Long-term data for Europe

EURHISFIRM

D4.1: Information system and documentation standards

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

AUTHOR:

Johan POUKENS[*] *(Universiteit Antwerpen)*

APPROVED IN 2018 BY:

Jan ANNAERT *(Universiteit Antwerpen)*

Wolfgang KÖNIG *(Goethe-Universität Frankfurt)*

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

[*] I am grateful to Oliver Watteler and Wolfgang Zenk-Möltgen of GESIS for sharing their insights on metadata standards during our meeting in Köln on May 9 2018. Their input was paramount for my understanding of the Data Documentation Initiative metadata standard and the CESSDA Core Metadata Model. Jan Annaert, Frans Buelens, Elisa Grandi, Wolfgang König, Uwe Risch, Angelo Riva and Oliver Watteler commented earlier versions of this report and I have to thank them as well for their valuable feedback. David Smadja provided examples helpful in assessing the coverage of geographic datasets.

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 Scope and structure

In WP4, documentation about existing historical datasets and printed serial sources on European companies will be collected to compile an inventory of the available source material. To produce, share and store the documentation that will feed into the data and sources inventory, a data documentation (or metadata) standard and information system will be developed. **The purpose of this report is to select a metadata standard that will serve as the basis for developing an EURHISFIRM documentation standard.** This is only the first proposal, as the development of the EURHISFIRM documentation standard is a recursive trajectory. Future developments might warrant extra standards, depending on the services offered by EURHISFIRM and its data structure.

Section 2 introduces some general functions and types of metadata and data documentation standards. In section 3, existing metadata standards, will be analysed and assessed according to the requirements of EURHISFIRM. Five metadata standards were included in the analysis. Dublin Core and DataCite are general standards; da|ra, Statistical Data and Metadata Exchange (SDMX) and Data Documentation Initiative (DDI) are discipline-specific standards for the social and economic sciences. **DDI Lifecycle** is our standard of choice because is it more suitable for the documentation of historical datasets and serial sources than DDI Codebook or any of the other standards included in the discussion. DDI Lifecycle has many elements, but EURHISFIRM would only require a subset of the full DDI specification. This will warrant the creation of a DDI profile for EURHISFIRM. Section 4 proposes to build a DDI profile from the bottom up during the compilation of the inventory. This requires a software package that supports the full DDI Lifecycle standard. In section 5, the proprietary **Colectica Platform** is presented as a possible information system for collecting, organizing, storing and sharing metadata for EURHISFIRM.

2 Introduction to metadata and data documentation standards

Simply put, metadata are "data about data". The International Organisation for Standardisation (ISO) definition of metadata is "data that defines and describes other data" (Organisation for Economic Co-operation and Development, 2007, p. 73). Generally, four types of metadata are distinguished (Pearce-Moses, 2005; Riley, 2017).

- Administrative metadata refer to the acquisition of information objects such as a printed volume or a dataset (hereafter referred to as "resources"). They are instrumental for managing collections, for instance by capturing information about the provenance of a resource (e.g. purchase or loan) or conditions to access and re-use datasets (e.g. privacy issues and intellectual property rights).
- Descriptive (or reference) metadata cover the intellectual or logical content and form of the resource. Bibliographic citations and library catalogue records are common types of descriptive metadata. They facilitate resource discovery and identification (e.g. via a DOI or call number).
- Structural metadata describe the relationship between parts of a resource (e.g. chapters in a book or tables in a relational database). They support the understanding of complex resources.

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

- Preservation metadata provide the information necessary for preserving long-term access to the resource, for instance checksums for safeguarding the authenticity of digital objects.

Metadata are most useful if they are universally understandable by both human and software applications. This is accomplished through standardisation. Hence, a metadata standard is "an agreed list of common metadata items and the standardisation of terminology and definitions for these items" (OECD, 2007, p. 76). Again, three types of metadata standards are distinguished (Pearce-Moses, 2005; Gregory, Heus, & Ryssevik, 2009).

- Content standards are formal rules that specify the content, order and syntax of information in data elements to promote consistency (e.g. ISO-8601 for the representation of dates).
- Data structure standards contain formal guidelines specifying the elements or categories into which data are to be organized (e.g. Encoded Archival Description for encoding archival inventories in XML).
- Data value standards are lists of normalized terms (controlled vocabularies) which can (and often must) be used for describing data elements to ensure consistency (e.g. ISO-639 for the representation of languages).

