gesis Leibniz-Institut für Sozialwissenschaften

GESIS Papers

2018|18

Microdata Information System MISSY:

Benefits for Research with Official Microdata, DDI-Based Implementation, and Evaluation with Regard to FAIR Criteria

Jeanette Bohr, Anne Balz, Florian Thirolf & Matthäus Zloch

GESIS Papers 2018|**18**

Microdata Information System MISSY:

Benefits for Research with Official Microdata, DDI-Based Implementation, and Evaluation with Regard to FAIR Criteria

Jeanette Bohr, Anne Balz, Florian Thirolf & Matthäus Zloch

GESIS – Leibniz-Institut für Sozialwissenschaften 2018

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Dauerbeobachtung der Gesellschaft
German Microdata Lab
Postfach 12 21 55
68072 Mannheim
E-Mail: jeanette.bohr@gesis.org

ISSN: 2364-3781 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

Table of Contens

Summary 4

1 Introduction 5

1.1 Structure and objectives of the paper 5

1.2 Documentation requirements for official microdata 5

2 Metadata on official microdata for research in MISSY: A service offered by the German Microdata Lab at GESIS 7

2.1 Purpose of MISSY: Added value for researchers 7

2.2 Contents of MISSY 8

2.2.1 Documented microdata 8

2.2.2 MISSY metadata schema 8

2.2.3 Access to metadata via MISSY 10

3 Technical implementation 14

3.1 Summarizing description of the implementation 14

3.2 Architecture 15

3.2.1 Data model 15

3.2.2 Database 16

3.3 Import interfaces 17

3.4 MISSY editor 18

3.5 Representation and export 19

4 MISSY with respect to the FAIR criteria 20

4.1 Findability 20

4.2 Accessibility 21

4.3 Interoperability 21

4.4 Reusability 22

4.5 Conclusion: Is MISSY FAIR? 22

References 24

Appendix: MISSY metadata schema 25

Summary

This paper presents the microdata information system (MISSY). MISSY is a service of the German Microdata Lab (GML) for empirically working scientists conducting research using microdata from official statistics. MISSY provides detailed metadata on individual data sets from the German (Microcensus) and European official statistics (e.g. EU-SILC, EU-LFS) and aims to facilitate the use of the data through user-friendly and quickly accessible data documentation.

We address the documentation requirements of official microdata, elaborate the benefits of structured metadata for researchers and describe the resulting objectives and contents of MISSY. Subsequently, we introduce the specific technical implementation: A general description of the technical infrastructure as well as the basic data model (DDI-based) and the import/export interfaces of the database. Finally, we discuss MISSY with regard to the FAIR criteria and show how MISSY contributes to official microdata being "FAIR".

1 Introduction

1.1 Structure and objectives of the paper

This paper examines the microdata information system MISSY[1] from different perspectives.

MISSY – a service offered by the German Microdata Lab at GESIS – is aimed at empirically working researchers who use microdata from official statistics, and meets the specific documentation requirements of this data for scientific purposes. In addition, MISSY is a technical system that provides interfaces to support the internal preparation processes of the metadata and to produce the representation of its content in the web portal.[2]

After the introductory section on the documentation requirements of microdata from official statistics for research, Section 2 explains the objectives of MISSY and the added value of a structured metadata service for researchers. Also, it discusses the scope and structure of the contained metadata as well as the structure of the MISSY online service (MISSY web). Section 3 presents the technical implementation of MISSY – in addition to a general description of the technical infrastructure, it explains the underlying (DDI-based) data model and the import and export interfaces of the database. Finally, Section 4 discusses MISSY with respect to the FAIR (findable [F], accessible [A], interoperable [I], and reusable [R]) criteria, and shows how MISSY contributes to making official statistics microdata "FAIR."

This paper is aimed at different target groups. Researchers who work with official microdata can get an overview of the metadata available online in MISSY (Section 2). The description of the technical infrastructure (Section 3) has a more information science perspective and is addressed, for example, to interested metadata producers and system developers. The information on FAIR (Section 4) is intended for all users and providers of research data who are interested in implementing the FAIR guidelines for metadata.

1.2 Documentation requirements for official microdata

Selected microdata[3] are made available to science by official statistics for research purposes. An important, and already established, data source for empirical social and economic research is the German Microcensus, which provides the official representative statistics on the population and the labor market in Germany. With the development of European survey programs, European statistics are increasingly offering the social sciences international comparative datasets (e.g., the Labour Force Survey [EU-LFS] or the European Statistics on Income and Living Conditions [EU-SILC]).[4]

Official microdata for scientific purposes is published by official statistics in anonymous form called *scientific use files*. These files usually are repeated cross-sections or longitudinal studies that can be used for analyses over time. The data cover the standard topics of social structure analysis: household

---

1 See https://www.gesis.org/en/missy/.

2 The department of Knowledge Technologies for the Social Sciences (WTS) at GESIS is responsible for the technical development of this system.

3 Microdata contain information on persons, households, dwellings, or companies; and in official statistics, they are the basis for the production of aggregated data. Under certain conditions, the microdata are made available for research purposes.

4 For further available microdata on European statistics, see http://ec.europa.eu/eurostat/web/microdata/overview.

and family, labor market participation, education, migration, and socio-economic situation. Their advantages are their sample size – which makes possible the study of small subpopulations – high response rate, and the recording of all persons in a household.

An appropriate analysis of the data requires an extensive knowledge of the data. Since official statistics microdata is not collected primarily for research purposes, but rather for political and administrative purposes, the available documentation material often does not meet the needs of researchers. Therefore, a necessary prerequisite for a competent data analysis is an extensive induction into the data and the documentation material. The more frequently official microdata is used by social research, the higher the benefit of central access to the required metadata. In this context, the German Microdata Lab (GML)[5] of GESIS provides a comprehensive scientific service for microdata for German and European statistics, which includes routines for data preparation, the online metadata portal MISSY, and workshops for working with the data. The use of the routines prepared by the GML enables researchers to read and process the data correctly without great effort, and thus leads to a relevant simplification of work. Eurostat microdata files are delivered to researchers as comma separated values (CSV). The raw data files include neither variable nor value labels. The setup files can read CSV-data into SPSS or Stata and attach variable and value labels. For the German Microcensus, setup files are also available for SAS. In addition, a number of tools are available (for SPSS and Stata) to operationalize social science concepts and official constructs with the data.

As a research institution, the GML is independent of statistical offices and other data producers and acts as an intermediary between empirical social research and official statistics. This service for official data is provided in cooperation with the Federal Statistical Office and Eurostat.[6]

[5] See https://www.gesis.org/en/en/gml/gml-home/.

[6] See also the workshops and user conferences organized jointly with the German Federal Statistical Office: https://www.gesis.org/angebot/veranstaltungen/veranstaltungsarchiv/german-microdata-lab/.

2 Metadata on official microdata for research in MISSY: A service offered by the German Microdata Lab at GESIS

2.1 Purpose of MISSY: Added value for researchers

MISSY originated from the idea to provide all relevant information for the scientific use of official data centrally and systematically in an online metadata portal. As a service of the GML at GESIS, MISSY is addressed to all scientists who work with data from official statistics, and it facilitates the use of official microdata. MISSY was established and continuously expanded during several project phases, among others financed by the *Federal Ministry of Education and Research* and the *German Research Foundation*.[7]

The first version of MISSY focused on the metadata for the scientific use files of the German Microcensus (Bohr et al. 2010). As part of the Data without Boundaries Project (DwB)[8] – an EU project to improve access to official microdata within the EU – the second version of documentation on microdata for European statistics was developed (DwB 2015a). At the beginning of 2016, the two MISSY services were integrated into a combined system with a shared technical infrastructure, so all the structured metadata services provided by the GML were now accessible online in one system. The documentation refers to the scientific use files made available for research purposes. The metadata is updated regularly by the latest survey rounds (usually annually).

