

1660²⁰²⁶

SOEP Survey Papers
Series G – General Issues and teaching Materials

Die Tabellen des SOEP-Metadatensystems

Knut Wenzig

Running since 1984, the German Socio-Economic Panel (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing. The SOEP Survey Papers is comprised of the following series:

Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at <http://www.diw.de/soepsurveyspapers>

Editors:

Dr. Jan Goebel, DIW Berlin
Dr. Christian Hunkler, DIW Berlin
Prof. Dr. Philipp Lersch, DIW Berlin and Humboldt-Universität zu Berlin
Dr. Levent Neyse, DIW Berlin and Berlin Social Science Center (WZB)
Prof. Dr. Carsten Schröder, DIW Berlin and Freie Universität Berlin
Prof. Dr. Sabine Zinn, DIW Berlin and Humboldt-Universität zu Berlin

Please cite this paper as follows:

Knut Wenzig, 2026. Die Tabellen des SOEP-Metadatensystems. SOEP Survey Papers 1660: Series G – General Issues and teaching Materials. Berlin: DIW Berlin/SOEP



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

© 2026 by SOEP

ISSN: 2193-5580 (online)

DIW Berlin

German Socio-Economic Panel (SOEP)

Anton-Wilhelm-Amo-Str. 58

10117 Berlin

Germany

soeppapers@diw.de

Die Tabellen des SOEP-Metadatensystems

Knut Wenzig

2026

Inhaltsverzeichnis

1	Einleitung	3
2	Beschreibung der Tabellen	3
	Kerntabelle	3
	Datensatz-Dokumentation	4
	Fragebogen-Dokumentation	4
	Verbindungstabellen	5
	Themen und Konzepte	6
	Kontrollierte Vokabulare	7
3	Beziehungen zwischen den Tabellen	7
	Hierarchie der Datensatz-Seite des Modells	7
	Hierarchie der Fragebogen-Seite des Modells	8
	Verbindung beider Seiten	8
	Variablenfluss	8
	Fremdschlüsselbeziehungen zwischen Datensätzen	8
	Anhänge	8
4	Entity-Relationship-Diagramm (ERD)	9
5	Technische Details	11
	Grundprinzip: ein git-versioniertes CSV-Repository	11
	Ablagestruktur	11
	Die soep-meta-Konvention	11
	Export nach paneldata.org	12
	Bezug zu Metadatenstandards	12
	Anmerkung zur Konsistenz	13
	Bekannte Eigenheiten und Designentscheidungen	13
6	Zusammenfassung und Ausblick	14
7	Danksagung und Hinweise	14
8	Anhang: Spaltenreferenz	15
	Kerntabelle	15
	Datensatz-Dokumentation	16
	Fragebogen-Dokumentation	21
	Verbindungstabellen	24
	Themen und Konzepte	26
	Kontrollierte Vokabulare	27

1 Einleitung

Das Sozio-oekonomische Panel (SOEP) ist eine seit 1984 laufende repräsentative Längsschnittstudie, die am DIW Berlin durchgeführt wird. Jährlich werden dieselben Personen und Haushalte zu Themen wie Einkommen, Erwerbstätigkeit, Gesundheit, Bildung und subjektivem Wohlbefinden befragt. Im Laufe von vier Jahrzehnten sind dabei Zehntausende von Variablen in Hunderten von Datensätzen entstanden, die unterschiedlichen Erhebungsinstrumenten, Zielpopulationen und Versionierungsständen zuzuordnen sind.

Die Verwaltung der Metainformationen dieser Studie – von der Fragebogenstruktur mit Frage- und Antwortformulierungen über Datensätze mit Variablenbezeichnungen, Wertelabels bis hin zu Datensatzbeziehungen – stellt eine eigenständige Infrastrukturaufgabe dar.

Dieser Beitrag beschreibt das Metadatensystem des SOEP und enthält in den nächsten Abschnitten eine Beschreibung der Tabellen mit ihren Funktionen, wie diese Tabellen zusammenhängen und einige technische Details, beispielsweise das Prinzip der Selbstdokumentation (vgl. Seite 11). Im Anhang findet sich eine Spaltenreferenz für jede Tabelle.

