gesis Leibniz-Institut für Sozialwissenschaften

GESIS *Papers* 2016|06

The extended metadata schema of the VRE soeb3

Uwe Jensen & Stefan Schweers

GESIS Papers 2016|06

The extended metadata schema of the VRE soeb3

Uwe Jensen & Stefan Schweers

GESIS – Leibniz-Institut für Sozialwissenschaften 2016

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Abteilung „Datenarchiv für Sozialwissenschaften“
Unter Sachsenhausen 6-8
50667 Köln
Telefon: 0221 / 47694 -430
Telefax: 0221 / 47694 -8430
E-Mail: uwe.jensen@gesis.org

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

Acknowledgement

This text is a translation of the original German working paper "Das erweiterte Metadatenschema der VFU soeb 3" (Jensen and Schweers. 2014).

We want to thank Dr. Kristi Winters for proof-reading the English text version and her valuable comments and improvements.

Table of contents

1 Introduction

The aim of the "Virtual Research Environment" (VRE) [1] project was to develop a work platform that supports collaboration of distributed scientific institutions in collaborative projects. Virtual research environments can only be developed if they are closely related to research and projects or are community specific.

For this purpose, infrastructure development and professional scientific research were linked in terms of the joint research project soeb3 (Berichterstattung zur sozioökonomischen Entwicklung Deutschlands / Reporting on Socio-Economic development of Germany) funded by the Federal Ministry of Education and Research (BMBF; 2013 to 2016). This project serves as use-case for the project-based, research-oriented development of the Virtual Research Environment (VRE).

The VRE project was based on conceptual work from the sub-project "Collaborative Data analysis and virtual working environment" (VirtAug) during the concept phase[2] to soeb. 3. During the concept development, the pilot project should examine and document the practical case of socio-economic reporting, and

> "how the shared data access and data related cooperation between social scientists from various quantitative-empirically oriented research institutions and in particular a collaborative evaluation of microdata of research data centres could be better organized and technically supported in the future." (cf. Bartelheimer, Schmidt. 2011) [3]

The implementation phase, funded by the German Federal Ministry of Education and Research (BMBF), began in June 2012 and lasted until the end of June 2014. The resulting portal "VRE soeb 3" and the modular implemented tools support the project operation. In particular, the collaborative use of social scientific microdata across the scientific performance chain; it extends across data access to archiving research results. In addition, further components facilitate VRE communication and coordination in the research association soeb3.

Under a common user interface, the portal integrates the developed metadata editors (see below) and other tools such as forum, wiki, file management, calendar, announcement and consultation component, and - by means of the middleware - the storage and searching of metadata (archive function).

The data management focuses on supporting the archiving, documentation and re-use and collaborative use of syntaxes for evaluation (Syntax-sharing). Besides syntax files and (shared) research data all other file types, such as output files, spreadsheet or text documents are saved regardless of format and are made discoverable by search and content via metadata.

It should be possible, under the existing usage permissions, to create logical links between different types of data (e.g. between syntax and used research data), in order to understand the research process coherently and transparently. A crucial prerequisite for the use of a common working environment is that the data are well described using metadata (see. SOFI. 2013).

Beyond the research network on Socio-Economic Reporting the project's aim is also to find solutions of general importance for quantitative, empirically-based social science research. On 24.01.2014 the project group presented "Virtual Research Environment for Socio-Economic Reporting" (VRE soeb 3) in

---

[1] http://www.soeb.de/vfu-soeb-3/

[2] http://www.soeb.de/vfu-soeb-3/konzeptphase/

[3] All resources and publications of the project "VFU soeb 3" were published in German language.

Berlin, the first operational version used in the VRE a broader professional public and put it in the context of five panel discussions[4].

Implemented for productive operation, the "Virtual Research Environment for Socio-Economic Reporting" portal is accessible for the research network's registered members[5]. An idea of how the different areas and functions of the VRE are associated is described by Schmidt (2014). The user oriented introduction to using the VRE provides the component's main structure and relationships including exemplary processing sequences.

The virtual research environment was developed by an interdisciplinary project group that consists of the following partners:

- Sociological Research Institute Göttingen (SOFI) eV at Georg-August University Göttingen (project leadership)
- GESIS - Leibniz Institute for the Social Sciences eV , Mannheim
- Society for Scientific Data Processing Göttingen mbH (GWDG), Göttingen
- Research Data Centre of the Federal Employment Agency at the Institute for Employment Research (IAB), Nuremberg
- Göttingen State and University Library (SUB) of the Georg – August University Göttingen

in cooperation with

- D-Grid-gGmbH, Dortmund
- Research Data Centre of the Federal Pension Insurance , Berlin and Würzburg
- Research Data Center of the Socio-Economic Panel , German Institute for Economic Research (DIW), Berlin

GESIS defined the required metadata for the VRE and specified them in related metadata schemas that are described in this report. Based upon the specialized scientific specifications as defined with the project partners, further collaborative IT tools were developed and implemented; in particular three metadata editors by GESIS. The editors and the implemented metadata elements have been documented in the report "Die Metadateneditoren der VFU soeb 3" (Jensen, Schweers, and Carevic. 2014).

The present report describes the metadata elements of VRE soeb 3's extended metadata sets. To document the relevant data analysis, a metadata schema has been developed for each of the objects: Study, Data Usage and Syntax. Files that are stored in the VRE soeb 3 archive can also be assigned to these main objects.

Chapter 2 provides an overview of the various schemas and describes how the metadata elements in the schema tables are specified. Further, the required, recommended and optional elements and automatic generated metadata are described across the schemas.

Chapters 3 to Chapter 5 specify all the elements of the three metadata schema (Study Description, Data Usage and Syntax) and their associated sub-schema needed to document the related files stored in the VRE archive.

[4] http://www.soeb.de/vfu-soeb-3/abschlusstagung/

[5] https://vfu.sofi.gwdg.de/

2 The extended metadata schema of VRE soeb 3 at a glance

VRE soeb 3's extended metadata schema consists of three subschemas: Study, Data Usage and Syntax. These schemas describe objects and their associated files, which are the subject of conventional workflows when conducting data analyses. In the soeb 3 research network these analyses are carried out in eighteen scientific work packages of more than thirty-five scientists from seventeen institutions. Thus the documentation and development of analysis-based metadata in the VRE soeb 3 supports overall the collaborative use of social and economic research microdata. The implemented tools and metadata elements of the production system VRE soeb 3 are documented in the report "Die Metadateneditoren der VFU soeb 3" (Jensen, Schweers, and Carevic. 2014).

Below we present the metadata elements of VRE soeb 3's extended metadata schema. Using this metadata, researchers can document things systematically: e.g. initial data used in data analyses, objectives of the investigation, methodological strategies and analysis parameters of the data use, the statistical methods applied, the concrete syntaxes needed and their development. At the same time they promote informational networking between the scientific work packages; allowing, for example, searching the VRE archive for all data sets documented by the work packages. In particular, the value-added services of the archive (cf. ibid. chapter 8) allows - based on the captured metadata - common development, sharing and re-use of analysis syntaxes. To ensure good scientific practice, VRE soeb 3 metadata document relevant information of the research process from the initial data up to the results of the data analysis. The information captured with the metadata is verifiable, re-usable and permanently secured in VRE archive. Thus they are also available for upcoming infrastructure purposes after the end of the project.

The metadata elements are organized into three schemas defining a list of structured information. Each describes the following objects and their associated files.

The metadata schema **Study Description (SD)** describes the initial data and documentation of the data sets (of a study) that are used for data analysis in soeb 3 network project. The data provider's documents and / or data files can be associated with the study description and stored in the archive (as far as reasonably and legally-permissible). Files uploaded to the VRE archive store are described by the subschema **SD-Upload**. The contents of the metadata elements are edited with the metadata editor Study (cf. ibid. chapter 3).

The metadata schema **Data Usage (DU)** contains those items that describe methodical concepts and parameter of the data analysis. Data sets, which were generated by soeb 3 work packages from initial data and stored in the VRE archive, are assigned to this schema. Such stored data files and related text documents about data use are described by the elements in the subschema **DU-Upload**. The contents are captured with the metadata editor Data Usage (cf. ibid. chapter 4).

The metadata schema **Syntax (SX)** describes syntax files and its associated output files in various stages of their development. This schema is used to assign syntax files and in particular output files belonging to syntax files that are stored in the VRE archives. Uploaded syntax files and output files stored to the archive are documented with the subschema **SX-Upload**. The metadata are processed with the editor Syntax (cf. ibid. chapter 5).