Metadata and metadata standards for the documentation of scientific data have four functions (Gregory et al., 2009; Willis, Greenberg, & White, 2012):

- Discovery and access of data: Metadata facilitate finding and retrieving datasets. Metadata standards ensure sufficient information is available to users. Resource discovery is also made more effective by applying controlled vocabularies in metadata schemas.
- Interchange (exchange): Metadata standards allow users (both human and machine) to share information. Achieving semantic interoperability (i.e. exchange of data between, computer systems) is an important goal of metadata standardisation. Currently, the XML (Extensible Mark-up Language) and RDF (Resource Description Framework) standards facilitate semantic interoperability. DTDs (document type description) and XSDs (XML schema definition) formally describe the structure of an XML document. XML encoded (meta)data can be validated against a DTD or XSD to check conformity to a standard. RDF (Resource description framework) offers next level semantic interoperability. Unlike XML, the RDF model can be used for interchanging data between applications that use with differing underlying schemas (Decker et al., 2000).[1]
- Interpretation: Metadata help users understand data correctly for re-use. This is intricately linked to amount of detail contained in the metadata. Again, standards ensure sufficient information is recorded.
- Preservation: Metadata support long-term archiving of research data.

[1] https://www.w3.org/RDF. See Hoekstra et al. (2018) for use cases of RDF in socio-economic history.

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

In short, complete metadata are instrumental for meeting FAIR Data Principles (Wilkinson et al., 2016). The FAIR Data Principles were formulated to support the reuse of research data. Its four foundational principles (Findability, Accessibility, Interoperability and Re-usability) apply not only to data, but also to metadata. This can be achieved by selecting and using a standard appropriate for the form or discipline of the research data.

3 Metadata standards overview

"The nice thing about standards is that you have so many to choose from."[2] There are indeed many metadata standards out there. A few standards are generic (e.g. Dublin Core), but most standards are tailored to specific disciplines (e.g. Darwin Core for biodiversity) or resource types (e.g. MARC standards for books and other resources typically preserved in libraries or ISAD(G) for archival sources).[3] The following overview is limited to generic and social science specific metadata standards for documenting research data.[4]

Each standard will be analysed and assessed according to the requirements of EURHISFIRM. These requirements are determined by the objectives of EURHISFIRM and the role of the data and sources inventory in reaching these goals. The objective of EURHISFIRM is to collect, collate and connect, align and share detailed, high-quality long-term micro data on European companies. The inventory must contain in-depth information to assess the semantics and quality of available historical datasets and serial printed sources and to feed the elaboration of a common data model. This involves meticulous documentation of the sources of historical data, reasons for collecting data, data collection methods and procedures, database scheme and data structure, coding schemes and data modifications. Hence, the selected standard must accommodate all of these elements.

3.1 General standards

3.1.1 Dublin Core Metadata Initiative

Website: http://dublincore.org

The **DCMI Metadata Element Set** (DCMES) is an international standard (ISO-15836) since 2009. It consists of a broad and generic ("domain-agnostic") set of 15 elements or properties (e.g. creator, title, publisher or date; see appendix for all properties in DCMES). It was initially intended to facilitate the discovery of electronic resources, but can be used to describe a wide range of resources including digital and physical resources (e.g. datasets and printed sources). The elements are optional and repeatable.

[2] https://en.wikiquote.org/wiki/Andrew S. Tanenbaum

[3] An extensive list of discipline specific metadata standards and specifications is maintained by the Digital Curation Centre: http://www.dcc.ac.uk/resources/metadata-standards.

[4] As part of our research for this report, we also looked into the use of data documentation standards by financial databases such as US Stock Database from CRSP (Center for Research in Securities Prices) and the London Share Price Database (LSPD). This yielded no results, as might be expected because these commercial standalone datasets have little to gain from rigorous, standardized documentation of their data.

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

In 2000, **Qualified Dublin Core** introduced element refinements to make the meaning of an element narrower or more specific, as well as three additional elements (audience, provenance and rightsHolder).[5] DCMES and Qualified Dublin Core were superseded in 2008 by the **DCMI Metadata Terms** (DCTERMS). Elements and element refinements became properties. For instance, the properties "dcterm:spatial" and "dcterm:temporal" refine the property (element) "dcterm:coverage". Dublin Core currently also maintains a controlled vocabulary for recording the resource type in a consistent manner (DCMI Type Vocabulary or DCMITYPE).