2 Beschreibung der Tabellen

Das Metadatensystem umfasst 24 Tabellen, die in sieben thematische Gruppen gegliedert sind. (Vgl. Tabelle 1) Im Prinzip bilden sie eine Struktur verbundener Tabellen, wie sie aus relationalen Datenbanken bekannt sind. Viele Tabellen verfügen auch über Primärschlüssel, mit denen einzelne Zeilen in einer Tabelle eindeutig identifizierbar werden.

Tabelle 1: Übersicht der Tabellengruppen im SOEP-Meta-Datenmodell

Gruppe	Tabellen
Kerntabelle	studies
Datensatz-Dokumentation	logical_datasets , datasets_versions , variables , variable_categories , codebook , variables_images
Fragebogen-Dokumentation	questionnaires , questions , answers , questions_images
Verbindungstabellen	logical_variables , generations , attachments , dataset_links
Themen und Konzepte	topics , concepts
Kontrollierte Vokabulare (studienbezogen)	analysis_units , conceptual_datasets , periods , versions
Kontrollierte Vokabulare (studienübergreifend)	questionnaire_types , scale_types , minedition_types

Kerntabelle

[studies.csv](#) definiert eine Studie mit einem Identifikator in der Spalte [study](#) und einem Titel (in der Spalte [label](#)) pro Studie. Innerhalb eines einzelnen Repositories enthält sie typischerweise nur eine Datenzeile. Der Primärschlüssel ist [study](#); er dient als oberster Anker für

alle anderen Tabellen. Die Spalten **description** und **html_description** sind im Produktivsystem derzeit ungenutzt; die Studienbezeichnung für paneldata.org wird aus einer separaten Markdown-Datei (**study.md**) bezogen.

Datensatz-Dokumentation

logical_datasets.csv ist die zentrale Registrierungstabelle für alle Datensätze. Sie enthält je eine Zeile pro Datensatz mit zweisprachiger Bezeichnung und Beschreibung, Angabe der Analyseeinheit, des Datensatztyps und des Erhebungszeitraums sowie den Primär- und Fremdschlüsseln des Datensatzes. Die Spalte **minedition** gibt die Zugangsstufe des Datensatzes an (siehe Abschnitt zu bekannten Eigenheiten). Die Tabelle ist versionsagnostisch: Sie beschreibt Datensätze über alle Versionen hinweg.

variables.csv enthält je eine Zeile pro Variable pro Datensatz-Version. Sie bildet zusammen mit **variable_categories.csv** und **codebook.csv** die Kerndokumentation eines Datensatzes. Der Primärschlüssel ist die Kombination aus **study**, **dataset**, **version** und **variable**. Neben zweisprachiger Bezeichnung und Beschreibung enthält die Tabelle die Spalten **type** (Datentyp), **concept** (Verweis auf **concepts.csv**) und **minedition** (Zugangsstufe der Variable). Die Sortierung der Variablen in der Tabelle entspricht idealerweise der Sortierung der Variablen im Datensatz; in SOEP-Core wird dies erzwungen.

variable_categories.csv enthält je eine Zeile pro gelabeltem Wert einer Variable (je Datensatz und Version) – also die Wertelabels für kategoriale Variablen und/oder Missing-Werte. Der Primärschlüssel umfasst **study**, **dataset**, **version**, **variable** und **value**. Numerische Variablen ohne Wertelabels haben keine Einträge in dieser Tabelle. Die Standardsortierung ist aufsteigend nach Variable und Wert; die Zeilen einer Variable sollten nicht unterbrochen werden.

codebook.csv steuert Struktur und Texte des Codebook-PDFs. Sie enthält je eine Zeile pro Variable oder Abschnittsüberschrift. Zeilen ohne **variable**-Eintrag ermöglichen, Abschnitte zu definieren, die nur aus Überschrift und Text bestehen; Zeilen mit **variable**-Eintrag entsprechen genau einer Variablen. Neben Abschnittstexten enthält die Tabelle die Wellenangaben (**waves**, **years**) und Kontaktinformationen. Die Reihenfolge der Zeilen ist relevant und entscheidet über die Reihenfolge der Abschnitte und Variablen im Dokument; ein Abschnitt darf nicht unterbrochen werden. Der Primärschlüssel besteht aus **study**, **dataset**, **version**, **section** und **variable**.

datasets_versions.csv listet auf, welcher Datensatz im Release welcher Version veröffentlicht wurde. Der Primärschlüssel ist die Kombination aus **study**, **dataset** und **version**. Diese Tabelle steuert, welche Datensätze in einen Versionsexport aufgenommen werden.

variables_images.csv verknüpft Bilder mit Variablen, etwa Abbildungen mit deskriptiven bivariaten Statistiken, die die Entwicklung der Variablenwerte im Zeitverlauf zeigen. Im Unterschied zu **attachments.csv** ist sie auf Variablen spezialisiert und unterstützt separate URLs für die deutsch- und englischsprachige Bildversion. Die Tabelle ist versionsagnostisch: Sie kennt keine **version**-Spalte. (Eine Ergänzung der **version**-Spalte ist für eine künftige Revision vorgesehen.)