The metadata schemas **Study Description** and **Data Usage** were developed with reference to the elements of metadata standards from DDI 2 (DDI Alliance. 2014). By contrast, currently there is no systematic, coherent metadata structure for a detailed documentation of syntax files and output files as well as the project-based context of collaboratively conducted data analyses (cf. Friedhoff et al. 2013:. 12 ff.). For this reason the metadata schema "**Syntax**" was developed from scratch to cover the needs of the VRE soeb 3.

The implemented elements have been used since the end of 2013 in the production system of VRE soeb 3 (Jensen, Schweers, and Carevic. 2014). Conclusions on the practical use of metadata and which "changes actually take place in work processes through the technical possibilities can only be made visible by future user studies" (Schmidt. 2014B).

The documented schemas and elements in this report are incorporated into the ongoing development of the DDI Alliance (cf. Jensen. EDDI 2014). Chapters 3 to 5, which specify and further describe the conceptual aspects of the three metadata schemas: Study Description, Data Usage and Syntax.

Figure 1 shows the relationship between the three metadata schemas, their files, their recording and processing by respective editors and the reference to the VFU archive.

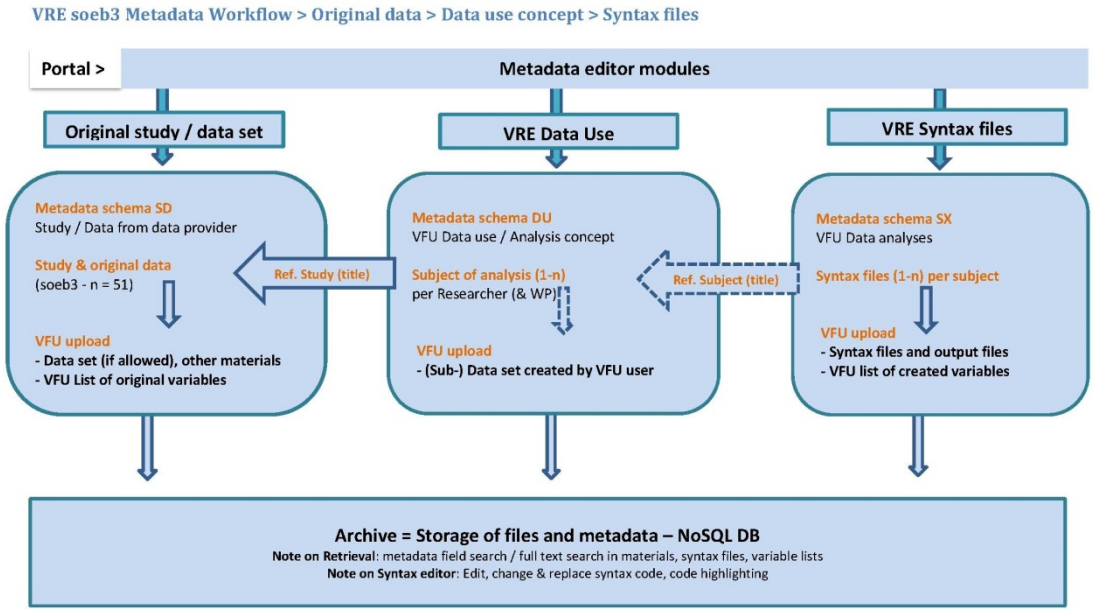

Figure 1: Overview of the editors and metadata schemas (Jensen 2014)

2.1 Representation of the metadata elements in the schema tables

The tables in Chapters 3 to 5 describe all elements of the respective metadata schemas. Additionally each schema of the objects Study Description, Data Usage and Syntax includes two more tables (sub-schema) with elements that document the file types assigned to the relevant main object while saving it at the archive.

Each table contains information corresponding to the entries in the table header, which are explained below.

No.	Element name	Definition and remarks (MD-SD)	FA

Figure 2: Column names of metadata schema tables

- **Column No.**

  It specifies the serial number of the adjacent element name.

  In case a line entry is not associated with any number, it is an introductory remark at a logical group.

  In case that a line entry is assigned with an integer number (e.g. 5.), it is usually an umbrella term without the need to capture any related content. The last column is empty in this case. Umbrella terms serve as context information that will be preceded by the following sub-elements (e.g. 5.1 / 5.2), which contain the captured content.

- **Column "Element name"**

  Each field holds the name of the metadata element. In some cases global names of previous elements are carried along with the name of the current element to retain the context.

- **Column "Definition and remarks (schema abbreviation)"**

  Abbreviations in brackets designate the current schema, e.g. MD-SD stands for Metadata Schema of the object Study Description.

  The content in this column has a dual function. First, the element is defined. Additional notes explain, where required, an element's substantive or formal peculiarities, show logical dependencies on other elements, refer to the repeatability of elements and provide examples.

  For practical reasons, this report will omit presenting the logical relationships between the elements (A/C; Child/Attribute) and the attribute "occurrence" (number of possible or mandatory instances of an element) separately in columns. Specifying 1 - n in the text means, in this context, that the element must occur once (1), and can occur several times (n). 0-n indicates that the element does not have to occur (0), but can occur more than once.

  Furthermore, controlled vocabularies, used as standardized list entries in the editors, are listed in this column. Unless vocabularies originate from other schemas, they are listed here as a source and indexed in the reference list.

- **Column „FA" (Functional Aspects)**

  In the last column the elements are connected in terms of functional aspects.

  This includes in particular the indication about mandatory (M) and recommended (R) or optional (without coding) entries.

  Concerning the required information, metadata are distinguished between those that must be captured by a person (M), and those that are automatically generated (A) by a technical system of the VRE archive (AA) or VRE portal (AP).

  For clarity the mandatory, the recommended and the optional elements are not shown separately in short form as a table. Instead, they are jointly explained in Section 2.2 in terms of their content and across all schemas.

  The corresponding information is also encoded in all schemas in the last column. All abbreviations used are shown in the next overview.

  Overview of shortcuts to encode functional specifications in column "FA"

  - AA Required information that is "A"utomatically generated by the "A"rchive system
  - AP Required information, which is "A"utomatically generated by the "P"ortal system
  - M Mandatory information that is created by a person
  - R Recommended information
  - FT Free text field
  - CL Calendar function
  - CV Controlled vocabulary
  - URI Uniform Resource Identifier. Unique identifier for identifying sources.
    The form of URL used in the schema refers to a source on the Web.

2.2 Mandatory, recommended and optional elements

The mandatory, recommended and optional elements as well as automatically generated metadata are described across all schemas in the following sections.

Mandatory elements (M)

When creating an object Study, Data Usage or Syntax the following two elements are mandatory:

- Title (of the object)

  The title refers to the given object with meaningful content. This is required for sensible indexing of objects and is used as a central metadata in the archive search. For this purpose conventions for assigning file names and titles were proposed in the VRE soeb 3 (Schmidt 2014a: 5).

- Work Package

  A VRE member can be a member of one or more scientific or administrative work packages. To assign the processing of an object to the related work context, the member selects the respective work package from a predefined list.

If files are associated with one of the objects Study, Data Usage or Syntax by storing them in the VRE archive, following information on these elements is required:

- Work package (see above)
- Reference Title < Study, Data Usage or Syntax >

  Referencing the respective object is done by selecting from a list of all the titles already assigned to one of the three main object types.
- Title (of the file) (see above)
- Filename (see above title of the object)

  Enter the full name of the file (file descriptor, file format) stored at the VRE archive. It can be taken when uploading to the VRE or can be edited before saving.

Automatically generated administrative metadata

Furthermore, some elements are automatically generated by the portal system or archive storage system, which are necessary to manage the object. At the same time, these metadata are also necessary because they should automatically appear as context information when editing an object in one of the editors (cf. Jensen, Schweers, Carevic. 2014: Chapter 6).

The automatically generated (technical and administrative) metadata include the following elements, which are used in all three schemas and their file-related sub-schemas.

Administrative information generated by the VRE portal (AP)

- First name, last name of the person who created or recently edited the object (and associated files).

  These metadata are taken from information provided by the registered person in the portal that uses a metadata editor to create or edit the object.
- Work Package

  Information about a work package can be derived through registered memberships of the respective person. If one has membership in several APs a selection list will be offered to assign the WP fitting to the work context.
- Placeholder for a VRE project title, e.g. "VFR soeb 3"

Administrative information generated by the archive (AA)

- Storage date (object Study Description, Data Usage or Syntax)
- System ID (ditto)
- Information in the context of the editor function "**Publish**" (ditto)

  Editors function with Yes / No switch. This feature is intended to avoid that incomplete or incomprehensible metadata is published, e.g. because the processing was interrupted.