To reduce the effort of individual researchers, the relevant metadata and the syntax routines for data preparation and analysis created by the GML are made quickly retrievable in MISSY. One of the benefits of the MISSY system is the possibility to link information between years and countries, which eliminates time-consuming searches for information in year-specific and country-specific PDF documents. The systematic preparation of metadata leads to significant time savings, especially if several survey years or countries are being analyzed. By using the program routines offered, time-consuming recoding work related to the implementation of social science concepts is no longer necessary.

MISSY supports researchers in different phases of the research process:

- Finding suitable data: Answering the question of whether the studies documented in MISSY are suitable for answering a specific research question (e.g., using a thematic search for variables or testing whether case numbers are sufficient for specific years or countries).
- Preparing data: For example, using setup files to read and label the raw data in statistics software programs, or using tools to implement social science classifications.
- Understanding data: Detailed information at the study-, country-, variable-, and question-level, as well as comparability over time and between participating countries.

The target group of MISSY mainly consists of empirically working social scientists who compare different European countries or who work with several survey rounds. The usage figures of the past years prove an intensive use of both the metadata on the German Microcensus and the documentation on European microdata (Dragon & Zvezdanova 2018: 7).

---

[7] For the history of the MISSY project, see https://www.gesis.org/en/missy/general-information/project-history/.

[8] See http://www.dwbproject.org/.

2.2 Contents of MISSY

2.2.1 Documented microdata

The metadata of the following surveys are currently available in MISSY:

- National data:
  - MZ (German Microcensus)
- European data:
  - EU-LFS (European Union Labour Force Survey)
  - EU-SILC (European Union Statistics on Income and Living Conditions)
  - AES (Adult Education Survey)
  - CIS (Community Innovation Survey)
  - SES (Structure of Earnings Survey)

The metadata for the Microcensus, EU-LFS, EU-SILC, and AES are updated regularly following the publication frequency of the microdata. For CIS and SES, only the metadata created in the Data without Boundaries project (project duration 05/2011 to 04/2015) is currently available. The system is open for the documentation of further datasets that are used by the research community.

The metadata for the Microcensus are prepared in close cooperation with the Microcensus group of the Federal Statistical Office (Statistisches Bundesamt). Since the scientific use files of the Microcensus are jointly developed by the GML and the Federal Statistical Office, part of the metadata already is recorded during data preparation. Metadata on EU data is produced after the publication of the microdata files by Eurostat.

2.2.2 MISSY metadata schema

The metadata offered in MISSY includes all important information relevant to analyze the studies. The MISSY metadata schema is based on the expertise of the GML in the field of official microdata and decisions about the design of the schema were made in accordance with the recommendations of the international documentation standard DDI[9]. The use of the DDI standard ensures consistent documentation of the various datasets. In addition, the possibility of a flexible adaptation for the specific documentation requirements of individual datasets has been taken into account[10]. The metadata schema for EU data was defined in consultation with the Data without Boundaries project (see DwB 2015b).

MISSY contains metadata on several levels, organized in a hierarchical structure (see Figure 1). Each level provides a set of metadata elements that can be selected or hidden according to documentation requirements (see the complete MISSY metadata schema in the Appendix).

[9] Data Documentation Initiative (DDI), see https://www.ddialliance.org/.

[10] Details on the implementation of the metadata schema in the MISSY data model can be found in Section 3.2.1.

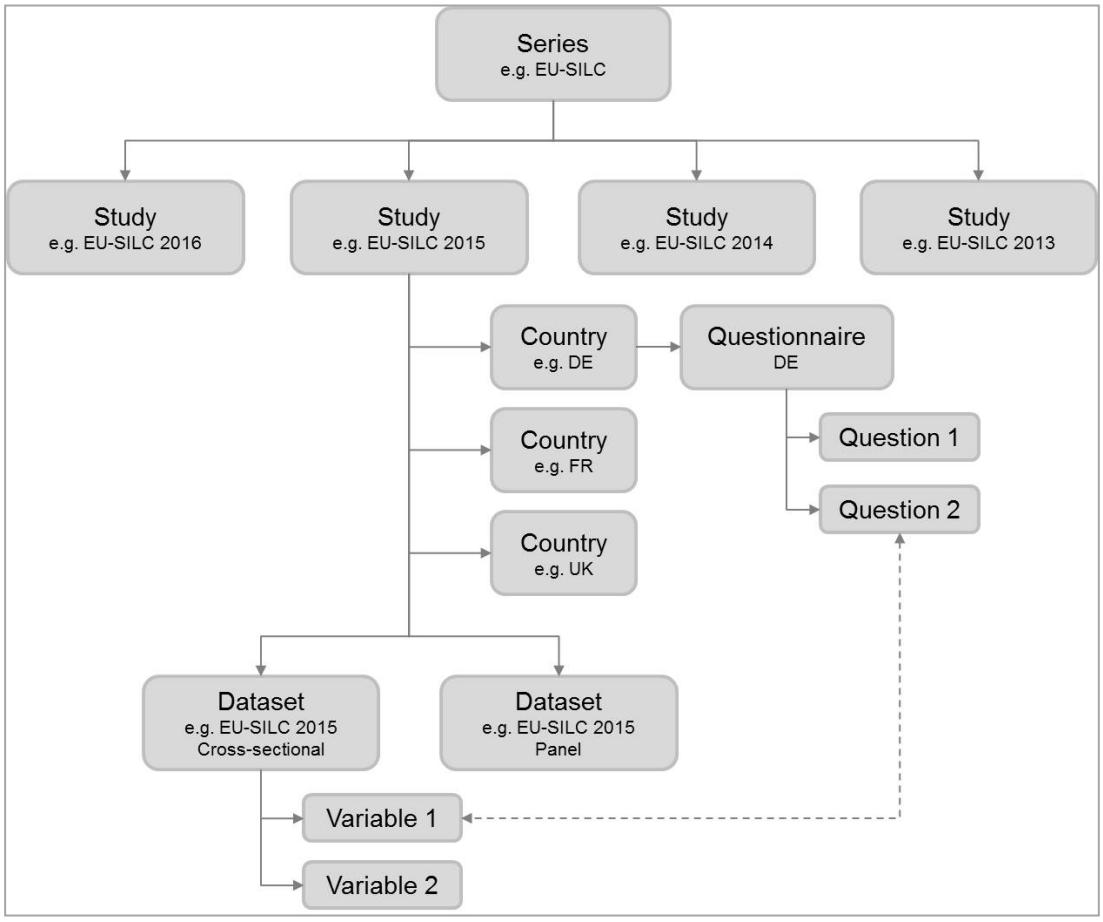


Figure 1: MISSY metadata schema (simplified visualization)

At the top level (Series), general information on the study collection is collected (e.g., abstract information on data access and legal principles). The second level (Study) refers to a single instance of the series (usually a survey year), which provides year-specific information (e.g., on sampling, data collection, extrapolation). For EU data, various information is collected for each of the countries available in a year (Country). For each study, the underlying question texts from the country-specific questionnaires (Questionnaire) also are documented at the country level. A mapping between variables and national questionnaires illustrates how the different concepts were collected. In the case of EU data, the question texts may differ between the participating countries. Currently, only the question texts for the UK and DE are included as examples. The third level (Dataset) distinguishes between the available datasets of a study that are available for analysis (for the EU-SILC, e.g., cross-sectional and panel). The dataset level contains the metadata for all variables in the dataset. In addition to variable-specific information and explanations, simple unweighted frequencies are provided for each variable (or descriptive statistics for metric variables). The underlying question texts also are linked.

Each of the variables also is assigned to a thematic classification that applies to all documented variables and years of a study collection (not shown in Figure 1). The assignment to a thematic classification is necessary to enable variables to be linked across studies and to generate thematically-sorted variable lists.

2.2.3 Access to metadata via MISSY

The structured organization of metadata in the database also is evident in the presentation of metadata in the MISSY web.