Fragebogen-Dokumentation

questionnaires.csv enthält je eine Zeile pro Erhebungsinstrument (Fragebogen). Sie dokumentiert Typ (**qtyp**), Modus, Erhebungsperiode, Stichprobengruppe sowie zweisprachige

Bezeichnung und Beschreibung. Die Spalte **publish** steuert, ob ein Instrument nach paneldata.org exportiert wird. Der Primärschlüssel ist die Kombination aus **study** und **questionnaire**.

questions.csv enthält alle Frageitems eines Fragebogens – ausgenommen Antwortoptionen für Einfachauswahl-Fragen (in **answers.csv**), die Zuordnung zu Variablen (in **logical_variables.csv**) und fragebogenbezogene Metadaten (in **questionnaires.csv**). Eine Zeile entspricht einem Frageitem. Alle Items einer Frage werden durch die Fragenummer in der Spalte **question** zusammengehalten; eine Frage darf nicht unterbrochen werden, und die Reihenfolge der Zeilen ist relevant. Die Spalten **scale** und **answer_list** verweisen auf **scale_types.csv** bzw. **answers.csv**; **concept** verweist auf **concepts.csv**. Der Primärschlüssel ist die Kombination aus **study**, **questionnaire**, **question** und **item**. Zwei Spalten (**template.id**, **old.template.id**) enthalten Punkte im Spaltennamen – eine historisch gewachsene Eigenheit, die in einigen Werkzeugen zu Problemen führt.

answers.csv enthält je eine Zeile pro Antwortoption pro Antwortliste – die Wertelabels für Single-Choice-Fragen. Sie ist über die Kombination aus **study**, **questionnaire** und **answer_list** mit **questions.csv** verbunden. Der Primärschlüssel umfasst zusätzlich **value**.

Ein Beispielfragebogen, anhand dessen die Funktion der einzelnen Spalten erklärt wird, findet sich hier: https://git.soep.de/mds/Enter_DoR_metadata/-/tree/master/questionnaires/soep-meta-2026

questions_images.csv verknüpft Bilder mit Fragen eines Fragebogens, beispielsweise Screenshots oder Ausschnitte aus gedruckten Fragebögen. Sie ist das Gegenstück zur **variables_images.csv** auf der Fragebogen-Seite des Systems. Eine Zeile enthält jeweils die URL eines Bildes (mit separaten Spalten für die deutsch- und englischsprachige Version) zu einer Frage; pro Frage sind mehrere Bilder möglich. Der Primärschlüssel ist die Kombination aus **study**, **instrument**, **question** und **url**. Eine Besonderheit ist der Spaltenname **instrument**: Er entspricht der Spalte **questionnaire** in den übrigen fragebogenbezogenen Tabellen und folgt der Namenskonvention des paneldata.org-Exports. Die Tabelle referenziert Fragen auf Fragenebene (ohne Item-Spalte); ein Fragebild kann daher mehreren Frageitems entsprechen.

Verbindungstabellen

logical_variables.csv ist die Brückentabelle zwischen der Datensatz-Seite und der Fragebogen-Seite des Systems. Sie verknüpft Variablen in **variables.csv** mit den Frageitems in **questions.csv**, aus denen sie entstanden sind. Die Tabelle hat keinen eigenständigen Primärschlüssel. Die Variablen-Seite wird durch **study**, **dataset** und **variable** identifiziert; die Fragen-Seite durch **study**, **questionnaire**, **question** und **item**. Die Tabelle ist versionsagnostisch. Sie erlaubt mehrere Modellierungen: den Standardfall einer 1:1-Zuordnung (in Mermaid-Notation `||:||`) von Frageitem und Variable, den flexiblen Fall eines Items, das nicht nur eine Variable (z. B. in zwei Datensätzen) erzeugt, sowie den Legacy-Fall einer Variablen, die aus zwei verschiedenen Fragebögen gespeist wird.