  A corresponding flag indicates to the person, whether the current object state has previously been published in the VRE.
- Information in the context of the editor function "**Duplicate**" (ditto)

  An editor function for duplication of the complete metadata set of an object under a new title, e.g. to reuse (individual) content in the new object, while other content is to be changed.

The new object is linked to the originating object via the following metadata:

- Title of the original object Study Description, Data Usage or Syntax
- System-ID of the original object
- Ref object ID.

  Due to the reference the new object is navigable linked with the original object. By clicking the link all the metadata of the originating object are displayed.

- Information in the context of the editor function "**Permissions**" (ditto)
  - Date of approval

    Date on which the approval was granted

- Information in the context of linking a file to the object Study Description, Data Usage or Syntax
  - Object <SD, DU or SX> Title (selection list)

    Assigning a file to a title of a reference object Study Description (SD), Data Usage Description (DU), or Syntax file Description (SX), both stored in the VRE archive.
  - Ref Parent ID

    Reference, which associates the stored file to the system ID of the reference object.

    The referenced files will be displayed as "associated files" together the reference object in the respective processing context of the editors.

- Further information about the linked file:
  - File size
  - Storage date
  - Fingerprint
  - Date of approval

Recommended (R) and optional elements

This distinction between recommended and optional elements touches particularly the metadata schema Study Description. For transparent and replicable documentation of the initial data used at the study level, we recommended compiling certain individual elements or groups of individual elements. Depending on the needs, context and traceability of the data further elements can be documented, e.g., by referencing materials supplied by the data provider (data documentation, methods reports, literature on the data).

By contrast, it is recommended as a rule to report (step by step) all metadata elements of the objects (and the associated files) Data Usage and Syntax to support the indexing and collaborative use of these objects and their components in the context of data analyses by the scientific work packages.

3 The metadata schema Study Description

3.1 Introduction

The schema Study contains the metadata elements with which the initial data (so-called "Ausgangsdaten" cf. Dickman et al 2010) are documented. These data are supplied by various data provider (data archives, research data centers, statistical offices, central bank data, OECD, etc.), that are used for data analysis in the project VRE soeb. Fifty data sets and databases from more than thirty-five data providers are used in soeb 3 data analyses. In addition, updates and new data to the respective data sets are taken into account. The metadata for all data sets used in the eighteen scientific work packages shall be retrievable within the VRE search functionality.

To ensure good scientific practice, the analysis procedures and syntaxes as well as the underlying (output) data are transparent and clearly documented. In the interests of safeguarding the VRE soeb 3 results, data analyses (and information on initial data) should be documented for review and reuse and permanently secured.

The formation context of social science research data, the subject-specific content and methodological characteristics have for a long time been described in archive contexts with the so-called Study Description of the DDI documentation standard (DDI Alliance. 2014). Therefore, the metadata schema of the study description is used as a generic approach for documenting initial data of soeb 3. The DDI 2 compatible schema of the GESIS Data Catalogue (Zenk-Möltgen; Habbel. 2012) serves as a reference schema and was adjusted to the needs of the VRE soeb 3. Additionally, elements of the da|ra Metadata Schema (Hausstein et al. 2012, 2014) were included, e.g. for documentation of persistent identifiers.

When considering the research interests in a VRE and the challenges for sustainable curation of data-related context information, it has to be weighed how extensively and accurately the initial data should be documented. On the one hand, 'broken links' might be considered in access to initial data (where it is already subject to contractual restrictions) and / or to their documentation because URLs are no longer available to relevant websites. On the other hand, the documentation effort is considerably reduced when the data provider supplies initial data with a persistent identifier (PID) which can be accessed via the URL permanently.

3.2 The metadata elements of the schema Study Description

Mandatory elements (M in column FA) are 1.3 and 7.1 Work Package Title.

The recommended minimum information concerns elements (cf. FT / CV in column FA) of the groups that document the access, origin and context of the data set in the initial data used for 3 soeb analyses. This concerns, in particular, the elements of the following groups:

- No. 5 Data provider: 5.1 Name 5.2 Language version of the data, 5.3 URL;
- No. 6 Data access: 6.1 Terms of use (+ URL), 6.2 Access type;
- No. 7 Study: 7.3 Study ID (+ type), 7.4 Content (+ URL);
- No. 9 Data publication: 9.1 Date, 9.2 Version in conjunction with
- No. 10 Persistent Identifier: PID 10.1, 10.3 PID URL, 10.4 Data citation
- No. 11-14 Methodology: 11 Reference period, (12) time and (13) and dimension of the data, 14 Basis of data collection (including universe, sampling method, survey method);

- No. 16-18 Data set: 16 Structure, 17 Anonymisation, 18 Formal characteristics (including number of variables, unit of analysis);
- No. 19-22: Data File: Structure of the files (+ URL), file format, file name, fingerprint.

To minimize capture efforts, especially for complex data sets such as the Socio-Economic Panel (SOEP), it is recommended to reference, whenever possible, the corresponding information page of the data provider by the respective URL element, in addition to the related item description.

An available Persistent Identifier (PID) for the data set should always be captured and the data provider's page (Landing Page) should be referenced by URL. This ensures that the initial data can be easily cited and are easy to find.

Optionally additional elements from the group No. 8 (Principal Investigator / institution), as well as information on further materials (number 23-26: usage condition; documents, literature) can be used when needed. Thus, references to data documentation or methods reports as supplied by data provider facilitates the traceability of the data and promotes their subsequent re-use in connection with re-analysis of soeb 3 results after the project ends.

Table 1: The metadata elements of the schema Study Description (MD-SD)

No.	Element name	Definition and comments (MD-SD)	FA
	Administration SD	Information about the VRE member and the work package context, in which the study description is processed in the VRE and other administrative elements	
1.	Person	Person who creates and edits the study description	
1.1	First name	First name of the person	AP
1.2	Last name	Last name of the person	AP
1.3	Work package	Work package context in which the object is edited • Selection list: Names of the work packages	M CV
1.4	Project title	Placeholder for VRE project name, e.g. "VRE soeb 3"	AP
2.	Object	Properties of the object "Study Description"	
2.1	Title	The title of the stored object "SD" • Automatically generated from 7.1 Study title	AA
2.2	ID	Unique human-readable ID of the SD • Format e. g. sb001, sb002, etc. • Select from ID list When duplicating the SD the study title is changed and a new SD ID will be assigned	CV
2.3	Date Update	Logical definition of the metadata status as a (new) version by the user. The user specifies the date on which the content was last updated. • Calendar function: Format JJJJ.MM.TT	CL
2.4	Date Description	Logical identification of the metadata status as (new) version. The user describes the state of this version or the changes to the content of a new version, e.g.	FT

No.	Element name	Definition and comments (MD-SD)	FA
		• State of the data provider's metadata according to the date the SD content is captured. • New version of the data set from a data provider that is used in the VRE	
2.5	Version number	Declaration of a numerical version number • Format depends on the versioning concept comprising 2 or 3 digits	FT
2.6	Storage Date	System date on which the file was stored in the archive applying the editor function "Save"	AA
2.7	Publish	Flag to control the publication of SD to avoid fragmentary information e.g. due to interruption of work • Selection: Yes / No	CV
2.8	Object-ID	System-ID of the created SD	AA
3.	Duplicate SD Origin	Details of the originating SD (source object), which are automatically transferred to the duplicate when applying the editor function "Duplicate"	
3.1	Title	Title of the originating SD, which is duplicated • Automatically taken from 2.1	AA
3.2	Object ID	System ID of the originating SD • Automatically taken from 2.8	AA
3.3	Ref Object-ID	Reference to the object ID of the originating SD. The duplicate is linked with the source object by the ID and via link navigable. It displays all the metadata of the originating object.	AA
4.	VRE use Data	Details on the use of the data from this study by the scientific VRE work packages and their member (s)	
4.1	Work packages Name	Specification, which Work Package uses the data set • Choose from a list of all AP (1-n)	CV
4.2	Work Package User	Specification, which person (s) of the AP use the data set • Enter: "Last name, first name" (1-n) for each AP	FT
4.3	Contract Data	Specification whether or not a contract for the use of that data set is necessary • Filter Yes / No	CV
4.4	Contract Note	Comments on specifics of the provision and use of data, e.g. as provided by project management. • If a contract for data use is necessary ("Yes" in field 4.3), the following information must be collected.	FT
4.5	Contract Signatories	VRE member who has concluded the contract (for an AP) • Enter: "Last name, first name" (1-n) for each contract / data set	FT

