata to users. Subjects and keywords are very important as points of entry in this respect. The discoverability of data is enhanced if subjects and keywords are derived from domain-relevant subject classifications and thesauri (i.e. controlled vocabularies) such as the *Standard-Thesaurus Wirtschaft* (Thesaurus of Economics, Leibniz Information Centre for Economics). The use of controlled vocabularies such as the Thesaurus of Geographic Names (Getty Research Institute) is also advisable for descriptions of the geographical coverage of sources and datasets. Controlled vocabularies furthermore support multilingualism, an important feature of a multi-national research infrastructure such as EURHISFIRM.

Data should also be citable. The metadata should include a citation to ensure that researchers who use the data properly attribute the source of the data and a persistent identifier. The presence of persistent identifiers is of critical importance to the FAIR Data Principles because it associates the metadata with the dataset they describe.[3]

4 EURHISFIRM metadata standards

This section will put forward the metadata standards from the DDI family which can support the various functions of metadata in the federated EURHISFIRM infrastructure. We explicitly use the plural because, at present, there is no one standard that serves all functions equally well.

4.1 DDI 2.5

We propose DDI 2.5 as the standard for documenting sources and datasets at the Study level during upload.[4] According to the current design of the EURHISFIRM infrastructure, upload of sources and datasets is performed at the local level - that is, by individual researchers or teams who have collected images or data but are not necessarily part of EURHISFIRM. Conceptually, it would be better to apply a metadata standard before upload. As researchers, however, we know all too well that data are collected first and foremost for private use. This precludes the necessity of extensively documenting datasets. Although funders of research increasingly emphasise the importance of data management and demand that research data is archived for future re-examination and reuse, the fairly recent articulation of the FAIR Data Principles shows that they have not yet accomplished a situation in which researchers automatically produce data documentation that meets “domain-relevant community standards” (Wilkinson et al., 2016). Documentation of data during upload should therefore be kept as simple and easy as possible.[5] This

[3] https://www.go-fair.org/fair-principles/f3-metadata-clearly-explicitly-include-identifier-data-describe
[4] “Study” is the term used by DDI to designate data collected in a single research project.
[5] This approach, i.e. documentation of data during upload, is also favoured by GESIS’ data repository, *datorium* (Linne, 2013).

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

favours DDI 2.5 over the much more extensive DDI 3.2. Moreover, there are open source applications available for producing data documentation according to DDI 2.5, while tools for producing data documentation according to the DDI 3.2 standard are generally not openly available and more complicated to use.

The choice for DDI 2.5 as the standard for documenting sources and datasets during upload in no way constitutes a compromise on the necessary level of detail. DDI 2.5 accommodates all of the necessary metadata elements at the level of the dataset or source (study), the data files and the variables. Of course, DDI 2.5 lacks the reference structure and the metadata elements for describing common characteristics of datasets (group) and variables (represented and conceptual variables). During upload, however, this is not an issue because uploads will most likely be stand-alone datasets.

4.2 DDI 3.2

We propose DDI 3.2 as the standard for storing data documentation after data harmonisation. According to the current design of the EURHISFIRM infrastructure, harmonisation of data is performed at the national level, that is, by a service provider who is part of EURHISFIRM. This makes the complexity of DDI 3.2 and the lack of openly available tools for producing, editing, storing and retrieving DDI 3.2 metadata less of an issue.

DDI 3.2 encompasses all of the metadata elements of the DDI 2.5 specification. It also offers, amongst other things, the possibilities of grouping subsequent iterations of a printed source or dataset (or otherwise related sources and datasets) into series and of mapping variables from different datasets to common data elements. The latter requires some explanation.[6] A Data Element is the common unit for transferring information. In the ISO/IEC 11179 Data Element Classification Structure, a Data Element is composed of three parts (Figure 4): an Object Class, a Property and a Representation (see also Section 8). In a more traditional data-modelling terminology, the Object Class conforms to the Entity about which certain characteristics or Attributes are recorded. For the Object Class “bonds”, for instance, possible Properties are “interest rate” and “coupon date”. The Representation of the possible values that these characteristics of bonds can take are either a percentage (e.g. 3,5%) or a month and day (e.g. January 1) (ISO/IEC, 1999).