generations.csv bildet den Variablenfluss bei der Datenaufbereitung ab: Welche Input-Variablen (typischerweise aus Rohdatensätzen) münden in welche Output-Variablen (aufbereitete Datensätze)? Die Tabelle dient auch zur Steuerung der Datenaufbereitung. Alle acht Spalten – vier für die Input-Seite (**input_study**, **input_dataset**, **input_version**, **input_variable**) und vier für die Output-Seite – sind Fremdschlüssel auf **variables.csv**. Die Tabelle hat keinen eigenständigen Primärschlüssel; jede Zeile repräsentiert eine Verknüpfung zwischen zwei

Variablen. Sie kann sich selbst referenzieren: Eine Generation kann als Input einer weiteren Generation fungieren, sodass mehrstufige Ableitungsketten modellierbar sind.

attachments.csv verknüpft URLs – z. B. PDF-Codebooks oder externe Dokumentation – mit Datensätzen, Variablen, Instrumenten oder Studien. Die Tabelle nutzt einen polymorphen Fremdschlüssel: Die Spalte **type** bestimmt exklusiv, welche der vier Zieltabellen (**studies**, **logical_datasets**, **variables**, **questionnaires**) pro Zeile aktiv ist. Für jede Zeile ist damit genau eine der vier Fremdschlüsselgruppen semantisch aktiv; die anderen Spalten sind leer. Dies erlaubt eine einheitliche URL-Verwaltung ohne Tabellenvervielfachung.

dataset_links.csv ist die maschinenlesbare Quelle der Fremdschlüsselbeziehungen zwischen Datensätzen. Sie enthält je eine Zeile pro Beziehung mit Quell- und Zieldatensatz sowie den verbindenden Schlüsselspalten beider Seiten: Der Join erfolgt durch Abgleich der in **source_key** aufgeführten Spalten (leerzeichengetrennt) mit den entsprechenden Spalten in **target_key** (gleiche Reihenfolge). Die Kardinalität wird als Paar aus Mermaid-erDiagram-Endpunktsymbolen (**source_card**, **target_card**) gespeichert, was eine verlustfreie Hin-und-Her-Übersetzung in Mermaid-Syntax ermöglicht. Freitext-Hinweise für Ausnahmefälle – polymorphe Fremdschlüssel, abweichende Spaltennamen – werden in den Beschreibungsspalten abgelegt. Diese Tabelle treibt die automatische Erzeugung des Entity-Relationship-Diagramms und ist Grundlage für die Generierung von Join-Code. Der Primärschlüssel ist die Kombination aus **study**, **source_dataset**, **target_dataset** und **label**.

Tabelle 2: Kardinalitätssymbole in **dataset_links.csv** in der Mermaid-Notation, die auch als Martin-Notation bzw. Krähenfuß-Notation bekannt ist.

Bedeutung	source_card	target_card
genau eine		
null oder eine	o	o
null oder viele	}o	o{
eine oder viele	}	{

Die Kardinalitäten enthalten Informationen in beide Richtungen, also sowohl von der source-Tabelle zur target-Tabelle als auch umgekehrt. Die Verbindung zwischen **questions.csv** und **answers.csv** wird laut **dataset_links.csv** also über die Spalten **study**, **questionnaire** und **answer_list** hergestellt. **o{** in **target_card** bedeutet, dass eine Zeile in **questions.csv** mit null oder vielen Zeilen in **answers.csv** verbunden sein kann. Umgekehrt bedeutet **}]** in **source_card**, dass eine Zeile in **answers.csv** mindestens mit einer oder mehreren Zeilen in **questions.csv** verbunden sein muss.

Themen und Konzepte

topics.csv definiert einen hierarchischen Themenbaum zur Klassifikation von Konzepten, die ihrerseits mit Variablen oder Frageitems verbunden sind. Jedes Topic kann ein übergeordnetes Topic als Elternelement (**parent**) referenzieren – die Tabelle ist damit selbstreferenzierend. Der Primärschlüssel ist die Kombination aus **study** und **name**.

concepts.csv definiert Konzepte, die inhaltlich gleiche Variablen über Wellen und Stichproben hinweg identifizieren. Ein Konzept wird in der **concept**-Spalte von **variables.csv** und **questions.csv** referenziert. Ein Konzept kann mehreren Topics zugeordnet werden; dies wird aktuell provisorisch dadurch gelöst, dass die Konzept-Zeile wiederholt wird. Der Primärschlüssel ist die Kombination aus **study** und **name**.