Dublin Core metadata can be expressed using HTML, XML and RDF.[6]

Dublin Core is a basic standard that is easy to understand and implement. It is one of the most widely used standards for describing scientific resources and has been used to document both datasets (Alam, 2014) and printed serials (Jones, 2001). Because of its simplicity, Dublin Core is also often used as an exchange format. OpenAIRE, for instance, requires literature repositories to comply to a Dublin Core application profile for automated harvesting purposes.[7] Several other metadata standards also provide mappings to DCMES or DCTERMS. Dublin Core, however, lacks the rich and complex metadata elements required for capturing research data specific metadata elements (Alam, 2014). As the EURHISFIRM standard has to be able to capture very rich metadata about a data set or source, Dublin Core is too narrow. If interoperability of the EURHISFIRM inventory with other repositories will be necessary or desired in the future, the mapping of the EURHISFIRM selected standard to Dublin Core or the development of a Dublin Core Application Profile will be in order.[8] Because of the widespread adoption of Dublin Core, this can advance the visibility and discoverability of sources on European countries.

3.1.2 DataCite

Website: https://www.datacite.org

DataCite is an international consortium that supports the citation of research data. It is the official DOI registration agency for research data. In the context of DataCite, research data should be understood as datasets in their broadest form (not only numerical data, but for instance also audiovisual recordings, images, models software and text). The **DataCite Metadata Schema** is a list of 19 core metadata properties (with their sub-properties, see appendix) chosen for an accurate and consistent identification of a resource for citation and retrieval purposes (DataCite Metadata Working Group et al., 2017). DataCite also includes controlled vocabularies for a number of (sub-)properties. Like Dublin Core, DataCite is domain agnostic. The current version of DataCite Metadata Schema (4.1) is only available as an XML schema (XSD), but mappings of older versions of the schema (3.1) to RDF are available.

Most of the pros and contras to Dublin Core also apply to DataCite. DataCite for instance adds a few extra fields to Dublin Core (e.g. funding reference), but nevertheless contains a relatively small number of properties. The inclusion of a few research data specific sub-properties (e.g. methods being a sub-property

[5] http://dublincore.org/documents/2000/07/11/dcmes-qualifiers

[6] http://dublincore.org/resources/expressions

[7] https://guidelines.openaire.eu/en/latest/literature/index.html. OpenAIRE is a project funded by the European Commission to promote open science and to enhance the discoverability and reusability of research publications and data. Research output stored in numerous local institutional repositories can be cross-searched via the OpenAIRE platform.

[8] A Dublin Core Application Profile defines a set of metadata for records in a specific application in globally defined vocabularies and models (e.g. RDF) to provide semantic interoperability between applications.

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

of description) makes it only marginally better than Dublin Core in this domain. We should bear in mind, however, that the main purpose of DataCite is resource discovery. Its small number of properties does in fact increase interoperability as it keeps technical barriers for implementation low. OpenAIRE therefore requires compliant data archives to use a slightly adapted version of the DataCite Metadata Schema.[9] DataCite also actively supports interoperability by maintaining a Dublin Core application profile.

3.2 Social science standards

3.2.1 da| ra

Website: http://www.da-ra.de

da| ra is the German registration agency for social science and economic research data operated jointly by GESIS and ZBW. In cooperation with DataCite, da|ra assigns DOIs to datasets and stores the associated metadata. The **da | ra Metadata Schema** is a list of core metadata properties chosen for the identification of data and retrieval purposes. The da|ra Metadata Schema is compliant with the DataCite Metadata Schema, but adds 13 properties to accommodate domain specific metadata for the social sciences and economics (see appendix). These include for instance several properties related to survey data (universes, samplings, timeDimension and collectionModes). In total, the da| ra Metadata Schema contains 32 properties. For interoperability purposes, da|ra provides mappings to Dublin Core, DataCite and DDI. da| ra also maintains a number of controlled vocabularies which align with DataCite and DDI controlled vocabularies.

da| ra better suits the discipline specific needs and practices of the social sciences than Dublin Core or DataCite. It offers better possibilities for the structured description of survey related metadata and datasets. The dataSet property, for instance, contains sub-properties for a summary description of a dataset. These include the type and number of units (e.g. companies), the number of variables and the type of data (e.g. csv) in the dataset. da| ra also includes a notes property for capturing information that does not fit into any of the specified categories, hence increasing the flexibility of the standard. For the purpose of EURHISFIRM, using the notes property could overcome some of the limitations of da| ra, for instance for the documentation of database schemes and data structures. Extensive use of notes, however, deprecates interoperability.