[6] From here onwards, specific terms used in standards such as Variable or Concept are capitalised.

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

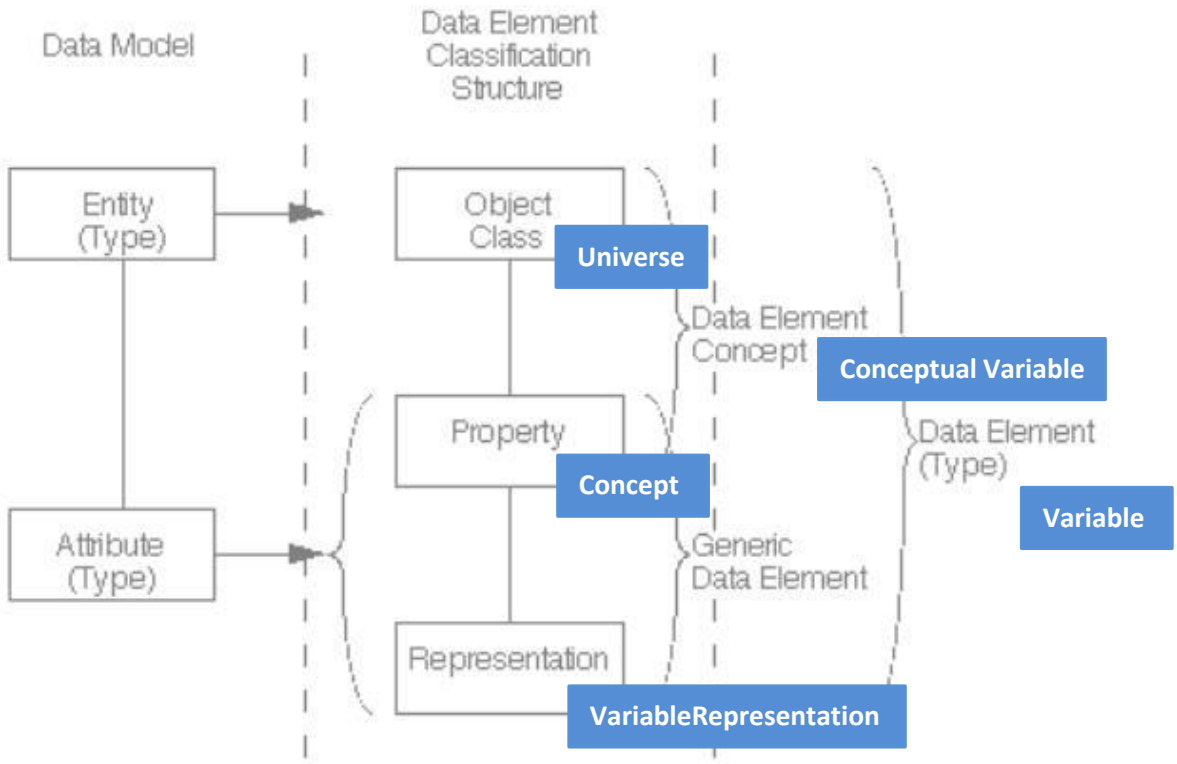


Figure 4: ISO/IEC 11179 Data Element Structure (ISO/IEC, 1999, p. 18).

The terms in blue textboxes represent the DDI equivalent of ISO/IEC 11179 terms. Definitions of ISO/EIC 11179 terms (in grey) are provided in Section 8 of this report.

In DDI terminology (represented by the blue labels in Figure 4), the Object Class is defined by the Universe; its Property is the Concept, and the Representation is the VariableRepresentation content used by the Variable that measures it. A Variable that references a Universe and Concept conforms to a Data Element. A Data Element Concept, in DDI 3.2 terminology, is a Conceptual Variable. The Conceptual Variable links a Universe and a Concept to create an abstract, reusable Data Element Concept without any particular Representation. A reusable expression of Representation can be recorded in a DDI 3.2 Represented Variable (DDI Alliance, 2014).[7]

The ISO/IEC 11179 Data Element Classification Structure is implemented in DDI 3.2 through the General Statistical Information Model (GSIM).[8]

Figure 5 shows how Variables, Represented Variables and Conceptual Variables are related to each other and to Concepts, Universes and Representations. The GSIM terminology differs slightly from the DDI terminology and is included in the blue labels (DDI Alliance, 2014; Nielsen & Dannevang, 2017).