Access to this metadata is hierarchically structured and facilitates a quick retrieval of the required information. After selecting a series, the metadata for the studies and datasets documented can be chosen from a navigation menu. In the case of EU data, selected information also is collected at the study level for each of the countries included in the data. (see Figure 2). The web view shows at a glance the countries for which data is available in the selected dataset. By clicking on individual countries, country-specific metadata can be compared.

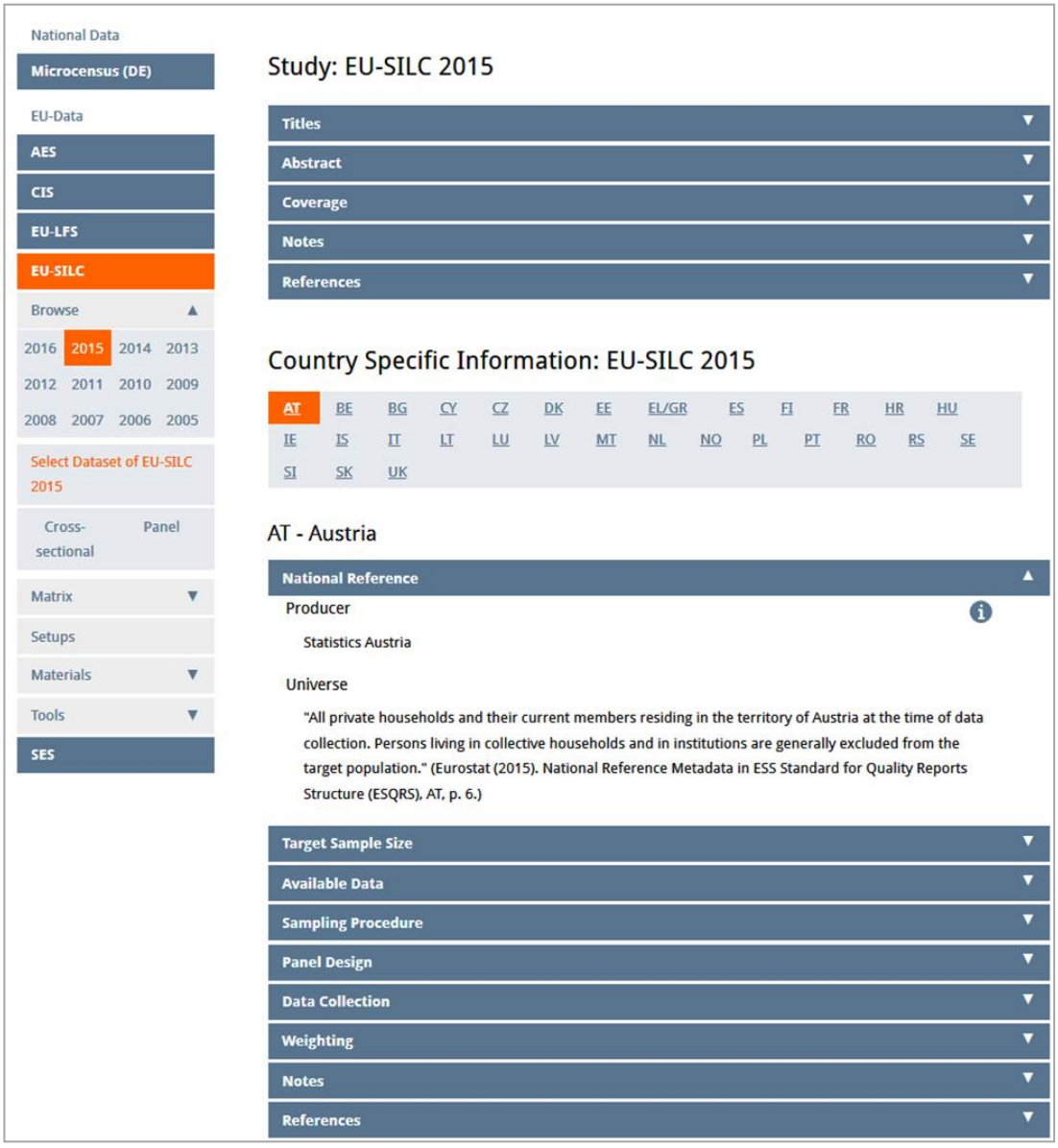


Figure 2: Screenshot country-specific information at study level (EU-SILC)

The variable lists of documented datasets can be accessed in the original sorting of the dataset, as well as through a thematic sorting (see Figure 3). A thematic classification of variables is available for all series documented in MISSY. Within the variable lists, researchers can click on individual variables to display their details. The left navigation menu enables switching between levels and selection, e.g., of other studies at any time.

Figure 3: Screenshot variable list original order (EU-LFS)

The metadata at the variable level can be retrieved via several tabs (Basic Information, Questions, Description, and Values/Statistics) (see Figure 4).

Basic Information contains, for example, information about the group of respondents and whether the variable belongs to an ad-hoc module. *Question* contains information about the underlying question texts, and *Description* offers further explanations and notes on the variable, as well as links to relevant official definitions.

The information under the *Values/Statistics* tab is used to research information about the values and missing categories of the variables and unweighted frequencies. With respect to EU data, the frequency counts of the variables in total and for each country are available.

The availability of variables over time also is documented at the variable level (see the *Availability* section in Figure 4). The basis for this availability is the assignment of each variable to a unique the-

matic concept, which is used over all the survey years of a series. This feature enables researchers to identify in which years the variable is available and to compare variable details over time.