3.2.2 Statistics Data and Metadata Exchange

Website: https://sdmx.org

Statistics Data and Metadata Exchange (SDMX) is an international standard (ISO-17369) designed to document statistical data and metadata. SDMX was primarily designed for exchanging aggregate data, typically time series data, between national and international statistics organizations (Gregory & Heus, 2007). Therefore, SDMX also addresses the harmonization of terms, classifications and concepts used in gathering and processing statistics.

[9] https://guidelines.openaire.eu/en/latest/data/use of datacite.html

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

SDMX is a standard for the publication and exchange of aggregated data, indicators and time series. The focus of EURHISFIRM, however, is microdata (although frequently in time series format). The unit of observation is the individual, in this case the company.

3.2.3 Data Documentation Initiative

Website: https://www.ddialliance.org

The Data Documentation Initiative (DDI) is standard for describing the data 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) to replace the OSIRIS codebook standard. The DDI standard is currently maintained by the DDI Alliance and is widely used by social science data archives. Although DDI is primarily designed to describe microdata (e.g. survey data), it is also possible to capture information about aggregate data in multidimensional tables (data cubes). DDI also provides a set of controlled vocabularies that can be used with DDI.

Since 2008, DDI has provided two versions of its standard: **DDI Codebook** (currently version 2.5) and **DDI Lifecycle** (currently version 3.2). DDI-Codebook (also referred to as DDI-C or DDI2.X) can be used to document a single data collection, DDI-Lifecycle (also referred to as DDI-L or DDI3.X) metadata supports the entire research data lifecycle (from planning to archiving). Because DDI aims to define a superset of all possible elements and attributes used to describe social science data resources, the standard in general and DDI-L is particular contains many elements. The current versions of DDI-C and DDI-L respectively have 351 and 1,154 elements.[10] The full DDI specifications are currently only available as XSDs (XML schemas). The DDI Alliance, however, is in the process of developing a number of RDF vocabularies. A subset of DDI-C and DDI-L elements will be made available in the DDI-RDF Discovery Vocabulary (Disco). This subset contains essential metadata for the discovery of microdata as linked (open) data. Disco is currently an unofficial draft.[11] The next generation of DDI, DDI4 will be a model-based standard which can be expressed both as an XML schema and an RDF vocabulary (Gregory & Wackerow, 2014).

DDI-C elements are organised in five sections which correspond to the chapters of a printed codebook:

1. Document description: bibliographic description on the DDI document itself (metadata about the metadata, or "mete-metadata") (Ryssevik, 2001).
2. Study description: information about the dataset. Study level descriptions typically cover the type of elements also included in Dublin Core, DataCite and da|ra (creator, title, coverage, etcetera).
3. Data files description: information about the individual files in the dataset.
4. Variable description: information about the variables in a data file.

[10] Field level documentation of the DDI standards is available online: http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field level documentation.html ; http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation

[11] http://www.ddialliance.org/Specification/RDF/Discovery

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

5. Other materials: documents containing information related to the dataset.

DDI-L elements are arranged according to modules which roughly relate to stages in the research data lifecycle (e.g. data collection or archiving) or publishing packages (e.g. studies or series, an example of a Series-level description of a resource in DDI3.2-XML is included in the appendices). The studyUnit publishing package describes a single study and most closely corresponds to DDI-C. In DDI-L, studies with shared features (e.g. subsequent waves of a longitudinal study) can be part of a Group which contains the mutual metadata for the studies in the group. Through the introduction of schemes, DDI- L also facilitates the reuse of descriptions for common concepts or variables.

The high number of elements in DDI is illustrative for the level of detail in which it is possible to describe data. Basic functionality for the identification of sources pertaining to European companies can be found in all of the above standards, but the EURHISFIRM selected standard must allow for an exhaustive documentation of various aspects of a dataset such as data collection (including the sources used in compiling the dataset), format, structure and variables. This is only possible in DDI. Furthermore, DDI is also the standard of choice for CESSDA. In 2001, CESSDA defined a list of mandatory and recommended DDI-C elements.[12] CESSDA's project group on Metadata Management recently (May 2017) agreed on version 1.0 of the Core Metadata Model which is compatible with DDI- L.[13] CESSDA members such as GESIS-Leibniz Institute for the Social Sciences naturally have aligned their own metadata schemas with DDI. GESIS's Datenbestandskatalog Metadatenschema for instance is DDI-C compliant (Zenk-Moltgen & Habbel, 2012) and metadata for historical time series contained in the histat database can be exported as DDI 3.1 XML. Since the EURHISFIRM standard of choice must meet CESSDA's requirements, DDI is a natural choice for EURHISFIRM.