Kontrollierte Vokabulare

analysis_units.csv definiert das kontrollierte Vokabular für Analyseeinheiten (z. B. Person, Haushalt), das in **logical_datasets.csv** verwendet wird. **conceptual_datasets.csv** definiert das Vokabular für Datensatztypen. **periods.csv** definiert Erhebungszeiträume; im SOEP-Core entspricht dies Erhebungsjahren. **versions.csv** definiert Versionen; im SOEP-Core entsprechen Versionen den Bezeichnungen der veröffentlichten Datenlieferungen (z. B. v39); die Tabelle wird derzeit noch nicht verwendet. Diese vier Tabellen sind studienbezogen und tragen eine **study**-Spalte. Ihr Primärschlüssel ist jeweils die Kombination aus **study** und **name**.

questionnaire_types.csv kategorisiert Erhebungsinstrumente nach Zielgruppe und Erhebungssituation (Haushalts-, Personen-, Jugendfragebogen usw.). **scale_types.csv** beschreibt die Antwortformate von Frageitems (numerische Eingabe, Einfachnennung, Textanzeige, Überschrift usw.). **minedition_types.csv** listet die Editions-Zugangsstufen in aufsteigender Restriktivität. Diese drei Tabellen haben keine **study**-Spalte und gelten studienübergreifend. Ihr Primärschlüssel ist jeweils ein einzelner Bezeichner (**qtype**, **scale** bzw. **value**).

3 Beziehungen zwischen den Tabellen

Hierarchie der Datensatz-Seite des Modells

studies.csv ist der Ausgangspunkt. Von dort führt ein Fremdschlüssel über die **study**-Spalte zu allen studienbezogenen Tabellen – darunter **logical_datasets.csv** als erste Ebene der Datensatz-Hierarchie.

logical_datasets.csv referenziert drei studienbezogene kontrollierte Vokabulare: die Analyseeinheit (**analysis_unit** → **analysis_units.name**), den Datensatztyp (**conceptual_dataset** → **conceptual_datasets.name**) und den Erhebungszeitraum (**period** → **periods.name**). Die Zugangsstufe (**minedition**) verweist auf **minedition_types.value**.

datasets_versions.csv verbindet einen Datensatz (aus **logical_datasets.csv**) mit einer Versionskennung (aus **versions.csv**). Damit ist bekannt, in welchen Versionen ein Datensatz enthalten ist.

variables.csv hängt über **study + dataset** an **logical_datasets.csv**. Ihre eigene **concept**-Spalte referenziert **concepts.name**; **minedition** referenziert **minedition_types.value**. Von **variables.csv** gehen drei weitere Verbindungen aus: nach **variable_categories.csv** (über den vollständigen Primärschlüssel **study + dataset + version + variable**), nach **codebook.csv** (auf zwei Ebenen – versionsweise und variablenweise, siehe unten) und nach **variables_images.csv** (versionsunabhängig über **study + dataset + variable**).

Die Verbindung von **variables.csv** zu **codebook.csv** hat eine Besonderheit: Sie existiert auf zwei Ebenen. Auf Versionsebene gilt eine **||:o{**-Beziehung – mehrere Variablen können zu einer Codebook-Version gehören, und Codebook-Zeilen ohne **variable**-Eintrag (reine Abschnittsköpfe) fallen nur unter diese Kante. Auf Variablenebene gilt eine **||:o|**-Beziehung – jede Variable erscheint in höchstens einer Codebook-Zeile.

Hierarchie der Fragebogen-Seite des Modells

`questionnaires.csv` hängt über `study` an `studies.csv`. Die Spalten `qtyp` und `period` referenzieren `questionnaire_types.qtype` bzw. `periods.name`. Unterhalb der Fragebogen-Ebene liegen `questions.csv` (über `study + questionnaire`) und `answers.csv` (über `study + questionnaire + answer_list`). `questions.csv` referenziert zusätzlich `scale_types.scale`, `concepts.name` (über `concept`) und `answers.csv` (über `answer_list`).