No.	Element name	Definition and comments (MD-SD)	FA
4.6	Contract Duration	Specifications of the duration of the contract	
4.6.1	Contract Duration from	Duration of the contract per user / AP • Date field from (dd.mm.yyyy)	CL
4.6.2	Contract Duration to	Duration of the contract per user / AP • Date field to (dd.mm.yyyy)	CL
	Data Provider	Source Information of the study, the data and conditions of data use	
5.	Data Provider	Administrative Information about the Data Provider	
5.1	Name	Name of the organization that provides the study / the data and material for use in the project	FT
5.2	Language	Language version of the study (data, materials), which is used in VRE: • German: DEU (Alphanumeric code according to ISO 639-2) • English: ENG (alphanumeric code according to ISO 639-2) • DEU + ENG (in case of mixed linguistic documentation)	CV
5.3	URL	Link to the website of the data provider	URI
6.	Data access	Information from the data provider to Terms of Use and provision of the data set	
6.1	Terms and Conditions	Legal or formal Terms of the data set (without costs) Selection list (CV): • Free (download; possibly taking into account copyright etc.) • Controls (Register and usually explicit agreement to Terms of Use of the data depositor) • Contract (contractual arrangements in writing) • Other	CV
6.1.1	URL	Link to the website informing on Data Usage conditions and data provisioning	URI
6.2	Access	Access to the degree of anonymization and access type Selection list (CV): • Absolute (PUF: Download, CD / DVD, or the like) • In fact, (SUF: Off-Site) • In fact (Guest stay On-site) • Project-Specific Basis (Guest stay On-site) • Formal (Controlled Remote Computing On-Site) Compare source of CV: FDZ Statistische Ämter Bund und Länder	CV

No.	Element name	Definition and comments (MD-SD)	FA
6.2.1	URL	Link to the website with specific information for data access	URI
	Bibliography	Information of the data provider to Bibliography	
7.	Study	Study (data set), used in a soeb 3 WP for data analysis	
7.1	Title	Title of study / data set (according to the data provider information)	M FT
7.2	Other Titles	Specifying additional study titles	FT
7.2.1	Other Titles Type	Type of additional title, for example, the original English title: Selection list (CV): • Original title • Alternative Title • Parallel Title • Subtitles • Project title Source CV: Metadatenschema DBK (Zenk-Möltgen. 2012)	CV
7.3	Study ID	Unique identifier of the study	FT
7.3.1	Study ID Type	Type of studies ID Selection list (CV): • Study No. • Data set Number • Other	CV
7.4	Content	Abstract, brief description of the study, encoded variables, Demography, etc. Information about the study should be compared with the Information about the data set	FT
7.4.1	URL	Link to the page the data provider with further information about the contents of the study	URI
8.	Principal Investigator or institution	Name of the person and / or institution that is (are) responsible for the study or data as an author • 8.1 and / or 8.2 should be specified For data collection see elements at no. 14	
8.1	PI Name	Name of person (Principal Investigator) Depending on the Concept: • 1-n fields for each pair of 8.1.1 and 8.1.2 or • free text field with input per line: last name, first name	
8.1.1	PI first name	Karl	FT
8.1.2	PI last name	Maier	FT
8.2	Institution	Name of the institution, the person is associated with, or which is responsible for study / data • Institute for Comparative Social Research; German Bundesbank, Genesis database	FT

No.	Element name	Definition and comments (MD-SD)	FA
9.	Data publication	Details from the data provider to publication, version, PID, citation of the data set	
9.1	Publication Date	Date on which the data provider published the data set • Format JJJJ.MM.TT	CL
9.2	Publication Version	Version number of the data set at the time of publication by the data provider	FT
9.3	Version Reason	Description of the reason for the new version of the data set.	FT
10.	Persistent Identifier	Persistent Identifier (PID) of the data set	
10.1	PID	E.g. DOI: 10.5684 / soep.v28 for all data of the SOEP wave 28	FT
10.2	PID Type	Type of Persistent Identifier (PID) Selection list (CV): • ARK • DOI • EAN13 • EISSN • Handle • ISBN • ISSN • ISTC • LISSN • LSID • PURL • UPC • URL • URN Source CV: da|ra Metadata Schema 3.0 (Table 3.1.10) (Hausstein et al. 2014)	CV
10.3	PID URL	Reference to data provider's landing page for the PID • e.g. SOEPv28.1: http://dx.doi.org/10.5684/soep.v28.1	URI
10.4	Data citation URL	Link to page with instruction to cite the research data of this study (or general data citation rules) • Example SOEP: http://www.diw.de/sixcms/detail.php?id=diw_01.c.320144.en#277524	URI
	Methodology	Details from the data provider about methodological context and design of data generation	
11.	Reference period	Time period that cover the data content (fieldwork re. surveys; reference period re. statistics)	
11.1.1	Period - from	Information about the period as range: • From: Format JJJJ.MM.TT	CL

No.	Element name	Definition and comments (MD-SD)	FA
11.1.2	Period - to	• To: Format JJJJ.MM.TT	CL
11.2	Period (from - to)	Option to specify a time period when this can't be specified in the calendar mode. • Seasons, e.g. "autumn 1989"	FT
12.	Time dimension	Temporal extent and frequency of data collection	
12.1	Temporal Dimension (controlled)	Temporal extent of data collection – Characterizes Selection list (CV): • Longitudinal section: Trend, cohort, time series • Longitudinal section: Panel • Cross-section • Other Source CV: da|ra Metadata Schema 3.0 (Table 3.1.5) based on the DDI recommendations (Hausstein et al. 2014)	CV
12.2	Temporal Dimension (free)	Option to describe the temporal extent when CV contains no matching terms	FT
12.3	Frequency	Frequency of data collection • monthly, quarterly, annually, etc.	FT
13.	Dimension Space	Describes the geographical unit that underlies the sample (selection)	
13.1	Geographic space (controlled)	Spatial extent in the design of data collection (names of geographical units) •Names of States in accordance with ISO 3166-1 e.g. DE •ISO 3166-2 for their regions, e.g. DE-BY (Bayern)	CV
13.2	Geographic space (free)	Indication of territorial units (within higher ISO units 13.1) or of non-standard units: • spatial territorial units e.g. spatial level "model regions" (Raumordnungsregionen ROR;) , micro census circuit regions (MZKR) or • Northern Germany or closer specifications (FRG without Berlin / W.)	FT
14	Data collection Bases	Methodological bases and methods of data collection	
14.1	Universe	Description of the (statistical) unit the sample is based upon • for example, Households with people who had completed the age of 18 at the time of the survey	FT
14.2	Sampling procedure	Description of the selection process for the creation of -sample (sampling)	FT

No.	Element name	Definition and comments (MD-SD)	FA
		The type of the sample and sample design used to select the survey respondents that represent the population; • e.g. 1980-1992 and 1998 multi-stage stratified by-case selection from all households with persons who completed 18 years at the time of the survey (ADM sample design)	
14.3	Sampling procedure (controlled)	Description of the method by which the data was collected (types) Selection list (CV): • Quantitative interview (Questionnaire) • Qualitative Interview (Guide) • Process-generated data • Observation • Experiment • Content analysis Source CV: da|ra Metadata Schema 3.0 (Table 3.1.7) based on the DDI recommendations (Hausstein et al. 2014)	CV
14.4	Sampling procedure (open)	Free description of the survey method, when the CV provide no suitable term	FT
15.	Data collection Implementation	Name of the institution / person who collected the data, e.g. survey institute Entries are possible in 15.1.x and / or in 15.2. • Facilities such as Statistical offices and German Bundesbank involved in preparing and offering data should be entered as an institution above under Principal Investigator (8.1) and Institution (8.2)	
15.1	Person	Name of the person who collected the data	FT
15.1.1	First name	Elke	FT
15.1.2	Last name	Maier	FT
15.2	Institution	Name of institution that collected the data	FT
	Dataset	Details of the data provider to substantive, formal and technical aspects of the dataset. Substantive Information about the data should be compared with Information about study content 7.4	
16.	Data set Logical	Logical aspects of the data set (content, anonymization, structure, etc.) – Procedure for capturing: • Complex data: description at 16.1 "Structure" with reference to the data provider page and / or • Brief description in the individual sections at 19.x (Data set technical)	
16.1	Structure	General description structure, especially regarding complex data	FT