DDI-L is better suited for documenting historical datasets and printed serial sources than DDI-C (or any of the standards described above). A few examples can sufficiently illustrate this point:

- The possibility to group studies makes DDI-L better suited for serials description. Printed serial sources can run for very long periods of time. Over time, its form or contents may change. The *Belgisch Staatsblad,* for instance has been published exclusively on paper from 1830 until 1996, was available on paper and online from 1997 until 2002 and is only available online since 2003. By grouping different manifestations of a resource, only unique characteristics need to be recorded at the study level in DDI-L. Mutual characteristics are described once at the group level and inherited by the studies of a group. In DDI-C, however, shared metadata would have to be repeated.

- DDI-C does not allow a date range in the publicationDate element. For the description of serials, however, this is of cardinal importance. In DDI-L, date ranges can be recorded in all date elements by using a start and/or end date (hence also allowing open date ranges for running serials). Date elements can also be repeated, for instance for handling interruptions in the publication of a serial or issues missing in a collection.

[12] https://www.ddialliance.org/sites/default/files/cessda-rec.pdf

[13] https://www.cessda.eu/Consortium/Communication/News/CESSDA/CMM-1.0-CESSDA-Metadata- Management

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

- DDI-C follows ISO-8601 conventions for formatting dates. ISO-8601, however, assumes all dates are expressed in the Gregorian calendar. While most predominantly catholic or protestant countries in western and central Europe adopted the Gregorian calendar well before 1815, some countries in eastern and southern Europe retained the Julian calendar until the twentieth century (Latvia and Lithuania adopted the Gregorian calendar in 1915, Bulgaria in 1916, Estonia in 1918, Romania in 1919 and Greece in 1923). Dates in nineteenth-century printed sources from these countries can be expected to be expressed in the Julian calendar. The inclusion of an historicalDate element with a calendar attribute in DDI-L avoids the need to convert dates to the Gregorian calendar. The NonISODate element could also be used to record uncertain dates because it allows dates expressed in a non-ISO compliant structure (e.g. a non-serial source that has been published somewhere between 1890 and 1893 could be recorded here as [1890-1893] whereby the square brackets are a conventional method to indicate uncertainty).

DDI-L also integrates ISO-11179. This is a framework for the specification and standardization of data elements. By mapping the core structures Concept, Universe, ConceptualVariable, and RepresentedVariable to the ISO-11179 Data Element Classification Structure, DDI-L provides functionality for the harmonization of variables in support of data semantics analysis. For instance: the conceptual variable (or data element concept) 'price of securities' can have several representations (e.g. 'opening price' or 'closing price'). The RepresentedVariable ties the concept ('price') and the universe or object ('securities') to its different representations.

4 EURHISFIRM DDI-L profile

The DDI standard was developed by data archives to capture information about survey data. The elements and logic of the standard are geared towards this data type (Ryssevik, 2001). For instance, the utility of elements pertaining to questions in a survey for EURHISFIRM is limited. It is a best practice to work with an established set of DDI objects rather than with the entire specification if this is not necessary. A DDI-L profile contains the elements and attributes used and/or not used by a community or organisation (Ionescu, 2009).

We propose to incrementally build a DDI-L profile for EURHISFIRM as WP4 progresses, adding modules and elements as and when needed. The first task at hand involves compiling an inventory of data and sources on listed companies in the post-1815 period in the countries of the consortium (task 4.2). The primary goal of the inventory is to discover and identify datasets and printed serial sources. This is typically done with the Group, StudyUnit and Archive modules. In addition, we also recommend extensive documentation of the DDI instance (i.e. the document description section in DDI-C) because the data and sources inventory will represent a considerable intellectual effort in its own right. The DDI Instance can be credited and cited properly using the Instance module. Moving further into WP4, necessary additional modules will probably include LogicalProduct for the documentation of variables and the analysis of data semantics and DataCollection for assessing source quality. Software tools in support of such a bottom-up approach are available. DDIProfileSXLT for instance is a software tool for creating DDI-L profiles from

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

existing DDI instances.[14] This profile should be documented in detail in a user manual that will ensure a correct use of the standard by users from different countries and traditions.