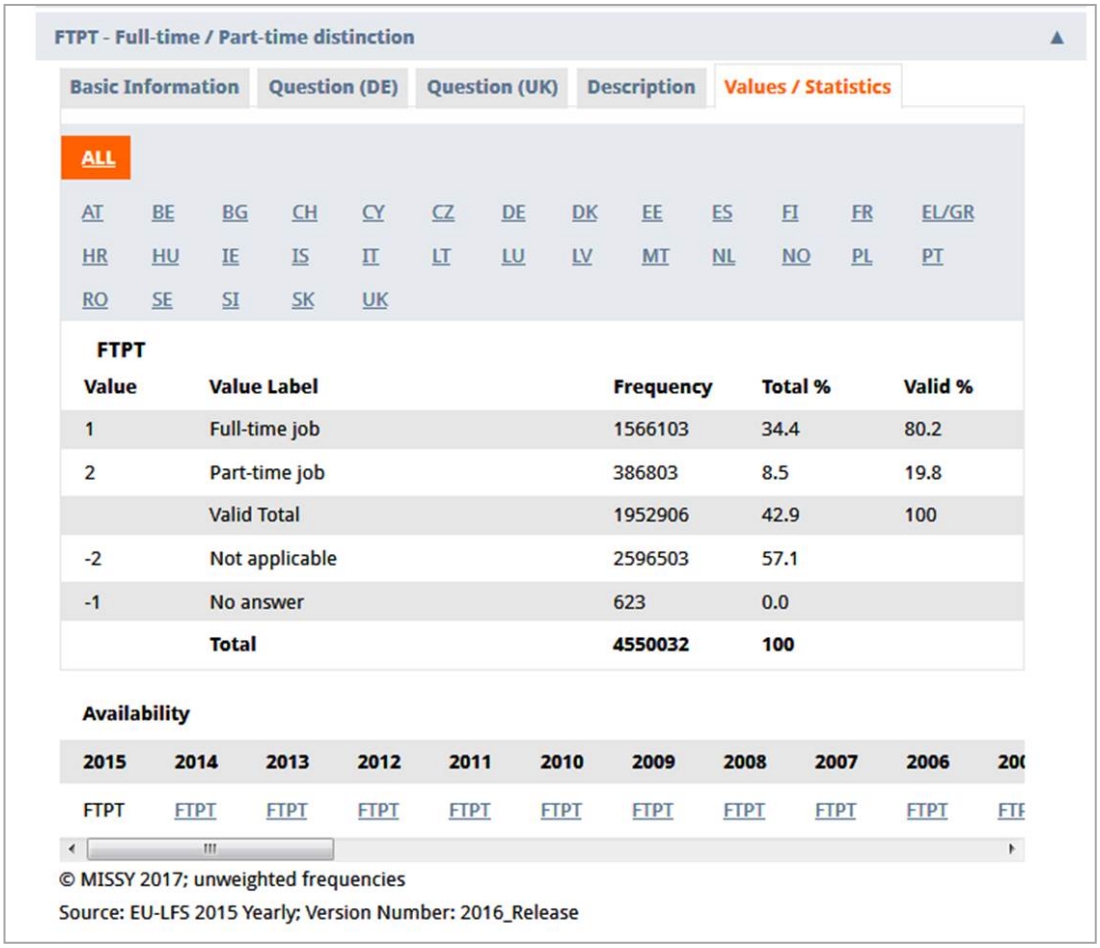
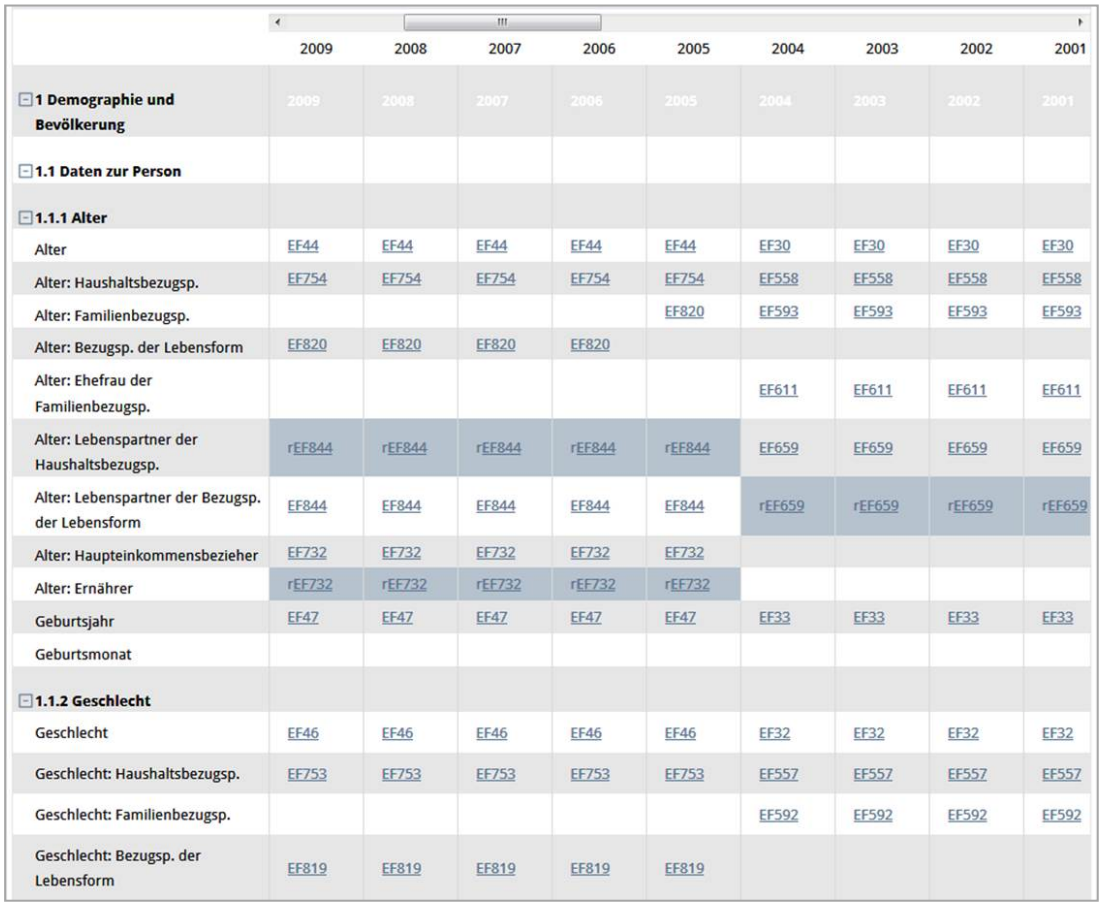


Figure 4: Screenshot variable information (EU-LFS, detail)

In addition, a variable-time correspondence matrix is available for each series documented in MISSY. This matrix provides a thematically structured overview of all variables and studies and links the variables over time. The overview of the variables over time can be used to find out whether a variable is available for all studies of interest.

With respect to the Microcensus in particular, a number of breaks have occurred in variables over time, e.g., variables were substantially changed or were no longer collected. Thus, the matrix for the Microcensus also contains information about whether a new comparable variable exists if a variable is not recorded further. The new variable may be used to make comparisons (blue fields in Figure 5).

For the documented EU data, additional overviews are available that show the availability of datasets for countries over time.



	2009	2008	2007	2006	2005	2004	2003	2002	2001
1 Demographie und Bevölkerung	2009	2008	2007	2006	2005	2004	2003	2002	2001
1.1 Daten zur Person									
1.1.1 Alter									
Alter	EF44	EF44	EF44	EF44	EF44	EF30	EF30	EF30	EF30
Alter: Haushaltsbezugsp.	EF754	EF754	EF754	EF754	EF754	EF558	EF558	EF558	EF558
Alter: Familienbezugsp.					EF820	EF593	EF593	EF593	EF593
Alter: Bezugsp. der Lebensform	EF820	EF820	EF820	EF820					
Alter: Ehefrau der Familienbezugsp.						EF611	EF611	EF611	EF611
Alter: Lebenspartner der Haushaltsbezugsp.	rEF844	rEF844	rEF844	rEF844	rEF844	EF659	EF659	EF659	EF659
Alter: Lebenspartner der Bezugsp. der Lebensform	EF844	EF844	EF844	EF844	EF844	rEF659	rEF659	rEF659	rEF659
Alter: Haupteinkommensbezieher	EF732	EF732	EF732	EF732	EF732				
Alter: Ernährer	rEF732	rEF732	rEF732	rEF732	rEF732				
Geburtsjahr	EF47	EF47	EF47	EF47	EF47	EF33	EF33	EF33	EF33
Geburtsmonat									
1.1.2 Geschlecht									
Geschlecht	EF46	EF46	EF46	EF46	EF46	EF32	EF32	EF32	EF32
Geschlecht: Haushaltsbezugsp.	EF753	EF753	EF753	EF753	EF753	EF557	EF557	EF557	EF557
Geschlecht: Familienbezugsp.						EF592	EF592	EF592	EF592
Geschlecht: Bezugsp. der Lebensform	EF819	EF819	EF819	EF819	EF819				

Figure 5: Screenshot Matrix Microcensus (detail)

In addition to the metadata provided in a structured form, MISSY provides syntax routines for data preparation and analysis, as well as various kinds of background information. For example, MISSY provides the setup routines prepared by the GML for importing and labelling raw data into statistical programs as download files. Also, a number of microdata tools are available for the operationalization of social science concepts and official constructs. These tools are produced mainly by GML, and in some cases, also by external researchers. Researchers can download the routines for various statistics packages and apply them to the data. The use of these routines enables researchers to process the data correctly without extensive effort so that their data analysis can be started quickly.

In addition, original documentation such as national survey instruments or the Eurostat user guidelines, are made available or linked as PDFs. Furthermore, information on official classifications, definitions of official statistics, and related literature is provided.

3 Technical implementation

3.1 Summarizing description of the implementation

The extent and complexity of the metadata to be documented for the different series required the use of an appropriate data model, for example, to map country-specific information at the study level. The DDI-RDF Discovery Vocabulary (Disco) standard provided a good fundament, which was extended at various points.

The technical infrastructure of MISSY includes different components that run on the same server (see Figure 6): a relational database, a web server, the MISSY editor, and MISSY web.

The central component of the architecture is the MISSY database, which can be filled with information at the variable level via various import interfaces. In addition to the automated imports, the MISSY editor can be used for entering and maintaining information, and for managing many administrative tasks. The database contents are represented in the MISSY web and the PDF codebooks. Exports to various DDI formats also are possible.