[7] See also: https://www.ddialliance.org/standards/relationship-to-other-standards
[8] https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model

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

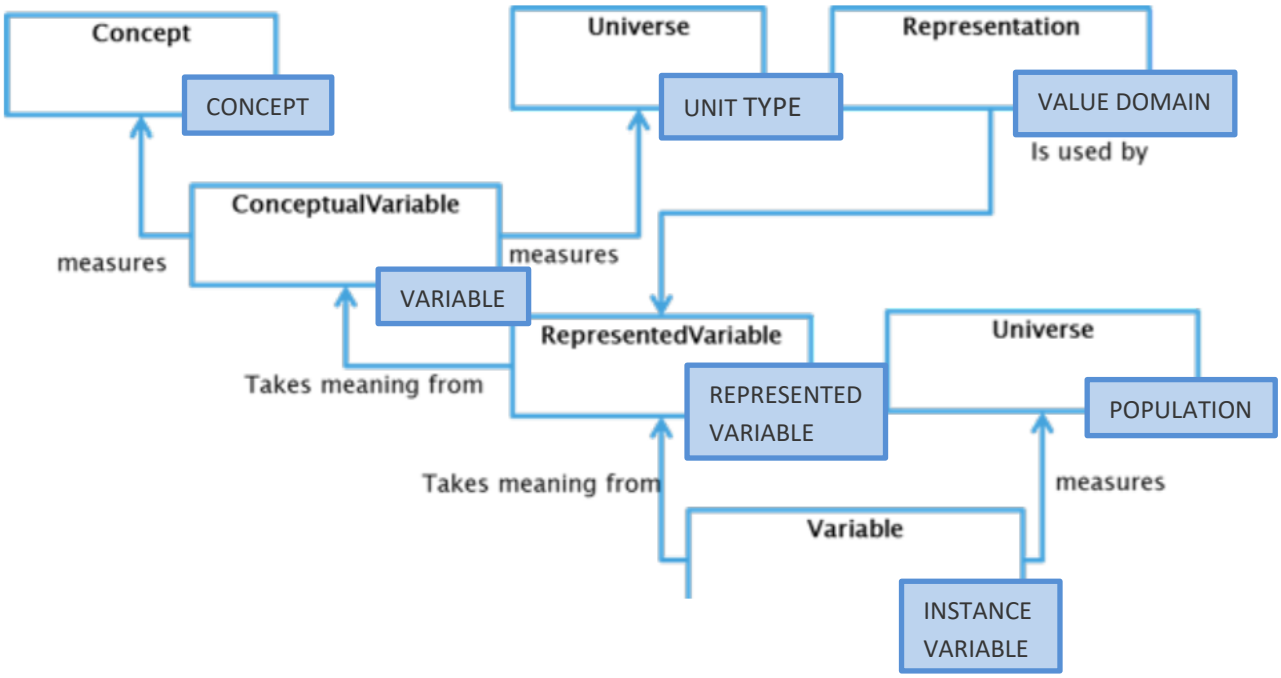


Figure 5: Mapping of DDI and GSIM Variables (Nielsen & Dannevang, 2017, p. 12)

Figure 6 illustrates how a Variable from the official price lists of the Brussels Stock Exchange - the last (or previous) price of a share - can be described with reusable DDI 3.2 metadata.