In addition to developing a DDI-L profile, we will also evaluate the suitability of DDI controlled vocabularies, as well as other controlled vocabularies for documenting historical data and select those most appropriate for inclusion in the EURHISFIRM DDI-L profile. For instance, the Getty Thesaurus of Geographic Names (Getty Research Institute) will be most suitable for historical data because it includes the possibility to reference former states (e.g. Prussia or Austria-Hungary) and former names of places (e.g. Breslau for Wroclaw in Poland).[15] Other comprehensive geographic datasets such as GeoNames[16] and the GeoNames Search of the US National Geospatial-Intelligence Agency[17] do not offer the same historical depth as the Thesaurus of Geographic Names (TGN) at the level of countries and regions. Some tests show that historical variations of cities, however, are covered in both, partially overlapping, datasets. These datasets also include more foreign language variants of a place which can be of particular interest to EURHISFIRM at a later phase, when datasets using a particular language to describe places (for instance French in the Paris Database) are added to the infrastructure. For the present purpose, data documentation, TGN will probably suffice because this requires mostly high-level description of geographic coverage (e.g. at the national or regional level). Moreover, TGN allows external contributions of missing places to the dataset. Contributions can be made by registered users via a web form and are processed and incorporated in the dataset within approximately two months.

5 Information system

The task at hand also requires an information system for collecting, organizing, storing and sharing metadata. As DDI is an established standard in the social sciences community, a fair number of software tools, both proprietary and open source, have been developed to support compliant data documentation. Editors are used for documenting datasets, repositories for storing metadata records and catalogues for searching through a set of records.

5.1 Open source: The Dataverse Project

A dataverse is a virtual archive or collection of related datasets and their metadata (e.g. datasets pertaining to the same project). In the Dataverse Project, Harvard University's Institute for Quantitative Social Science (IQSS) has created an open source web application to share, preserve, discover, cite and analyse research data (King, 2007).[18] Institutions can install Dataverse on their own servers (the software can be downloaded from GitHub) or create their own cloud-based dataverse within the Harvard Dataverse Network. Cloud-based dataverses are customisable to a certain extent (e.g. select optional and mandatory metadata elements and create templates). Researchers can easily upload and document datasets via a web browser. Workflows for contributing, checking and publishing datasets are available.

[14] https://www.ddialliance.org/node/919
[15] http://www.getty.edu/research/tools/vocabularies/tgn/index.html
[16] http://www.geonames.org
[17] http://geonames.nga.mil/namesviewer
[18] https://dataverse.org

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

Dataverse metadata are much more extensive than that of similar repositories for datasets (Dataverse contains 100 metadata elements, Figshare, for instance, only 12) (Farnel & Shiri, 2014). Dataverse metadata consist of a combination of generic citation metadata, discipline specific metadata (e.g. geospatial metadata or social science metadata) and file level metadata. Citation metadata, geospatial metadata and social science metadata are Dublin Core, DataCite and DDI Codebook 2.5 compliant. A mapping to DDI Codebook 2.5 is available and Dataverse can also export metadata as DDI-XML. Dataverse's lack of support for DDI Lifecycle, however, makes it less suitable as an information system for metadata management in EURHISFIRM. All of the above limitations of DDI Codebook also apply to Dataverse. Moreover, Dataverse metadata do not leverage the full potential of the DDI Codebook specification. For instance, elements for variable description, crucial for data semantics analysis, are not covered by Dataverse metadata elements. Finally, Dataverse only supports the ISO-3166 controlled vocabulary for country names. It is not possible to reference TGN or other geographic datasets in geospatial metadata elements (the keywords elements group in the citation metadata does allow referencing discipline specific controlled vocabularies).

5.2 Proprietary: Colectica Platform

Colectica offers editing, storing and sharing functionality through its Colectica Platform.[19] Colectica Designer is the editing client at the heart of the platform. It supports all versions of DDI (including DDI 2.5 and DDI 3.2). Colectica Designer requires Colectica Repository and Colectica Portal for publishing data documentation to a server and indexing it for search via a website. Colectica is proprietary software and each user would need an individual licence to create or edit metadata. A licence to Colectica Designer costs 59 USD per month, perpetual licences are also available for 2000 USD. Free software, however, is available for viewing metadata (Colectica Reader). Judging from the user manuals and use cases available online, Colectica Designer offers most of the functionality required for EURHISFIRM. Further testing is warranted for evaluating for instance the support of controlled vocabularies. A trial has been requested from the developer for this purpose.

A common concern with proprietary software is vendor lock-in. If EURHISFIRM metadata were stored in proprietary formats, this would result in substantial switching costs and reduce possibilities for interoperability. In the case of Colectica, however, this is not an issue. Colectica exports metadata in various open DDI-XML formats which can be re-used by other DDI editors or DDI compliant repositories.