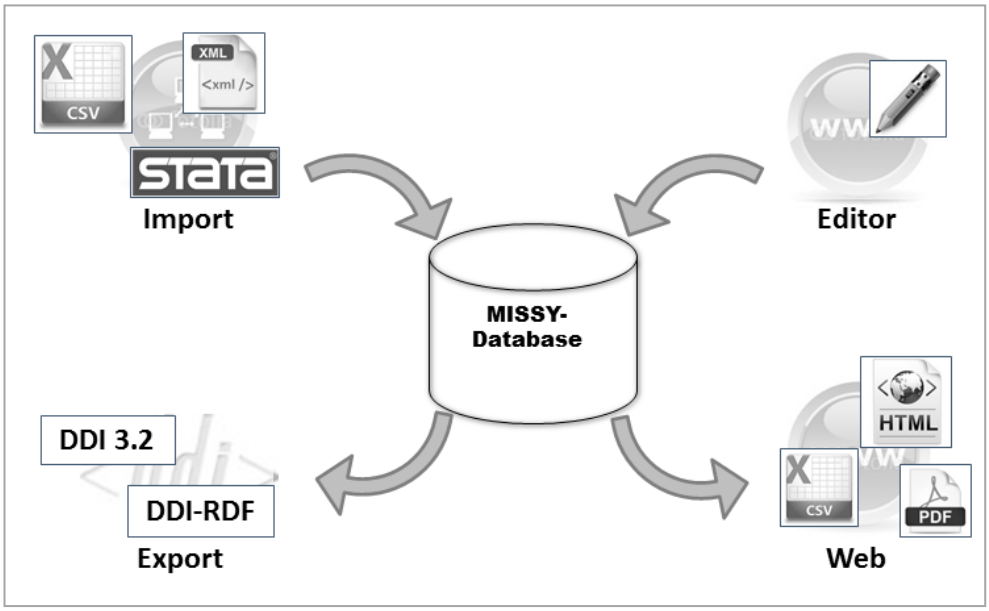


Figure 6: Technical infrastructure of MISSY

MISSY is mostly web-based. The MISSY-Editor, as the central tool for administration, can be used online and enables working on the database without local software installations. Therefore the access is flexible and location-independent. This facilitates work across locations, even in international teams.

3.2 Architecture

With respect to the technical implementation of MISSY, various best-practice approaches and software development patterns are used, e.g., the documentation via comments directly in the code and the publication of essential parts (such as the implementation of the data model) for re-use under an Open Access license.[11] In addition, a modular layer architecture was implemented to separate the different components into logically separate parts. The exchange of information between individual layers is standardized so that each of the components can be replaced fairly easily by new versions or other technologies. An overview of the technologies used can be found in Figure 7.

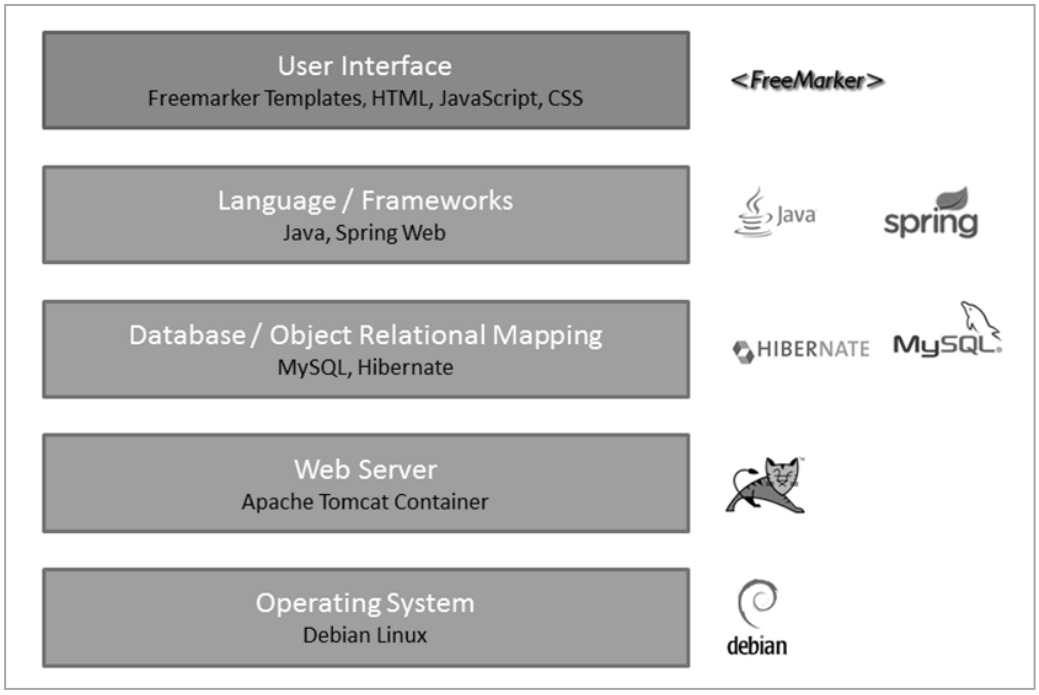


Figure 7: Technology stack

3.2.1 Data model

The MISSY data model is the logical conversion of the metadata schema described in Section 2. It is the core of the entire system and is used by all software components. It is based on DDI-RDF[12], a specification that represents the most important levels of the documentation of research and survey data (study, variable, question, questionnaire, statistics, etc.). DDI-RDF (Resource Description Framework) is a slim model for the structured representation of metadata in a format that also is machine-readable. Figure 8 shows the main entity types and dependencies of this model. The diagram indicates how individual entities are linked to each other and the cardinality of the connections.

[11] For information on the implementation of the data model, see https://github.com/missy-project/disco-model-impl; for information on the documentation of the code, see http://ddi.git.gesis.org/disco-model/.

[12] See https://www.ddialliance.org/Specification/RDF.

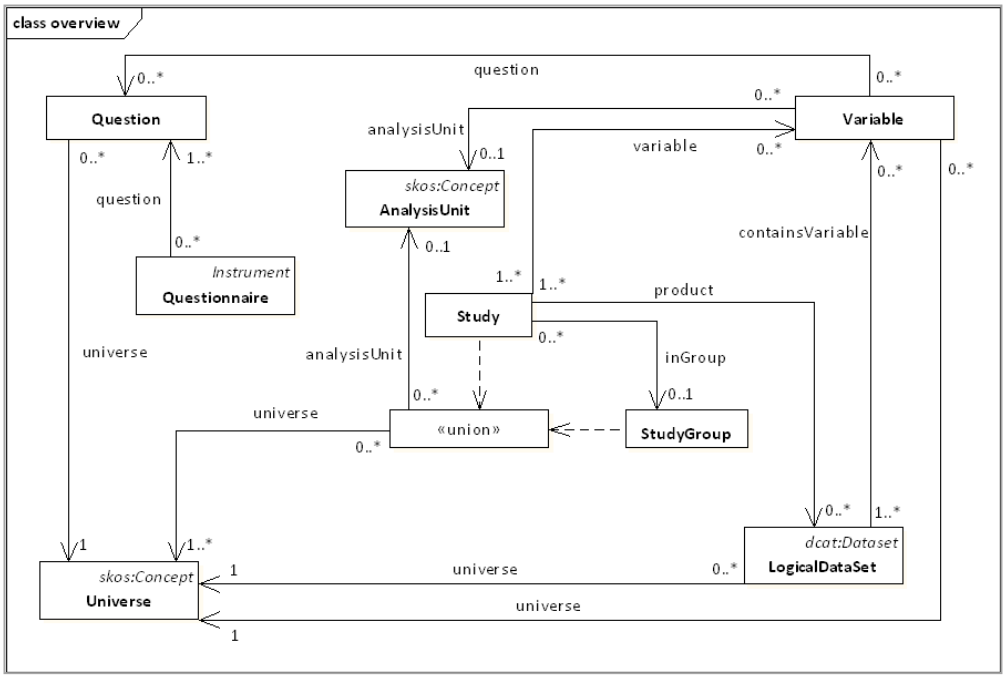


Figure 8: Section of the MISSY data model

The DDI-RDF model was developed in parallel to the MISSY infrastructure as part of the MISSY project. Through the collaboration of GESIS employees on the DDI-RDF vocabularies, the requirements of the MISSY project were directly incorporated into the development and thus provided important input for the specification (Bosch et al. 2015). However, since the documentation requirements of MISSY were too extensive and could not and should not be entirely covered by the standard, the model was extended project-specifically[13].