`questions_images.csv` ist auf zwei Ebenen angebunden: an `questionnaires.csv` über `study + questionnaire/instrument` und an `questions.csv` über `study + questionnaire + question` – also auf Fragenebene, ohne Item-Spalte. In beiden Fällen weicht der Spaltenname ab (`questionnaire` → `instrument`). Pro Frage sind mehrere Bilder möglich.

Verbindung beider Seiten

Die Brücke zwischen Datensatz- und Fragebogen-Seite ist `logical_variables.csv`. Eine Zeile in dieser Tabelle verknüpft eine Variable (`study + dataset + variable`) mit einem Frageitem (`study + questionnaire + question + item`). Damit ist für jede Variable nachvollziehbar, aus welchem Frageitem sie entstanden ist – und umgekehrt, welche Variablen ein Frageitem erzeugt hat.

`concepts.csv` stellt eine zweite, abstraktere Verbindung dar: Sowohl `variables.csv` als auch `questions.csv` referenzieren `concepts.name`. Konzepte erlauben es, Variable und Frageitem direkt zu verbinden, ohne den Umweg über `logical_variables.csv` nehmen zu müssen – etwa um thematisch verwandte Variablen verschiedener Wellen als Gruppe zu identifizieren.

Variablenfluss

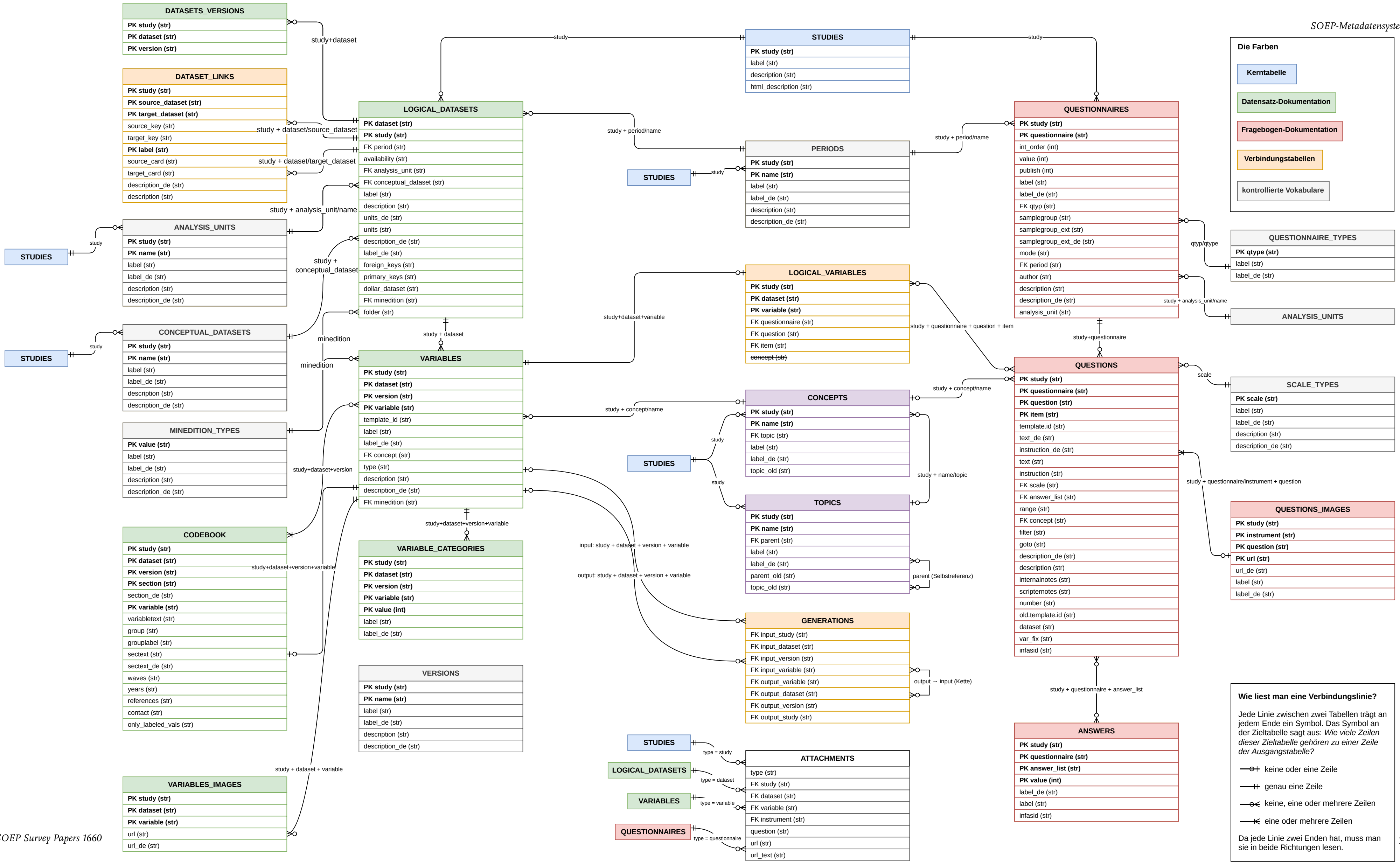
`generations.csv` ergänzt die Verbindung von Variablen über Datenaufbereitungsschritte hinweg. Jede Zeile beschreibt eine Ableitung: Eine Input-Variable (typischerweise im Rohdatensatz) führt zu einer Output-Variable (typischerweise im aufbereiteten Datensatz). Da der Prozess mehrstufig sein kann – ein Output kann seinerseits Input einer weiteren Ableitung sein – kann `generations.csv` sich selbst referenzieren. So lassen sich vollständige Ableitungsketten rekonstruieren.