No.	Element name	Definition and comments (MD-SD)	FA
		structures (hierarchical, multi-dimensional), with references to documentation for understanding of the data set. • For example, "The SOEP has a detailed documentation of the collection and the generated data"	
16.2	Structure URL	e.g. SOEB.v28 - Online - Dokumentation und Neuerungen	URI
17	Anonymization	Information of the data depositor about the anonymisation of the data	
17.1	Degree	Type of anonymization (and data access type) Selection list (CV): • Absolute (PUF) • In fact (SUF) • In fact (On-Site) • Formal (On-Site)	CV
17.2	Degree Description	Description of the anonymisation of the data	FT
18	Data set Formal	Description of formal properties of the dataset	
18.1	Number of Variable	Number of variables in the data set (regardless of the allocation to individual files)	FT
18.2	Unit of analysis Type	Type of units to which the data set applies statements Selection list (CV): • individual • Organization • Family / in the same household • Household / living unit • Event / Process • Geographical Unit • Unit of time • Text unit • Group • Object • Others Source CV: da|ra Metadata Schema 3.0 (Table 3.1.8) based on the DDI recommendations (Hausstein et al. 2014)	CV
18.3	Unit of Analysis No. of Type	Number of units per type of object (regardless of the allocation to individual files) • e.g. Number of respondents (number of cases)	FT
18.4	Data Collection type	Formal type of data collection Selection list (CV): • Survey data • Process data • Statistics	CV

No.	Element name	Definition and comments (MD-SD)	FA
		• Indicators • Other	
19.	Data set Technical	Description of the technical properties of the data set, which consists of 1-n data files. A large number of data files properties can be described by text files applying the related URL a sub element of 20 "Data File Structure".	
19.1	Data File Number	Number of files with equal format that make up the data set	FT
20	Data File Structure	Description of the structure of the data files with complex data sets	
20.1	Description	Systematic overview about e.g. file name, file format, file size, language version for each data file	FT
20.1.1	URL	Reference to the source at the data provider	URI
21.	Single file	Description of a single data files (0-1)	
21.1	File names	File names for each data file (1-n)	
21.2	File size	Memory size of the object no. 21.1	

21.3	Data File
	Format
		Technical format of the data file (s) part of data set
		Selection list (CV):
		• File format - usually quantitative data
		  ○ .RData
		  ○ .Sav
		  ○ .Dta
		  ○ .xls
		  ○ .csv
		  ○ .Sas
		• File Formats - qualitative Data
		  ○ Audio
		  ○ Video
		• File formats – Text
		  ○ Text (.txt, .doc, .pdf, etc.)
		  ○ File formats (other)
		• Other
			CV

22	Data File Fingerprint	Description of the authenticity of a digital object (single file) by a checksum	
22.1	Fingerprint Object	Checksum of a single object no. 21.1 (optional)	FT
22.2	Fingerprint List of files	Overview of files and their checksums, e.g. to soep.v28 section "MD5 fingerprint of each file"	FT
22.2.1	List URL	Example SOEP: Stata English | TXT, 16.16 KB	URI

No.	Element name	Definition and comments (MD-SD)	FA
22.3	Fingerprint Procedure	Technical method by which the checksum is created (possibly with an indication of directory) Selection list (CV): • UNF • MD5 • SHA • RIPEMD-160 • Tiger • HAVAL • Whirlpool • LM hash • NTLM	CV
	Material	Data provider's documentation of the study and the data	
23	Terms and Conditions	Information for use and citation of the material (without data)	
23.1	Terms of use Others	Other rights of use of the website content and citation for materials to the study (e.g. methods report) Selection list (CV): German terms • Copyright • Legal • Copyrights • Imprint English terms • Disclaimer • Terms of Condition • Condition of Use	CV
23.2	Terms of use URL	Reference to the data provider's source	URI
24	Documents	Information about documents, e.g. data documentation, methodology report, measuring instrument etc.	
24.1	Documents Overview	General description of the documentation, in particular for complex data structures of a data set.	FT
24.1.1	Overview URL	Reference to the data provider's source	URI
25.	Documents Single description	Information about individual documents / documentation e.g. the documents (0-n) should to be captured for reasons of transparency and traceability	
25.1	Title	Specify the document title (according to the data provider information)	FT
25.2	Document	Type of document	CV

No.	Element name	Definition and comments (MD-SD)	FA
	Type	Selection list (CV) • Study Description / Summary Data • Data set Description • Survey design and instruments • Questionnaire • Methodological report / Methods Description • Codebook / Data Manual / Data Report • Codeframe • Key list / List of key codes / Key code description • Variable List • Other	
25.3	Document Content	More details on the document content	FT
25.3.1	Document URL	Reference to the data provider's source	URI
	Literature	References to literature that refers to this study / data	
26	Publications	Bibliographic citation	FT
26.1	URL	Reference to the data provider's source	URI

3.3 Metadata on files uploaded and linked to a Study Description

Mandatory elements (M in column FA) are 1.3 Work Package, 2.1 Reference, 3.1 Title and 3.4 Filename.

It is recommended to provide information to all other editable elements (FT / CV in column FA).

Table 2: Metadata elements documenting the storage of files in the VRE (where legally permissible) that are related to a Study Description (SD-Upload).

No.	Element name	Definition and comments (SD-Upload)	FA
	Administration SD Upload	Information about VRE member and work package context, in which the upload of files into the VRE archive is carried out	
1.	Person	Member of the VRE, which stores the file in the VRE archive	
1.1	First name	First name of the person	AP
1.2	Last name	Last name of the person	AP
1.3	Work package	Context in which the object is processed • Selection list: Names of the work packages	M CV
	File	Information about the file, which is stored and linked to a study description (SD)	
2	File type	Type of file that is stored in the archive Selection list (CV): • Data File • Document	CV
2.1	Reference Title SD	Linkage of the file with the title of a study description • Selection list: Title of Study Description (MD-SD No. 2.1.)	M CV
2.2	Ref Parent ID	Linkage of the file with the system ID of the SD	AA
3	Property	More details on the stored file	
3.1	Title	Title of the stored file (according to the data provider information)	M FT
3.2	Type	Specification of the object type as specified at no. 2 Information is refers either to 3.2.1 (Data File) or 3.2.2 (Document)	
3.2.1	Type data file	Type of data file Selection list (CV): • File is part of a data set • Single Data File	CV
3.2.2	Type Document	Type of the document Selection list (CV):	CV

No.	Element name	Definition and comments (SD-Upload)	FA
		• Study Description / Short description of data • Data set description • Survey design and instruments • Questionnaire • Methods report / Methods Description • Codebook / Data Manual / Data Report • Codeframe • Key list / List of key codes / Key code description • Variable List • Other	
3.3	Content	Description of the stored file	FT
3.4	Filename	Full filename (descriptor, format) according to the data provider information	M FT
3.4.1	File size	Specifications in GB / MB / Kbyte	AA
3.4.2	Storage Date	System date on which the file was stored in the archive applying the editor function "Save"	AA
3.4.3	Fingerprint	Fingerprint of the data provider (if available)	FT
3.5	Approval Terms of use	Terms of use of the file by additional VRE member(s) and / or WP(s) after approval • SD No. 23.1 Terms of use according to the data provider regulations	FT
3.6	Approval	Assignment of defined VRE rights of use per person / per work package • According to VRE role and rights management	CV
3.6.1	Approval Date	Date on which the approval was issued. Declaration is set automatically after storing of 3.6	AA

4 The metadata schema Data Usage

4.1 Introduction

With the schema Data Usage, metadata describe the use of the initial data, in particular the subject of investigation and methodological parameters of the planned data analyses in the project VRE soeb 3. The Data Usage schema is formally separated from metadata schema Syntax. With Data Usage information on the planning of analyses are documented, while the Syntax metadata document the data analysis performed (see chapter 5).

Information on data use should enable researchers from VRE soeb 3 which issues and subjects of investigation are examined with which data and parameters, e.g. explore with the help of the system searches. Thus, the following aspects of collaborative data analyses can be systematically investigated and answered due to metadata searches:

- The study data will be used for which subjects and investigation objectives?
- What data set type is used, for example: Cross-Section, Longitudinal, or Wave/Year?
- For which time period? For which geographical area? What is the unit of analysis?
- What is the smallest territorial unit at which the data are analysed, e.g. county level, ROR?
- What analyses methods and procedures are used by users or in the work package?
- Which data sets are generated from the initial data?