All the software components of the MISSY system are based on the data model, and therefore speak "the same language," which facilitates orientation when working with the underlying resources and makes solutions partially reusable.

As part of the MISSY project, the first modular and extensible implementation of this model was realized in Java, together with an implementation for saving data in a relational database. It can be used "out-of-the-box" – i.e., without fundamental adaptations – for other projects and is published on GitHub under an open-source license.

3.2.2 Database

The implementation of the data model in the MISSY database is strongly normalized, i.e., elements are reused, and not duplicated, wherever possible. This strong normalization enables efficient storage (with regard to the required storage space). In addition, the implementation represents the data model almost one-to-one. The different entity types are found in the tables and their attributes in the corresponding columns. The native implementation enables an intuitive access to the database. The identification of the database elements themselves is handled via technical identifiers (UUID). All objects can be uniquely identified via these identifiers. Also, linked objects can be found easily.

[13] See https://github.com/missy-project/missy-model.

Table 1 shows some statistics on the elements contained in MISSY (retrieved May 2018).

Table1: Size of the database

Element Name	Number
DB-tables	157
Series	6
Studies	82
Questionnaires	140
Questions	9,063
Variables	35,912
Descriptive Statistics	523,990
Categories	2,567,013

3.3 Import interfaces

An interface for the statistics software Stata was developed for importing variable lists and categories, frequencies, and descriptive statistics into the MISSY database. This interface uses the Stata system files of a dataset and aggregates them into a metadata file that contains the values mentioned above. With respect to the documentation of international surveys, values are created for each country included in the data and for all countries in total. The metadata is completed with the identifiers (UUID) of the elements already present in the database. In a second step, SQL-statements are generated from this file, which are used to import the values into the database. Access to the microdata files is linked to special access and data protection conditions, which require an adequate IT infrastructure. Therefore, this part of the application is not available online; rather, it is operated at GESIS on an appropriately protected server.

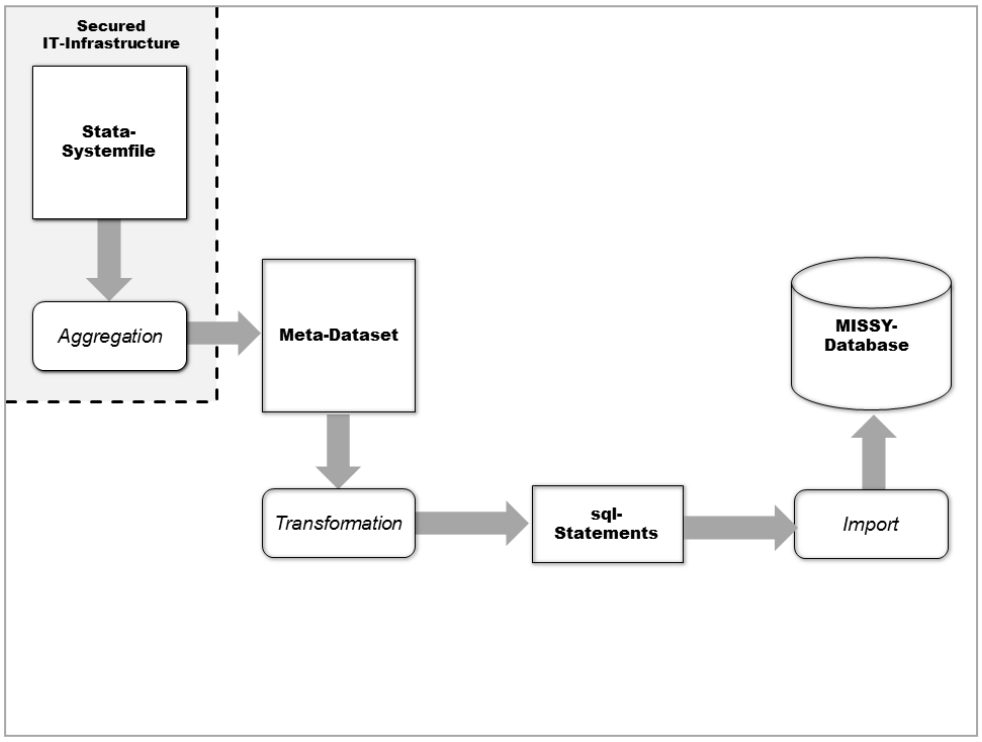


Figure 9: Scheme of the MISSY import process

Additional metadata at the variable level that is systematically documented in MISSY – such as a thematic classification or comments – also can be imported in bulk. This procedure also applies to questionnaires, whereby questions can be associated directly with their corresponding variables. This metadata is prepared locally in Excel- or CSV-files and can be imported into the database using various Java classes.

3.4 MISSY editor

The web-based MISSY editor provides a graphical user interface for entering and modifying metadata. In addition, a number of functionalities for administrative tasks can be found. The MISSY editor is based on the programming language Java and is implemented in a Spring Framework.

In the data documentation section, various input forms are available for viewing and entering metadata at study, country, variable, and question level. An editor user navigates to the required position by selecting the series, study, and dataset type to find the required data in a systematically organized form.