Fremdschlüsselbeziehungen zwischen Datensätzen

Die Fremdschlüsselbeziehungen zwischen den durch das System dokumentierten Datensätzen (z. B. die Verbindung von `ppathl` zu `pl`) sind in `dataset_links.csv` dokumentiert. Jede Zeile beschreibt eine gerichtete Beziehung: Quell- und Zieldatensatz, die verbindenden Schlüsselspalten beider Seiten sowie die Kardinalität. Die Tabelle macht diese Struktur maschinenlesbar und erlaubt die automatische Generierung von Join-Code und Entity-Relationship-Diagrammen.

Anhänge

`attachments.csv` kann von vier verschiedenen Tabellen aus referenziert werden: von `studies.csv`, `logical_datasets.csv`, `variables.csv` und `questionnaires.csv`. Die Spalte `type` bestimmt, welche dieser vier Verbindungen für eine gegebene Zeile aktiv ist.



5 Technische Details

Grundprinzip: ein git-versioniertes CSV-Repository

Das Metadatensystem besteht aus einem einzigen git-kontrollierten Repository von CSV-Dateien, das als einzige Quelle der Wahrheit (*single source of truth*) fungiert. Alle nachgelagerten Repräsentationen – die paneldata.org-Website, PDF-Dokumente, der jährliche DDI-Codebook-XML-Export und zukünftige DDI-Lifecycle-4-Exporte – werden aus diesem Repository abgeleitet.

Die Wahl des CSV-Formats ist bewusst: CSV-Dateien sind werkzeugagnostisch, versionierbar mit Standard-Git-Werkzeugen und für Mitarbeitende ohne Datenbankwissen direkt editierbar. Eine Datenbank würde die Einstiegshürde erhöhen und die Versionierung erschweren; proprietäre Formate würden die Langzeitverfügbarkeit gefährden.

Ablagestruktur

Die CSV-Dateien folgen einer einheitlichen Verzeichnisstruktur innerhalb eines Repositories. Studienweite Tabellen – darunter auch die kontrollierten Vokabulare und `variables_images.csv` – liegen im Verzeichnis `meta/` (z. B. `meta/logical_datasets.csv`, `meta/questionnaires.csv`). Datensatz-spezifische Tabellen liegen unter `datasets/{datensatz}/{version}/` (z. B. `datasets/pl/v36/variables.csv`). Fragebogen-spezifische Tabellen – `questions.csv`, `answers.csv`, `logical_variables.csv` und `questions_images.csv` – liegen unter `questionnaires/{fragebogen}/` (z. B. `questionnaires/ph2020/questions.csv`). Diese Struktur erlaubt es, Metadaten eines einzelnen Datensatzes oder Fragebogens als geschlossene Einheit zu behandeln und im Falle der Datensätze unabhängig voneinander zu versionieren.

Für `generations.csv` gibt es keine strikt einheitliche Ablagekonvention: Die Datei soll möglichst dort liegen, wo der Output-Teil der Zeile hinweist – also unter `datasets/{datensatz}/{version}/generations.csv`. Studienweite Generations-Dateien, die mehrere Datensätze überbrücken, werden separat unter `meta/{kategorie}_generations/` abgelegt.