Matched data or planning memos files can be linked to the Data Usage schema. Data files could be compiled by project members e.g. from databases like Genesis (Federal Statistical Office) or be prepared as subset data for analysis from other initial data. The descriptions of such 'self-created' (project driven) data are logically associated with the Schema Data Usage, while the source of these data – the initial data - are always documented by the schema Study description.

4.2 The metadata elements of the schema Data Usage

Mandatory elements (M in column FA) are 1.3 Work Package, 4.1 and 5.1 Reference Title.

It is recommended to provide information to all other editable elements (FT / CV in column FA).

Table 3: Metadata elements of the schema Data Usage (MD-DU)

No.	Element Name	Definition and comments (MD-DU)	FA
	Administration DU	Information about the VRE member and the work package context, in which the data use description (MD-DU) is processed in the VRE and other administrative elements	
1	Person	Person who creates and edits the Data Usage	
1.1	First name	First name of the person	AP
1.2	Last name	Last name of the person	AP
1.3	Work package	WP context in which the object is edited	M

No.	Element Name	Definition and comments (MD-DU)	FA
		• Selection list: Names of the work packages	CV
2	Object Data Usage	Information about the object Data Usage	
2.1	Title	The title of the stored object Data Usage • automatically generated from field 4.1 Title	AA
2.2	ID	Unique human readable ID of the Data Usage • Format e. g. DU001, DU002, etc. • Selection from ID list When duplicating the Data Usage the title is changed and a new ID will be assigned	CV
2.3	Date Update	Logical definition of the metadata status as a (new) version by the user. The user specifies the date on which the content was last updated. • Calendar function: DD.MM.YYYY	CL
2.4	Date Description	Logical identification of the metadata status as (new) version. The user describes the state of this version or the changes to the content of a new version	FT
2.5	Version number	Declaration of a numerical version number • Format depends on the versioning concept comprising 2 or 3 digits	FT
2.6	Storage Date	System date on which the file was stored in the archive applying the editor function "Save"	AA
2.7	Publish	Flag to control the publication of Data Usage to avoid fragmentary information e.g. due to interruption of work • Selection: Yes / No	CV
2.8	Object-ID	System-ID of the created Data Usage	AA
3.	Duplicate MD-DU Origin	Details of the originating Data Usage, which are automatically transferred to the duplicate when applying the editor function "Duplicate"	
3.1	Title	Title of the originating Data Usage • Automatically taken from 2.1	AA
2.8	Object-ID	System-ID of the created Data Usage	AA
3	Duplicate DU Origin	Details of the originating Data Usage (source object), which are automatically transferred to the duplicate when applying the editor function "Duplicate"	
3.1	Title	Title of the originating Data Usage, which is duplicated • Automatically taken from 2.1	AA
3.2	Object ID	System ID of the originating Data Usage	AA

No.	Element Name	Definition and comments (MD-DU)	FA
		• Automatically taken from 2.8	
3.3	Ref Object-ID	Reference to the object ID of the originating Data Usage. The duplicate is linked with the source object by the ID and via link navigable. It displays all the metadata of the originating object.	AA
	Subject	Information about the subject of investigation, used study / data and methods aspects of the investigation	
4	Topic	Description of the subject / objective of the investigation	
4.1	Title	Title of the subject of investigation	M FT
4.2	Content	Description of the subject of investigation, for which the study and the initial data are used	FT
5	Initial Data	Description of the specific source data from a study that will or should be used for the analysis	
5.1	Reference Study	Selection of the study with a description of the initial data, which are used for the analysis (SD) • Selection list: Title of Study Description (MD-SD No. 2.1.)	M AA
5.2	Data set Type	Type of (sub-) data set (from 1-n data files) of the study, which is used for data analysis (without temporarily produced data sets that can be documented in the schema syntax) Selection list (CV): • Data subset (from single data file • Individual data set (original data file) • Self-generated data (data from database, etc.) • Matched data file (created in advance from several data files)	CV
5.3	Description Data set	Description of the selected data set (1-n data files) and its specific use for a planned analyses • e.g. Subset description, data set structure, description of self-generated DS, specific application of specialised formats (Long Format; Wide Format); use of spell data etc.	FT
6	Method	Methodological parameter, with which the object is examined	
6.1	Period	Specification of the time frame (reference period) , for which the data will be used	FT
6.2	Geographical Area	Description of the geographical area (country / subunit), for which the data will be used • Names (not code) in accordance with ISO 3166-1 for States and ISO 3166-2 for their regions	FT
6.3	Territorial Unit	Description of the smallest territorial unit (e.g. county level ROR) are used for data Selection list (CV) [Note: CV regard German units only]: • Administrative territorial unit	CV

No.	Element Name	Definition and comments (MD-DU)	FA
		• Region • District • Municipality • Statistical Area Unit • ROR (Raumordnungsregionen / spatial level model regions) • BIK Classification (System, describing the urban-rural relations at the community level) • Other territorial unit • Postcode • Tax Office District • Court District • Employment Agency District • Federal election district • State election district • Federal District • Chamber district HWK (Chamber of Crafts) • Chamber district IHK (Chamber of Industry and Commerce) • Other	
6.4	Analysis unit	Description for which unit (s) of analysis, the data will be used	FT
6.5	Methods	Description which methods are applied for the investigation / data analysis	FT
7	Data Analysis	Conceptual description of the planned data analyses with respect to the subject of investigation	
7.1	Analysis Concept	Concept specification of data analysis on concept level, e.g. described as data analysis plan. The operationalization of the concept is carried out by the syntaxes that are linked to the description of this Data Usage and the subject of investigation as documented.	FT

4.3 Metadata on files uploaded and linked to a Data Usage description

Mandatory elements (M in column FA) are 1.3 Work Package, 2.1 Reference, 3.1 Title and 3.4 Filename.

It is recommended to provide information to all other editable elements (FT / CV in column FA).

Table 4: Metadata elements documenting the storage of files in the VRE (where legally permissible), which are related with a Data Usage description (DU-upload)

No.	Element Name	Definition and comments (DU-Upload)	FA
	Administration DU-Upload	Information about VRE member and work package context, in which the upload of files into the VRE archive is carried out	
1	Person	Member of the VRE, which stores the file in the VRE archive	
1.1	First name	First name of the person	AP
1.2	Last name	Last name of the person	AP
1.3	Work package	Context in which the object is processed • Selection list: Names of the work packages	M CV
	File	Information about the file that is stored and associated with a Data Usage (MD-DU).	
2	File type	Type of file that is stored in the archive Selection list (CV): • Data File • Document	CV
2.1	Reference Title SD	Title of a subject of investigation with which the stored file will be linked • Selection list: Title of Data Usage (MD-DU No. 2.1.)	M CV
2.2	Ref Parent ID	Linkage of the file with the system ID of the selected DU	AA
3	Property	More details on the stored file	
3.1	Title	Title of the stored file	M FT
3.2	Type	Specification for the object type as specified at no. 2 • Information is refers either to 3.2.1 (Data File) or 3.2.2 (Document)	
3.2.1	Type Data File	Type of data file Selection list (CV): • Sub-data set (from single data file • Individual data set (original data file) • Self-generated data (data from database, etc.) • Matched data file (created in advance from several data files)	CV
3.2.2	Type Document	Type of the document	CV

No.	Element Name	Definition and comments (DU-Upload)	FA
		Selection list (CV): • Description of a self-generated data set • Description of a single subject of investigation • Description of the analysis planning • Other document	
3.3	Content	Description of the stored file	FT
3.4	Filename	Full filename (descriptor, format) according to the data provider information	M FT
3.4.1	File size	Specifications in GB / MB / Kbyte	AA
3.4.2	Storage Date	System date on which the file was stored in the archive applying the editor function "Save"	AA
3.4.3	Fingerprint	Fingerprint of the stored file (if implemented in VRE / or available from elsewhere)	
3.5	Approval Terms of use	Terms of use of the file by additional VRE member(s) and / or WP(s) after approval	FT
3.6	Approval	Assignment of defined VRE rights of use per person / per work package • According to VRE role and rights management	CV
3.6.1	Approval Date	Date on which the approval was issued. Declaration is set automatically after storing of 3.6	AA

5 The metadata schema Syntax

5.1 Introduction

The metadata schema Syntax contains elements for describing the tasks and contents of syntax files used in performing data analyses. Syntax file used during the work process of data analyses can be distinguished by the categories Data selection, Data processing, Data linkage and Data analysis (cf. Dickmann et al. 2010.: 9).