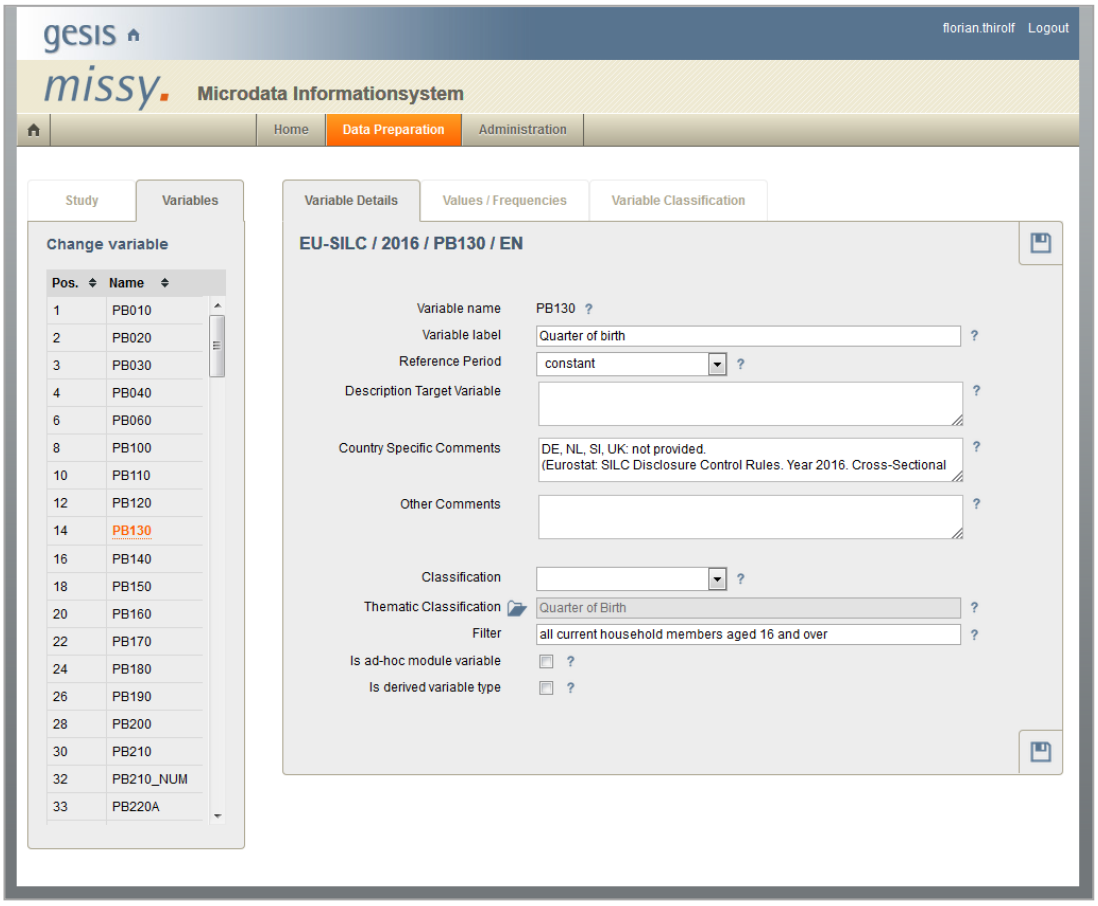


Figure 10: Variable details in the MISSY editor

The sub-navigation on the left can be used to switch between the data points (variables and studies). Depending on the type, documented information is collected as free text, single or multiple selection, or controlled vocabulary. Furthermore, an upload function is available that can be used to save questionnaires, quality reports, and other materials in a repository.

In the second section of the MISSY editor, various administrative functions are available, which are required during the documentation process. This series management enables the addition of new studies and questionnaires, and a modification of the thematic classification of variables and controlled vocabularies. A subset of metadata elements can be selected from the entire metadata schema to match the specific documentation needs of a series.

Users of the MISSY editor can be created via user management and can be assigned to predefined user profiles. In addition, access to individual studies and administrative functions can be adjusted individually.

3.5 Representation and export

The generation of web pages and PDF documents also is started by way of an administrative function. Graphical user interfaces are available for both output formats, which can be used to initiate output generation.

To create or update web pages, the editor user first selects the series and then defines the area of the website that is to be recreated, which can be restricted along the levels of the metadata schema and over time. The matrix views presented in Section 2 can be updated in the same way. The database is accessed using freemarker templates, which generate static html files that are stored server-side and therefore are retrievable for the users of the MISSY web without delay.

The process of creating PDFs is analogous. In this case, however, an XML-file containing the relevant information for the selected study is derived from the database, written to an XML-file, and transformed into PDF files using XSL-FO templates. This process can be used to generate variable lists in a thematic or original order (according to the order in the data set) and for frequency counts.

To enable other systems to use the metadata documented in MISSY, export interfaces in standardized formats were conceptually considered from the beginning of MISSY development. These export options are based on commonly implemented standards: for example, export to DDI-RDF is possible without additional mapping, since this mapping is implemented natively in the database. This makes it possible to merge the exported metadata with other standardized metadata. In the course of the development of MISSY, a mapping of DDI-RDF to DDI 3.2 was created, which is available to interested partners on request.

4 MISSY with respect to the FAIR criteria

The goal of MISSY is to make official statistical data easier to use and to facilitate scientific research with these data. A similar goal – to make data (and research objects in general) as easy as possible to re-use – is the objective of the working group composed of representatives of academia, industry, funding agencies, and scholarly publishers that has worked out guidelines to maximize the reusability of data (Wilkinson et al., 2016; see also Mons et al., 2017). The guidelines developed by this group request that data and its associated metadata should be FAIR – findable (F), accessible (A), interoperable (I), and reusable (R) – "both for machines and for people." These guidelines have received much attention and have been taken up by the EU Commission, among others. Broad agreement exists for accepting the FAIR principles, which have been endorsed by more than 100 organizations, including numerous data archives, universities, and GESIS.[14]

The FAIR guidelines are aimed primarily at data producers and archives. MISSY is a special case because GESIS neither produces nor distributes the official data, but exclusively provides the metadata and tools for using the data (see Section 2). However, the FAIR guidelines are designed deliberately to be modular and so flexible that they also can be applied to specific cases. It is explicitly stated that metadata can be FAIR even if the data on which it is based is so sensitive that it is not openly accessible (Wilkinson et al. 2016: 4), which means that the FAIR guidelines also can be applied to metadata and it can be checked to what extent MISSY complies with the FAIR guidelines.

The FAIR guidelines consist of four basic principles – findability (F), accessibility (A), interoperability (I), and reusability (R) – each of which is substantiated by several numbered sub-items. The following section explains the meaning of these principles and their associated sub-items, and to what extent MISSY fulfils them.

4.1 Findability

In general, data only have added value for researchers who are interested in re-use if the data can be found. Thus, the first principle (findability) requires that "Data and metadata should be easy to find by both, humans and computer systems" (SNF, 2017: 1). This principle is substantiated by the corresponding sub-items. This principle is fulfilled if the data and the metadata are searchable, and therefore, can be found (F4), are uniquely identifiable, and the data and metadata can be assigned to each other (F1, F3), and the data is well described (F2). Only when these aspects are fulfilled can the sought-after data be found.

The two aspects of the unambiguous identifiability of the data and the unambiguous assignment of the data to the metadata require the following: "(meta) data are assigned a globally unique and persistent identifier" (F1), and "metadata clearly and explicitly include the identifier of the data it describes (F3)" (Wilkinson et al., 2016: 4).

MISSY reaches its limits in meeting these criteria. The allocation of a "globally unique and persistent identifier (PID)" is a responsibility of data producers, and thus, of the statistical agencies. Such PIDs have not yet been assigned, and therefore, cannot be reported in MISSY. A clear identification of the data and a clear assignment to the metadata is nevertheless provided. With respect to the Microcensus, only one data release is available. Additionally, from the Microcensus 2005 onwards, the information about the data version is stored in a version variable that is documented in MISSY. EU data are

[14] See https://www.force11.org/datacitation/endorsements.

unambiguously identifiable by study name, year, and version number, which is information published with the data set and also specified in MISSY (for each variable). This approach ensures that the metadata is clearly assigned to the data.

MISSY also fulfills the searchability criteria: "(meta)data are registered or indexed in a searchable resource (F4)" (Wilkinson et al., 2016: 4). The metadata documented in MISSY can be searched via search engines and an internal search on gesis.org. Machine processing and searchability is provided by the use of semantic technologies (RDF) used in the "Web of Linked Data." RDF technologies are explicitly mentioned as a solution to meet the FAIR criteria (Mons et al., 2017: 51).

The last aspect of findability requires that data must be sufficiently described to make it clear to users (human and machines) what the data is. This criteria (F2: "Data are described with rich metadata") (Wilkinson et al., 2016: 4) is further defined in Section 4.4 "Reusability" (R1).

4.2 Accessibility

Apart from the fact that the data must be found in the first place, it also must be accessible to enable reuse, i.e., they "should be stored for the long term such that they can be easily accessed and downloaded or locally used by machines and humans using standard communication protocols" (SNF, 2017: 2). Two criteria of this accessibility requirement are easy and permanent access.

Easy access is described by the following criterion: "(meta)data are retrievable by their identifier using a standardized communications protocol (A1)" (Wilkinson et al., 2016: 4). MISSY fulfils this criterion of accessibility, since the metadata in MISSY can be found via search engines, and MISSY is linked via the homepage of the Federal Statistical Office and Eurostat (i.e., on the web pages of the data providers). The metadata is available online, and therefore, freely accessible on the MISSY web.

Permanent access is addressed by the following criterion: "metadata are accessible, even when the data are no longer available (A2)" (Wilkinson et al., 2016: 4). This aspect of the infinite availability of metadata via the MISSY websites cannot be guaranteed. However, it is possible to save the metadata from the underlying database or to export it to different formats for further use in other systems.

4.3 Interoperability

In addition to discoverability and accessibility, data must be *interoperable* so to be optimally reused, which means they should be compatible and linkable with other data. Thus, the criterion requires that data and metadata must be "exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems" (SNSF, 2017: 3). To make this possible, first, the data should be in a form that corresponds to a standard, and second, a link to other data (e.g., in the metadata) should be reported.