6 Relation to WPs in EURHISFIRM

The standardisation decision impacts and benefits the work of other work packages in EURHISFIRM. Standardised metadata allow those involved in other work packages to quickly locate and correctly interpret information pertinent to their own tasks from the metadata of a dataset. The labels and contents of elements in DDI are well defined leaving little room for conjectures about their meaning. The interpretation of (meta)data is furthered by referencing controlled vocabularies and including code lists in

[19] 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 metadata of a resource as much as possible. Terms from controlled vocabularies for instance can often be expressed in multiple languages.

Possible uses of standardised metadata include for instance:

- Ownership (e.g. intellectual property rights) and access restrictions (e.g. personal data) to resources for tacking stock of property rights and the development of an ethics code in WP3
- Coding systems (e.g. NACE-codes for classifications of economic sectors) used in resources for establishing national practices and developing data models in WP5
- Unique identifiers (e.g. ISIN-codes for the identification of securities) used in resources for matching data in WP6
- Language and script used in resources for automated extraction of data in WP7

7 Abbreviations

CESSDA: Consortium of European Social Science Data Archives

DCMI: Dublin Core Metadata Initiative

DDI: Data Documentation Initiative

DOI: Digital Object Identifier

GESIS: Leibniz Institute for the Social Sciences

ISAD(G): General International Standard for Archival Description

ISO: International Organization for Standardization

MARC: Machine Readable Cataloging

RDF: Resource Description Framework

TGN: Thesaurus of Geographical Names

XML: Extensible Markup Language

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

8 References

Alam, A. W. (2014). Dublin Core Metadata for Research Data: Lessons Learned in a Real-world Scenario with Datorium. In *Proceedings of the 2014 International Conference on Dublin Core and Metadata Applications* (pp. 64-73). Austin, Texas: Dublin Core Metadata Initiative. Retrieved from http://dl.acm.org/citation.cfm?id=2771234.2771242

DataCite Metadata Working Group, Ashton, J., Barton, A., Birt, N., Dietiker, S., Elliot, J., ... Zolly, L. (2017). *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1.* Hannover: DataCite. https://doi.org/10.5438/0014

Decker, S., Melnik, S., Harmelen, F. van, Fensel, D., Klein, M., Broekstra, J., ... Horrocks, I. (2000). The Semantic Web: the roles of XML and RDF. *IEEE Internet Computing,* 4(5), 63-73. https://doi.org/10.1109/4236.877487

Farnel, S., & Shiri, A. (2014). Metadata for Research Data: Current Practices and Trends. *International Conference on Dublin Core and Metadata Applications,* 74-82.

Gregory, A., & Heus, P. (2007). *DDI and SDMX: Complementary, not competing standards.* Tucson, Arizona: Open Data Foundation. Retrieved from http://www.opendatafoundation.org/papers/DDI_and_SDMX.pdf

Gregory, A., Heus, P., & Ryssevik, J. (2009). *Metadata* (RatSWD Working Papers Series No. 57). Berlin: German Council for Social and Economic Data. Retrieved from https://docs.google.com/viewer?url=http://www.ratswd.de/download/workingpapers2009/57_09.pdf

Gregory, A., & Wackerow, J. (2014, April). *DDI Discovery: An Overview of Current RDF Vocabularies.* Presented at the North American DDI (NADDI) Conference 2014. Retrieved from http://summit.sfu.ca/item/13930

Hoekstra, R., Merono-Penuela, A., Rijpma, A., Zijdeman, R., Ashkpour, A., Dentler, K., ... Rietveld, L. (2017). The dataLegend ecosystem for historical statistics. *Journal of Web Semantics, 50,* 4961. https://doi.org/10.1016Zj.websem.2018.03.001

Ionescu, S. (2009). *Creating a DDI Profile* (DDI Alliance Working Papers Series. Best Practices No. 6). DDI Alliance. Retrieved from http://doi.org/10.3886/DDIBestPractices06

Jones, W. (2001). Dublin Core and Serials. *Journal of Internet Cataloging,* 4(1-2), 143-148. https://doi.org/10.1300/J141v04n01_13

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

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

Pearce-Moses, R. (2005). *A glossary of archival and records terminology.* Chicago: Society of American Archivists. Retrieved from https://www2.archivists.org/glossary

Riley, J. (2017). *Understanding metadata: What is metadata, and what is it for?* Baltimore: National Information Standards Organization (NISO). Retrieved from http://www.niso.org/publications/understanding-metadata-riley

Ryssevik, J. (2001). *The Data Documentation Initiative (DDI) metadata specification.* Ann Arbor: Data Documentation Initiative. Retrieved from http://www.ddialliance.org/sites/default/files/ryssevik_0.pdf