Furthermore, the metadata elements in this schema document the creation and editing of syntax contents and their versions. The documentation of the syntax in the VRE thus serves both the internal quality assurance as well as the cooperation in the production and development of syntax files during the data evaluation. Due to a corresponding rights management, the syntax files can be shared with other researchers and / or scientific work packages in the soeb 3 project. For example, work package 11 provides syntax blocks or descriptive metadata to allow the harmonized generation of household variables wherever possible. To secure the origin of syntaxes and syntax blocks information with the author and regulation on citations can be documented in the corresponding metadata item "Terms of Use".

Finally, the output files created by applying a software specific syntax file, can be gathered, documented and managed during upload to the VRE archive by appropriate metadata. These include, for example, log files, listing, graphs or tables and spreadsheets.

5.2 The metadata elements of the schema Syntax

Mandatory elements (M in column FA) are 1.3 Work Package, 4.1 Reference, 7.1 Title, and 7.3 Filename.

It is recommended to have all other editable elements (FT / CV in column FA).

Table 5: The metadata elements of the schema Syntax (MD SX)

No.	Element Name	Definition and comments (MD-SX)	FA
	Administration SX	Information about the VRE member and the work package context, in which the syntax description (MD-SX) is processed in the VRE and other administrative elements	
1	Person	Person who creates and edits the Syntax description	
1.1	First name	First name of the person	AP
1.2	Last name	Last name of the person	AP
1.3	Work package	WP context in which the object is edited • Selection list: Names of the work packages	M CV
2	Object Syntax	Information about the object Syntax description	
2.1	Title	The title of the stored object Syntax • automatically generated from field 7.1 Title	AA
2.2	ID	Unique human readable ID of the Data Usage	CV

No.	Element Name	Definition and comments (MD-SX)	FA
		• Format e. g. DU001, DU002, etc. • Selection from ID list When duplicating the Data Usage the title is changed and a new ID will be assigned	
2.3	Date Update	Logical definition of the metadata status as a (new) version by the user. The user specifies the date on which the content was last updated. • Calendar function: DD.MM.YYYY	CL
2.4	Date Description	Logical identification of the metadata status as (new) version. The user describes the state of this version or the changes to the content of a new version	FT
2.5	Version number	Declaration of a numerical version number • Format depends on the versioning concept comprising 2 or 3 digits	FT
2.6	Storage Date	System date on which the file was stored in the archive applying the editor function "Save"	AA
2.7	Publish	Flag to control the publication of a- MD-SX to avoid fragmentary information e.g. due to interruption of work • Selection: Yes / No	CV
2.8	Object-ID	System-ID of the created MD-SX	AA
3	Duplicate MD-SX Origin	Details of the original MD-SX, which are automatically transferred to the duplicate when applying the editor function "Duplicate"	
3.1	Title	Title of the originating MD-SX • Automatically taken from 2.1	AA
3.2	Object ID	System ID of the originating MD-SX • Automatically taken from 2.8	AA
3.3	Ref Object-ID	Reference to the object ID of the originating MD-SX. The duplicate is linked with the source object by the ID and via link navigable. It displays all the metadata of the originating object.	AA
	Syntax property	Formal details about the syntax file	
4	Subject	Object, purpose, subject of the investigation, which is associated with the syntax. The claim is made at 4.1 (Title) and 4.2 (content) in the schema Data Usage description and linked via the following element (4.1 Reference Data Usage) to the Syntax description at hand.	
4.1	Reference Data Usage	Selection of the Data Usage description, which is related to this MD-SX • Selection list:	

No.	Element Name	Definition and comments (MD-SX)	FA
		Title of Data Usage (MD-DU No. 2.1.)	
5	Syntax file structure	Specification of the syntax file structure (single file, syntax set, module)	
5.1	File structure Type	Type of the file structure of the syntax Selection list (CV): • Single Syntax: File consists of a single syntax • Syntax set: File is part of a set of syntax files • Syntax module: e.g. a household typology, which can be associated with a syntax file	CV
6	Syntax Statistics Program	Statistics software with which the syntax is executable	
6.1	Name	Name of the software, to execute the syntax Selection list (CV): • STATA • SPSS • R	CV
6.2	Version	Software version with which the syntax is executable	FT
6.3	Packages	Additional package (and it's version), which is required to run the syntax (Packages / Ados)	FT
7	Syntax file & Storage	Informs on the title and the storage of syntax files (as described at MD-SX) at the VRE archive	
7.1	Title	Title of the syntax file	M FT
7.2	Description	General description of the syntax (abstract). A specification for each syntax function (cf. Dickmann et al. 2010: 9) is performed separately after • selecting a related category in item 9.1 and their descriptions in the element groups 10 to 14	E FT
7.3	Filename	Full filename (descriptor, format) of the stored file	M FT
7.4	Storage Date	System date on which the file was stored in the archive applying the editor function "Save"	AA
8.	Syntax sharing Approval	Regards the sharing of syntax files with other VRE members or work packages, by awarding appropriate permissions	
8.1	Approval Terms of use	Terms of use of the file by additional VRE member(s) and / or WP(s) after approval	FT
8.2	Approval	Assignment of defined VRE rights of use per person / per work package • According to VRE role and rights management	CV

No.	Element Name	Definition and comments (MD-SX)	FA
8.3	Approval Date	System date on which the approval was granted The date is set when the syntax file was stored in the archive applying the editor function "Save" (see 2.6 MD-SX)	AA
	Syntax Function Task	Rough categorization of syntax functions and task description of a single syntax file. Exact figures are given in element groups 10 to 14	
9	Syntax Function	Description of syntax function and the task to be performed by the syntax	
9.1	Function Category	Statement about the syntax task or function by category Selection list (CV): • Data selection: Create variable subset from initial data (details at section 10) • Data processing: Recode variable etc. (details at section 11 and 12) • Data linkage: Create data file for analysis (details at section 13) • Data analysis: Run analysis with syntax (details at section 14)	CV
9.2	Task Description	Description of the specific syntax task depending on the selected function category from 9.1	FT
9.3	Task Status	Development status of syntax Selection list (CV): • Design • Work in Progress • Final	CV
	Syntax Function Operationalization	Description of the execution of a task (the planned operationalization) depending on the selected category 9.1	
10	Syntax Data Selection	Description of the syntax, which read one or more data files (source / initial data)	
10.1	Initial Data Description	Specification of the read data set(s) (initial or subset data)	FT
10.2	Initial Data File name	Filename of the read data file (or data files)	FT
11	Syntax Data processing	Description of the syntax, which further processes the read data, for example, by recoding	
11.1	Data processing Description	General description of the data processing procedures. A specific description of the generation of individual new variables in the course of data processing is possible with the	FT

No.	Element Name	Definition and comments (MD-SX)	FA
		elements of the following group	
12	Data processing New variable	Description of the syntax (e.g. recode, generate), which creates or modifies one or more variable based upon selected variables from the source data file(s). • It is possible to document modifications of 0-n variables per syntax	
12.1	Concept	Technical description of how and for what purpose the new variable is formed (n = 1 each new variable), like • Concept description regarding a typology • Instructions for the recoding of source variable(s) to form a new variable	FT
12.2	Variable Name	Name for each new variable (n = 1)	FT
12.3	Variable Label	Label for each new variable (n = 1)	FT
12.4	Value	Value of the new variable (1-n each new variable)	FT
12.5	Value label	Value label (n = 1 per new Value)	FT
12.6	Missing Value	Missing value of the new variable (1-n each new variable)	FT
12.7	Missing Value Label	Missing Value Label (n = 1 per new Missing Value)	FT
13	Syntax Data Linkage	Description of the syntax, which creates the data file based upon read and / or modified variables to be applied for analysis (analysis file)	
13.1	Description	Description of the newly created data set (analysis file)	FT
13.2	Filename	Full filename (descriptor, format) of the data file, which is re-created by the syntax	FT
14	Syntax Data Analysis	Description of the syntax, which executes the data analysis	
14.1	Description	Description of the used computing and analysis method(s)	FT

5.3 Metadata on files uploaded and linked to a Syntax description

Mandatory elements (M in column FA) are 1.3 Work Package, 2.1 Reference, 3.1 Title and 3.4 Filename.

It is recommended to have all other editable elements (FT / CV in column FA).

For conceptual reasons, the upload of a syntax file is also included with this subschema. Metadata elements of this sub-schema concern only the description of the syntax file itself.