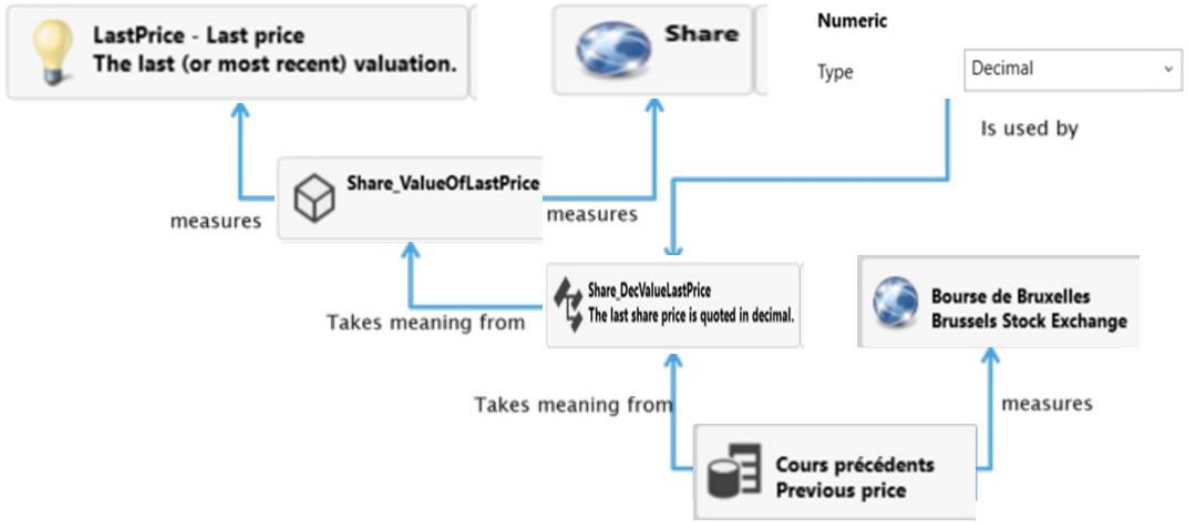


Figure 6: Example of harmonising Variables in DDI 3.2

Because DDI 3.2 is an implementation of the ISO/IEC 11179 Metadata Registry Standard (ISO/IEC, 2004), it is of special importance to EURHISFIRM. The EURHISFIRM Common Data Model is essentially a metadata

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

registry and DDI 3.2 presents an opportunity to implement it. It supports the mapping of data from diverse sources and databases to the Common Data Model, whilst retaining the lineage, provenance and characteristics of the original data sources. DDI 3.2 would also allow a mapping of historical corporate and financial data to present-day vocabularies such as the Financial Industry Business Ontology (FIBO).

FIBO is an RDF vocabulary or knowledge graph. In the RDF (Resource Description Framework) model, information is conveyed in the form of triples. A triple is composed of a Subject, Property (also called Predicate) and Object. The Subject is the resource (in RDF, any identifiable thing is a resource) which is being described, the Property is a characteristic of the resource and expresses the relationship between the Subject and the Object.[9] Take, for instance, the statement "*Société générale* ordinary shares have a nominal value of 1,000 Belgian francs". In this case, "*Societé générale* ordinary shares" is the Subject, "have a nominal value of" the Property and "1,000 Belgian francs" the Object. Resources with similar characteristics can be grouped into Classes (for instance, common shares).

An RDF ontology relates Classes and Properties into taxonomies. It contains definitions of Classes (the FIBO definition of common share is "a share that signifies a unit of ownership in a corporation and represents a claim on part of the corporation's assets and earnings") and Properties.[10] FIBO distinguishes Object Properties and Data Properties. The difference is in the Range, i.e. the possible values, that the Object can take. Data Properties can only have literal values (i.e. string, number or Boolean) while Object Properties can have a wider range of values (for instance, a monetary amount in the *Société générale* example, date or another Property). In the data model represented in Figure 4, Class and Property respectively corresponds to Entity and Attribute. In DDI 3.2, Class would map to Universe and Property to Concept and VariableRepresentation.

The ability of DDI 3.2 to leverage common, reusable metadata adds value for users too. For instance, Common metadata improves discoverability, as it refers a user from a specific variable in one dataset to other datasets with similar or related information (Nielsen & Dannevang, 2017).

4.3 Controlled vocabularies

In Task 4.4, two controlled vocabularies were used to document the topical and geographical coverage of printed sources and datasets. The Thesaurus of Geographical Names (Getty Research Institute) was used for geographical coverage. Subjects and keywords were taken from the *Standard-Thesaurus Wirtschaft* (Thesaurus for Economics, Leibniz Information Centre for Economics).