Standardization is specified by two requirements: that "(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation (I1)" and that "(meta)data use vocabularies that follow FAIR principles (I2)" (Wilkinson et al., 2016: 4).

MISSY fulfils both aspects of standardization. The MISSY data model is based on DDI-RDF, i.e., on the DDI Discovery Vocabulary (Disco) ontology, which includes the most important components of the DDI standard in RDF format that enables the metadata to be combined with other metadata that complies with this standard. Another major advantage of the data model is that metadata can be exported in various formats (DDI-RDF, DDI-XML, etc.) to enable the exchange of metadata with other documenta-

tion systems. These DDI files can be provided on request. The publication of metadata in XML format or the possibility of harvesting are currently not offered but are possible upon request.

The third point specifies the requirement for *linkability* by referring to other data records – it is required that "(meta)data include qualified references to other (meta)data (I3)" (Wilkinson et al., 2016: 4). MISSY only partially fulfils this requirement, since it includes the metadata of the official microdata without referring to other datasets. However, the individual datasets documented in MISSY are linked at the variable level within a study (see Section 2). This linking of variables over time is made possible by the *thematic classification*, which is available in machine-readable form.

4.4 Reusability

To be reused, data must be described in such a way that it can be understood, and it must be clear how it can be used, i.e.:

"Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans" (SNF, 2017: 4).

The first aspect of reusability, a sufficiently good description of the data, also facilitates the findability of the data and is, therefore, a concretization of criterion F2. The use of a standard is demanded again, which also overlaps with I1 and I2.

Three points are required for a good description of the data: that "meta(data) are richly described with a plurality of accurate and relevant attributes (R1)", "(meta)data are associated with detailed provenance (R1.2)" and "(meta)data meet domain-relevant community standards (R1.3)" (Wilkinson et al., 2016: 4).

As far as possible, the requirements of the rich description are fulfilled by MISSY, since it offers detailed metadata on series, study, and variable levels (see Section 2). MISSY also documents the information on how the data was generated, how it was processed, and how it can be reused.

MISSY uses a standardized metadata schema and a shared data model for all data. This data model is DDI-based and thus follows the standard for social science metadata (see Section 3). Thus, MISSY not only follows a consistent data model based on a standard but both the underlying DDI standard and the data model based on it have been published (see Section 3).

The second aspect of reusability, the clearly defined use of data, is described by criterion R1.1, which demands, that "(meta)data are released with a clear and accessible data usage license (R1.1)" (Wilkinson et al., 2016: 4).

The organization of access to data for scientific purposes is the responsibility of the agencies. MISSY documents how the microdata can be obtained. With regard to the metadata offered in MISSY, the content of MISSY is licensed under a Creative Commons license, and this information is documented on the MISSY web, which means that this criterion of clearly defined data access is met as far as possible.

4.5 Conclusion: Is MISSY FAIR?

MISSY itself is FAIR and helps to make official data more easily accessible and more usable.

The metadata in MISSY can be found via search portals that improve the findability of official microdata for research through extensive documentation and linking to the data provider (FAIR). The

metadata is freely accessible on the MISSY web (F**A**IR), and the MISSY data model is based on DDI-RDF, so export formats for exchanging metadata can be developed (FA**I**R). Furthermore, MISSY provides comprehensive metadata at the study and variable level, thus improving the usability of official microdata. In addition, both the underlying standard and the metadata schema are published and thus are both comprehensible and reusable (FAI**R**). Therefore, MISSY contributes to making official data FAIR.

References

Bosch, Thomas, Olof Olsson, Benjamin Zapilko, Arofan Gregory & Joachim Wackerow (2015): DDI-RDF Discovery - A Discovery Model for Microdata. IASSIST Quarterly, 38(4) & 39(1), 17-24. [http://iassistdata.org/sites/default/files/iqvol38_4_39_1_bosch.pdf]

Data without Boundaries (2015a): Deliverable D5.5: Final report & recommendations for the continuation of services for European OS Microdata. Work Package 5: Servicing European Researchers in the use of OS Microdata. [http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d5-5_eu-os-microdata-services-continuation_final-recommendations-report.pdf]

Data without Boundaries (2015b): Deliverable D5.4: Report and Databank Documenting Integrated EU OS Data. [http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d5-4_databank-documenting-integrated-eu-data_report.pdf]

Dragon, Iris & Mariya Zvezdanova (2018): Forschungsdatenzentrum „German Microdata Lab": Service für amtliche Mikrodaten Jahresbericht 2017. GESIS Papers 2018/07. [https://www.ssoar.info/ssoar/handle/document/56854?locale-attribute=en]

Mons, Barend, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz Olavo Bonino da Silva Santos & Mark D. Wilkinson (2017): Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use 37: 49–56 49, doi:10.3233/ISU-170824

Swiss National Science Foundation (2017): Explanation of the FAIR data principles www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

Wilkinson et al. (2016): Comment: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, doi:10.1038/sdata.2016.18

Appendix: MISSY metadata schema

Series Level

- Title
- Subtitle
- Release Year
- Version Number
- Data Publisher
- Abstract
- Selected Keywords
- Geographical Coverage
- Organization
- Universe
- Sampling
- Data Collection
- Anonymization
- Legal Basis
- Weighting
- Data Access: Conditions
- Data Access: Form
- Data Access: Contact
- Data Service
- Comparability Over Time
- Comparability Over Space
- Notes
- References

Study Level: General Description

- Title
- Subtitle
- Abstract
- Keywords for additional modules
- Geographical coverage
- Time period covered
- Notes
- References
- Additional modules
- Module: Sampling fraction
- Module: Type

Study Level: Country Specific Information

- Corresponding national Study
- Producer
- Universe
- Cross-sectional - Sampling units
- Cross-sectional - Actual sample size
- Cross-sectional - Achieved sample size
- Longitudinal - Sampling units
- Longitudinal - Actual sample size
- Longitudinal - Achieved sample size
- Ad-hoc module - Sampling units

- Ad-hoc module - Actual sample size
- Ad-hoc module - Achieved sample size
- Available data
- Source of sampling frame
- Sampling design
- Primary sampling units
- Secondary sampling units
- Stratification criteria
- Sampling method
- Number of rotational groups
- Units of observation
- Units of analysis
- Dates of collection: Start date
- Dates of collection: End date
- Participation mandatory
- Type of data source
- Interview mode
- Percentage of proxy interviews
- Design weight: Target
- Design weight: Method
- Non-response weight: Method
- Coverage adjustment weight: Method
- Weighting: Method
- Final weighting: Method
- Notes
- References
- Countries
- Changes to questionnaire
- Methodological advice
- New questions/variables
- New classifications
- New concepts
- Changes to typification
- Anonymization
- Year specific documentation

Study Level: Data Set

- Title
- Weighting
- Notes
- References

Study Level: National Questionnaire

- Question number
- Question text
- Question annotation
- Filter instruction
- Filter (literal)
- Filter (formal)
- Answer categories
- Participation mandatory
- Number in interview manual
- Number in laptop interview manual
- Assignment to variables
- Note

Variable Level: Basic Information

- Variable name
- Variable label
- Classification
- Reference period
- Thematic classification
- Unit of observation
- Filter
- Is ad-hoc module variable
- Sequential data record (start)
- Sequential data record (end)
- Digits of decimal point
- Subsample
- Program
- Sampling fraction
- Thematic comparability

Variable Level: Description

- Description target variable
- Country specific comments
- Comment: Change in question
- Comment: Change in category
- Comment: Methodological note
- Other comments

Variable Level: Values/Frequencies

- Statistics
- Value
- Value Labels
- Frequencies
- Total percent
- Valid percent

Variable Level: Classification

- Assignment to question
- Is generated variable?
- Generated variable: Type
- Generated variable: Concept
- Generated variable: Unit
- Generated variable: Recording details