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

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

Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science & Technology,* 63(8), 15051520. https://doi.org/10.1002/asi.22683

Zenk-Möltgen, W., & Habbel, N. (2012). *Der GESIS Datenbestandskatalog und sein Metadatenschema* (GESIS Technical Reports No. 2012-12). Köln: GESIS-Leibniz-Institut für Sozialwissenschaften. Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-292372

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

9 Appendices

9.1 Dublin Core Metadata Element set

Element	Definition
contributor	An entity responsible for making contributions to the resource.
coverage	The spatial or temporal topic of the resource.
creator	An entity primarily responsible for making the resource.
date	A point or period of time associated with an event in the lifecycle of the resource.
description	An account of the resource.
format	The file format, physical medium, or dimensions of the resource.
identifier	An unambiguous reference to the resource within a given context.
language	A language of the resource.
publisher	An entity responsible for making the resource available.
relation	A related resource.
rights	Information about rights held in and over the resource.
source	A related resource from which the described resource is derived.
subject	The topic of the resource.
title	A name given to the resource.
type	The nature or genre of the resource.

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

9.2 DataCite and Da|ra MetaData Schema

ID	DataCite property	ID	Da|ra property
1	Identifier	3	resourceIdentifier
2	Creator	7	creators
3	Title	4	titles
		5	otherTitles
4	Publisher	12	publisher
5	PublicationYear	10	publicationYear
6	Subject	17	classifications
		18	controlledKeywords
		19	freeKeywords
7	Contributor	26	contributors
8	Date	24	temporalCoverages
9	Language	15	resourceLanguage
10	ResourceType	1	resourceType
		2	resourceTypeFree
11	AlternateIdentifier	16	alternativeIDs
12	RelatedIdentifier	31	relations
13	Size	29	dataSets
14	Format	29	dataSets
15	Version		
16	Rights	14	rights
17	Description	6	collectiveTitles
		20	descriptions
		22	universes
		23	samplings
		28	collectionModes
18	GeoLocation	21	geographicalCoverages
19	FundingReference	27	fundingReferences
		8	dataURLs
		9	doiProposal
		11	publicationPlace
		13	availability
		25	timeDimensions
		30	notes
		32	publications

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

9.3 Example of a Series-level description in DDI3.2-XML

```
<Group versionDate="2018-06-19T12:06:15.7807216Z">
        <r:URN>urn:ddi:example.org:128d6947-e057-4b2f-8fa4-dd035e5f6ecb:1</r:URN>
        <r:Agency>example.org</r:Agency>
        <r:ID>128d6947-e057-4b2f-8fa4-dd035e5f6ecb</r:ID>
        <r:Version>1</r:Version>
        <r:UserID typeOfUserID="creator">Poukens, Johan</r:UserID>
<r:Citation>
        <r:Title>Le recueil financier</r:Title>
        <r:SubTitle>Annuaire des valeurs cotées aux bourses de Belgique</r:SubTitle>
        <r:Publisher>
                <r:PublisherName>Emile Bruylant</r:PublisherName>
                </r:Publisher>
        <r:PublicationDate>
                <r:StartDate>1893</r:StartDate>
                <r:EndDate>1975</r:EndDate>
        </r:PublicationDate>
</r:Citation>
<r:Abstract>
        <r:Content xml:lang="en">The Receuil financier covers governance and financial information about companies listed on the Brussels Stock Exchange. Governance information includes the location of the company's registered office, telegraphic address, names of managers and directors, balance sheet date, general assemby date, judicial statute and charter date, financial services providers, puropose, capital and a chronicle of important events. Financial information includes the most recently available balance sheet and operating results. Most volumes are accompanied by an appendix with additional information about managers and directors (Table alphabétique des administrateurs et commissaires), listing for instance their address and occupation.</r:Content>
</r:Abstract>
<r:Coverage>
        <r:TopicalCoverage>
                <r:Subject>Corporations</r:Subject>
                <r:Subject>Securities</r:Subject>
        </r:TopicalCoverage>
        <r:SpatialCoverage>
                <r:Description>
                        <r:Content xml:lang="en">Belgium</r:Content>
                </r:Description>
                <r:Country>BE</r:Country>
        </r:SpatialCoverage>
        <r:TemporalCoverage>
                <r:ReferenceDate>
                        <r:StartDate>1893</r:StartDate>
                        <r:EndDate>1975</r:EndDate>
                </r:ReferenceDate>
        </r:TemporalCoverage>
</r:Coverage>
</Group>
```

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