The aptness of **Thesaurus of Geographical Names** (TGN) for historical data has been discussed already in D4.1. In short, TGN offers a better coverage of former names of countries and regions than other geographic datasets such as GeoNames and GeoNames Search. In addition, TGN allows external contributions of missing places to their dataset.

The ***Standard-Thesaurus Wirtschaft*** (STW) is an open, bilingual (German-English) for economics and related disciplines. It covers macro economics and business economics. Related disciplines include,

[9] https://www.w3.org/TR/rdf-concepts
[10] https://spec.edmcouncil.org/fibo/ontology/SEC/Equities/EquityInstruments/CommonShare

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

amongst others, history and social sciences. It also includes classifications of industries and products which follow the standard classifications of the German Federal Statistics Office. Mappings to other general and subject-specific thesauri such as DBpedia (Wikipedia), the Integrated Authority File (German National Library) and the Journal of Economic Literature Classification Scheme (American Economic Association) are available.

Both thesauri were adequate for documenting the data and sources in Task 4.4. One of the minor shortcomings of STW was the lack of descriptors for specific stock exchanges (e.g. the London Stock Exchange). Since its editorial team is open to proposals from users, this should not be a major obstacle for future use in EURHISFIRM.

4.4 Persistent identifiers

The **Digital Object Identifier** (DOI) is an alphanumeric string assigned to uniquely identify an object. It is linked to a resource's metadata as well as to the URL of the website where the resource is accessible (Figure 7). The DOI is an ISO standard (ISO 26324) and the most commonly used persistent identifier by academic publishers. It is regarded as a good choice of persistent identifier for research data repositories (Duerr et al., 2011; Flathers, Kenyon, & Gessler, 2017).

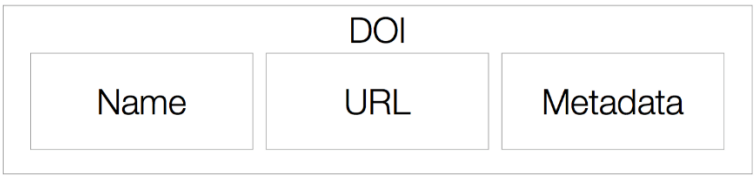


Figure 7: DOI structure (https://datacite.org/dois.html)

In order to be able to assign DOIs, EURHISFIRM would have to become a member of DataCite and also supply metadata to DataCite. DataCite is a global non-profit organisation that assigns DOIs for research data. Alternatively, EURHISFIRM could work with one of the current members of DataCite. EURHISFIRM member GESIS is also a member of DataCite and can assign DOIs. The metadata would have to be structured according to the DataCite Metadata Schema (DataCite Metadata Working Group, 2019). The DataCite Metadata Schema is described in more detail in D4.1. It is a very simple schema, however, and all mandatory elements can be mapped to DDI.

5 Software applications

DDI metadata are stored in complex XML documents. However, there are software applications available that aid researchers with no or little knowledge of XML and DDI in the documentation of datasets and sources according to DDI standards. This section puts forward two applications, Dataverse and Colectica, for respectively documenting datasets and sources at upload and harmonising and aligning data.

5.1 Dataverse

A dataverse is a virtual archive or collection of related datasets, files and metadata (e.g. datasets pertaining to the same project). In the Dataverse Project, Harvard University's Institute for Quantitative Social Science

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

(IQSS) has created an open-source web application to share, preserve, discover, cite and analyse research data (King, 2007). Institutions can install Dataverse on their own servers (the software can be downloaded from GitHub, a website which hosts software source code) or create their own cloud-based dataverse within the Harvard Dataverse Network.[11] The Dataverse software could serve different functions in the EURHISFIRM infrastructure. It registers DOIs for datasets, transforms tabular data to an archival format and offers features for visualizing data. Here, we focus mainly on the data documentation features during upload.