All elements with the information describing the syntax (content, function, etc.) and the associated syntax file are listed in the above schema Syntax (MD-SX). For this reason, reference is made in the following schema (SX-Upload) to the syntax element numbers in the schema MD-SX.

As in the case of the schema Study and Data Usage here also a file can only be linked with the description of Syntax if this was already compiled.

Table 6: Metadata elements documenting the storage of files in the VRE (where legally permissible), which are related with a Syntax file description (SX-Upload)

No.	Element Name	Definition and comments (SX-Upload)	FA
	Administration SX-Upload	Information about VRE member and work package context, in which the upload of files into the VRE archive is carried out	
1	Person	Member of the VRE, which stores the file in the VRE archive	
1.1.	First name	First name of the person	AP
1.2.	Last name	Last name of the person	AP
1.3	Work package	Context in which the object is processed • Selection list: Names of the work packages	M CV
	File	Information about stored file that is associated with a MD-SX	
2	File Type	Type of file that is stored in the archive. Selection list (CV): • Syntax file • Output files related to a syntax file	CV
2.1	Reference Title syntax	Title of the MD-SX, with which the uploaded file is linked. Only valid for the appropriate type under No. 2 Selection list: **Title** of Syntax Description (MD-SX No. 7.1.)	M
2.2	Ref Parent-ID	Linkage of the file with the system ID of the selected MD-SX	AA
3	Property	More details on the stored file	
3.1 (7.1)	Title	Title of the uploaded file • If it is a syntax file: MD-SX No. 7.1 is applied	M FT
3.2	Type	Specification of object type as specified at no. 2 Information regards either 3.2.1 (Syntax file) or 3.2.2 (Output File)	
3.2.1	Type	Type of the syntax file:	CV

No.	Element Name	Definition and comments (SX-Upload)	FA
(5.1)	Syntax file	Selection list (CV) re. MD-SX no. 5.1 • Single Syntax: File is a single syntax • Syntaxset: File is part of a set of syntax files • Syntax module: E.g. a household typology	
3.2.2	Type Output File	Type of the related output file Selection list (CV): • Documentation of the syntax flow: Log file • Data set: Work data set, finalised analysis file • Result file: Listings, table, spreadsheet, graphic, other • Variable list: VRE-generated variables	CV
3.3 (7.2)	Content	Description of the stored file • If syntax file: MD-SX No. 7.2 is applied	FT
3.4 (7.3)		Full filename (descriptor, format) • If syntax file: MD-SX No. 7.3 is applied	M FT
3.4.1	File size	Specifications in GB / MB / Kbyte	AA
3.4.2 (7.4)	Storage Date	System date on which the file was stored in the archive applying the editor function "Save" • If syntax file: MD-SX No. 7.4 is applied	AA
3.4.3	Fingerprint	Fingerprint of the stored file (if implemented in VRE / or available from elsewhere)	
3.5 (8.1)	Approval Terms of use	Terms of use of the file by additional VRE member(s) and / or WP(s) after approval • If syntax file: MD-SX No. 8.1 is applied	FT
3.6 (8.2)	Approval	Assignment of defined VRE rights of use per person / per work package • According to VRE role and rights management • If syntax file: MD-SX No. 8.1 is applied	CV
3.6.1 (8.3)	Approval Date	Date on which the approval was issued. Declaration is set automatically after storing of 3.6 • If syntax file: MD-SX No. 8.3 is applied	AA

References

Bartelheimer, Peter. Schmidt, Tanja. 2011
Modellprojekt "Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung" - VirtAug - Abschlussbericht. soeb-Arbeitspapier 2011-1.
http://www.soeb.de/fileadmin/redaktion/downloads/VirtAug/VirtAug_Abschlussbericht.pdf

Dickmann, Frank. Enke, Harry. Harms, Patrick. 2010
Technische Evaluation der Grid-Technologie für das Modellprojekt Kollaborative Datenauswertung und virtuelle Arbeitsumgebung – VirtAug. SOEB Arbeitspapier 2010-1.
http://www.soeb.de/fileadmin/redaktion/downloads/VirtAug/Expertise_VirtAug.pdf

DDI Alliance. 2014.
Data Documentation Initiative http://www.ddialliance.org/

DDI 2 - Codebook version
http://www.ddialliance.org/Specification/DDI-Codebook/

DDI Empfehlungen zu kontrollierten Vokabularen:
http://www.ddialliance.org/Specification/DDI-CV/

Friedhoff, Stefan. Meier zu Verl, Christian. Pietsch, Christian. Meyer, Christian. Vompras, Johanna. Liebig, Stefan. 2013.
SFB 882 Working Paper Series, 16. Bielefeld: DFG Research Center (SFB) 882. From Heterogeneities to Inequalities.
urn:nbn:de:0070-pub-25600355

Hausstein, Brigitte. Quitzsch, Nicole. Jeude, Kirsten. Schleinstein, Natalija. Zenk-Möltgen, Wolfgang. 2012. da|ra Metadata Schema Version 2.2.1 (Released August 8th, 2012)
GESIS-Technical Reports 2013|03
http://dx.doi.org/10.4232/10.mdsdoc.2.2.1

Hausstein, Brigitte. Schleinstein, Natalija. Koch, Ute. Meichsner, Jana. Becker, Kerstin. Stahn, Lena-Luise. 2014. da|ra Metadata Schema. Version 3.0. (Released March 11th, 2014)
GESIS-Technical Reports 2014|07
DOI: 10.4232/10.mdsxsd.3.0

Jensen, Uwe. 2014
Fachtagung "Das Portal, die Daten und wir - Eine Virtuelle Forschungsumgebung für die digitale Infrastruktur". Berlin 24.01.2014. http://www.soeb.de/vfu-soeb-3/abschlusstagung/

Panel 4: Gesucht, gefunden – Metadaten zur Nutzung von Forschungsdaten
Präsentation: Einführung in das Metadatenschema der VFU soeb 3
http://www.soeb.de/fileadmin/redaktion/downloads/VFU/VFU_Tagung_Jensen_final.pdf

Jensen, Uwe. EDDI 2014
6th Annual European DDI User Conference. London. December 2. -3.2014.
Presentation „Metadata Requirements to document Data Analyses and Syntax Files in a Virtual Research Environment (VRE) – The use case soeb 3"
http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14/paper/view/140/112

Jensen, Uwe; Schweers, Stefan; Carevic, Zeljko. 2014
Die Metadateneditoren der VFU soeb 3
GESIS-Technical Reports – 2014/14
http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2014/TechnicalReport_2014-14.pdf

Schmidt, Tanja. 2014a
Grundlegende Informationen zur Benutzung der Virtuellen Forschungsumgebung (VFU) für soeb 3 (Version 0.1). Januar 2014.
SOFI - Soziologisches Forschungsinstitut Göttingen an der Georg-August-Universität
http://www.soeb.de/fileadmin/redaktion/downloads/VFU/Grundlegende_Informationen_zur_Nutzung_der_VFU_v01.pdf

Schmidt, Tanja. 2014b
Fachtagung "Das Portal, die Daten und wir - Eine Virtuelle Forschungsumgebung für die digitale Infrastruktur". Berlin 24.01.2014: http://www.soeb.de/vfu-soeb-3/abschlusstagung/

Panel 2: Erprobt – Nutzungserfahrungen und Forschungspraxis
Präsentation: „Ergebnisse aus zwei Nutzungsstudien"
http://www.soeb.de/fileadmin/redaktion/downloads/VFU/VFU_Tagung_Schmidt_final.pdf

SOFI. 2013.
Soziologisches Forschungsinstitut Göttingen an der Georg-August-Universität
Projektbeschreibung: Verbundprojekt „Virtuelle Forschungsumgebung für die sozioökonomische Berichterstattung" (VFU soeb 3)
http://www.soeb.de/fileadmin/redaktion/downloads/VFU/Projektbeschreibung_VFU_SC_081013.pdf

VFU soeb 3. Projekt „Virtuelle Forschungsumgebung" (VFU)
VFU Projektseite:
http://www.soeb.de/vfu-soeb-3/

Über soeb 3 - Dritter Bericht zur sozioökonomischen Entwicklung in Deutschland:
http://www.soeb.de/ueber-soeb-3/

Zenk-Möltgen, Wolfgang. Habbel, Norma. 2012
Der GESIS Datenbestandskatalog und sein Metadatenschema - Version 1.8.
GESIS-Technical Reports 2012|01.
http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-01.pdf