Researchers can upload their datasets and add (or edit) descriptive metadata through a web form. The editable descriptive metadata conforms to the StudyUnit metadata in DDI 2.5. There is currently no support for DDI 3.2. Dataverses in the Harvard Dataverse Network have 100 metadata elements. Metadata are customizable through the web interface to a certain extent (e.g. select optional and mandatory metadata elements, hide elements and create templates). Much more customization (e.g. edit and add controlled vocabularies) is possible in Dataverse installations. The Dataverse software can also extract Variable metadata from tabular data files. During upload, data is extracted from the user-uploaded files and archived in an application-neutral format (plain text, tab delimited). The Variable metadata is also extracted and stored in a separate, relational database. Metadata can be exported as DDI 2.5 XML files. The tabular data ingest supports various formats (SPSS, STATA, R, Excel and CSV), but some formats have limitations. STATA currently is the best supported format for tabular data ingest.[12] Another limitation of Dataverse is that Variable metadata cannot be added or edited manually. It is not possible, for instance, to add labels to Variables originating from CSV files or to change the Representation.

One of Dataverse's strengths is its active community of users. There are currently 46 institutions or consortia of institutions that have their own Dataverse installation. In addition, over 3,300 dataverses have been created on the Harvard Dataverse platform. Two institutions from the EURHISFIRM consortium (the Royal Dutch Academy of Science's Internationaal Instituut voor Sociale Geschiedenis and Universidad Carlos III de Marid) also use Dataverse. The number of dataverses has been growing continuously. Hence, researchers will become more familiar with the practice of documenting and archiving datasets and with the Dataverse software. Moreover, the user-community also continuously develops the Dataverse software. Very recently, Data Curation Tool has been released as an external tool for editing Variable metadata in a dataverse.[13] Also, in CESSDA's ongoing DataverseEU project, a hosted dataverse-service for National Service Providers with support for DDI Lifecycle will be developed (Tykhonov, 2018). One of the participants of this project is the Data Archiving and Networked Services (DANS). DANS is CESSDA's Dutch National Service Provider and affiliated to EURHISFIRM partner Royal Dutch Academy of Science (KNAW). DANS also hosts DataverseNL which is used by several Dutch universities.

5.2 Colectica

Algenta Technologies, a developer of software for research, offers functionalities for creating, editing, storing and sharing metadata through its Colectica Platform.[14] Colectica Designer is the editing client at

[11] https://dataverse.org
[12] http://guides.dataverse.org/en/latest/user/tabulardataingest/index.html
[13] https://github.com/scholarsportal/Dataverse-Data-Curation-Tool
[14] https://www.colectica.com/software

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

the heart of the platform. Data documentation created in Designer can be published to a server with Colectica Repository and indexed for search via a website with Colectica Portal. Colectica supports all versions of DDI (including DDI 2.5 and DDI 3.2). It could be used in EURHISFIRM to transform the DDI 2.5 metadata from Dataverse to DDI 3.2 and edit or add Variable metadata. Principally, it could be used for aligning and harmonizing data by adding Group or Series and Conceptual Variable references. For instance, Colectica is used by Statistics Denmark (the Danish national statistical agency) for implementing their common metadata definitions (i.e Conceptual Variables).

6 Conclusion

The EURHISFIRM will rely heavily on data documentation or metadata. To meet FAIR Data Principles, metadata will have to conform to standards. Different standards for the social sciences were reviewed in D4.1 for the purpose of producing the data documentation during the design phase. The purpose of this report was choosing metadata standards for the post-implementation phase of EURHISFIRM. Because one standard cannot serve all functions, several specifications, coming mainly from the Data Documentation Initiative (DDI) family of standards, were proposed. For documenting stand-alone data during upload, the relative simplicity and availability of open-source editors are the main advantages of DDI Codebook (i.e. DDI 2.5). DDI Lifecycle (i.e. DDI 3.2), however, is better suited for the alignment and harmonisation of data by EURHISFIRM's National Focus Points. The use of controlled vocabularies and persistent identifiers further contributes to the discoverability of data. From our own experience during Task 4.4 and the literature, the Thesaurus of Geographic Names and the STW Thesaurus for Economics and the Digital Object Identifier come forward as the standards of choice in this respect. We also proposed applications for managing metadata during ingest and standardisation. Currently, Dataverse and Colectica offer the best perspectives in this respect.

The standards and applications proposed in this report should not be taken as definitive and irrefutable choices for EURHISFIRM, however. The selection and implementation of standards through specific software applications should be the object of regular review. Future choices regarding the services EURHISFIRM intends to provide, as well as developments of both standards and applications, may warrant the choice of alternative or additional standards and applications. The fact that the DDI standards are XML based is a great advantage in this respect. The use of mark-up language facilitates the transfer of information in case this would be necessary. The work of other Working Packages may also result in new decisions regarding standards. Working Package 5 in particular will provide metadata documentation for Common Data Model data elements that not only aligns with references and integrates DDI elements, but it will also extend DDI with metadata documentation for additional elements not included in DDI. Requirements for applications, on the other hand, will be designed by Working Package 9.

7 References

DataCite Metadata Working Group. (2019). *DataCite Metadata Schema documentation for the publication and citation of research data* (Version 4.2). Retrieved from https://doi.org/10.5438/bmjt-bx77

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

DDI Alliance. (2014). *Data Documentation Initiative (DDI) Technical specification. Part I: Technical documentation (Version 3.2)*. Retrieved from http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/HighLevelDocumentation/DDI_Part_I_TechnicalDocument.pdf

Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., … Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: An assessment and recommendations. *Earth Science Informatics*, *4*(3), 139. https://doi.org/10.1007/s12145-011-0083-6

Flathers, E., Kenyon, J., & Gessler, P. E. (2017). A service-based framework for the OAIS model for earth science data management. *Earth Science Informatics*, *10*(3), 383–393. https://doi.org/10.1007/s12145-017-0297-3

ISO/IEC. (1999). *Information technology - Specification and standardization of data elements - Part 1: Framework for the specification and standardization of data elements* (First version). Retrieved from https://www.oasis-open.org/committees/download.php/6233/c002349_ISO_IEC_11179-1_1999%28E%29.pdf

ISO/IEC. (2004). *Information technology. Metadata registries (MDR). Part 1: Framework* (Second edition). Genève: ISO/IEC.

King, G. (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research*, *36*(2), 173–199. https://doi.org/10.1177/0049124107306660

Linne, M. (2013). Sustainable Data Preservation using datorium – facilitating a scholarly Ideal of Data Sharing in the Social Sciences. In J. Borbinha, M. Nelson, & S. Knight (Eds.), *Proceedings of the 10th International Conference on Preservation of Digital Objects* (pp. 150–155). Lisbon: Biblioteca Nacional de Portugal.

Nielsen, M. G., & Dannevang, F. (2017). Towards Common Metadata Using GSIM and DDI 3.2. *IASSIST Quarterly*, *40*(2), 6. https://doi.org/10.29173/iq782

Organisation for Economic Co-operation and Development. (2007). *Data and metadata reporting and presentation handbook*. Paris: Organisation for Economic Co-operation and Development.

Tykhonov, V. (2018). *CESSDA DataverseEU Project*. Presented at the EDDI18 – 10th Annual European DDI User Conference (Berlin, 4-5 December 2018). Retrieved from http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi18/paper/view/387

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

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

8 Appendix: Definitions of ISO/IEC 11179 terms used in Figure 4

Term	Definition
Attribute	A characteristic of an **Object** or **Entity**
Concept	A unit of thought constituted through abstraction on the basis of characteristics common to a set of **Objects**. [ISO 1087]
Data	A **Representation** of facts, **Concepts**, or instructions in a formalized manner, suitable for communication, interpretation, or processing by humans or by automatic means. [ISO 2382-4]
Data Element	A unit of **Data** for which the definition, identification, **Representation**, and permissible values are specified by means of a set of **Attributes**.
Data Element Concept	A **Concept** that can be represented in the form of a **Data Element**, described independently of any particular **Representation**.
Entity	Any concrete or abstract thing of interest, including associations among things. [ISO/IEC 2382]. Also see **Object Class**.
Object	Any part of the conceivable or perceivable world. [ISO 1087]
Object Class	A set of **Objects**. A set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose **Properties** and behaviour follow the same rules.
Property	A peculiarity common to all members of an **Object Class**.
Representation	The combination of a **Value Domain**, datatype, and, if necessary, a unit of measure or a character set.
Value Domain	A set of permissible values.

Source: ISO/IEC (1999, pp. 2–9).

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