Die soep-meta-Konvention

Das Metadatensystem ist selbstbeschreibend: Es dokumentiert sich mit denselben Tabellen, mit denen es die eigentlichen Daten einer Studie beschreibt. Jede Spalte jeder Metadaten-Tabelle ist als Zeile in `variables.csv` eingetragen – mit `study = "soep-meta"` statt etwa `study = "soep-core"`. Analog registriert `logical_datasets.csv` die 24 Metadaten-Tabellen selbst als Datensätze mit `study = "soep-meta"`, während sie im Produktivsystem alle Datensätze einer Studie verzeichnet.

Diese Konvention hat mehrere Vorteile:

- **Keine zusätzliche Infrastruktur:** Das Schemawissen liegt in denselben CSV-Dateien wie die Objektmetadaten; kein separates Schema-Format ist erforderlich.
- **Maschinenlesbarkeit:** Das Repository `soep-meta` liefert die vollständige Systemdokumentation im selben Format wie die Nutzdaten.
- **Konsistenzprüfbarkeit:** Die Dokumentation kann gegen die tatsächlichen Tabellen validiert werden.

- **Wiederverwendbarkeit:** Werkzeuge, die das Metadatensystem lesen, können dieselben Werkzeuge nutzen, um das System selbst zu verstehen.

Die Ablagestruktur wird maschinenlesbar in [file_locations.csv](#) vorgehalten. Sie ist nicht Bestandteil der produktiv genutzten Metadaten-Repositories.

Das ERD im Mermaid-Format ([ERD.md](#)) kann algorithmisch erzeugt werden. [ERD.drawio](#) ist eine manuell bearbeitete Variante für Publikationszwecke. (Vgl. Abschnitt 4)

Im Repository <https://git.soep.de/mds/soep-meta> befindet sich ebenfalls eine README.md-Datei, die insbesondere geeignet ist, die Informationen über die Spezifikation der CSV-Dateien und Wissen über die Struktur der Metadaten auch für agentische Nutzung zur Verfügung zu stellen.

Export nach paneldata.org

Die meisten Tabellen des Systems werden nach paneldata.org exportiert; die technische Spezifikation des Exports ist je Tabelle in einem eigenen Repository dokumentiert (<https://github.com/paneldata/data-specification>). Beim Export finden mehrere Umbenennungen statt, die die interne Terminologie in die paneldata.org-Terminologie überführen:

- [logical_datasets.csv](#) wird als [datasets.csv](#) exportiert; die Spalte [primary_keys](#) wird zu [primary_key](#).
- [questionnaires.csv](#) wird als [instruments.csv](#) exportiert; die Spalte [questionnaire](#) wird zu [name](#).
- In [questions.csv](#), [answers.csv](#) und [logical_variables.csv](#) wird die Spalte [questionnaire](#) in [instrument](#) umbenannt; in [questions.csv](#) wird zusätzlich [question](#) in [name](#) umbenannt.
- In [variables.csv](#) wird die Spalte [variable](#) in [name](#) umbenannt.

Der Export ist gefiltert: Datensätze werden nur exportiert, wenn ihre [minedition](#) eine exportierbare Zugangsstufe hat und sie in der zu exportierenden Version enthalten sind ([datasets_versions.csv](#)); Instrumente nur bei entsprechendem [publish](#)-Status. Einige Tabellen werden nicht exportiert und dienen anderen Zwecken: [studies.csv](#), [codebook.csv](#), [dataset_links.csv](#) und [scale_types.csv](#).

Bezug zu Metadatenstandards

Das Metadatensystem treibt den jährlichen DDI-Codebook-XML-Export (~500 MB). DDI-Codebook deckt die Variablenebene gut ab – Bezeichnungen, Wertelabels, fehlende Werte, Analyseeinheit – kann aber strukturelle Konzepte wie Primärschlüssel, Fremdschlüssel sowie Beziehungen zwischen Datensätzen nicht ausdrücken und keine Fragebögen abbilden.

Ein Export nach DDI-Lifecycle 4 ist in Planung. DDI-Lifecycle 4 unterstützt über [CaselIdentification](#) Primärschlüssel und über [RecordRelationshipType](#) Datensatzverbindungen. Die in [dataset_links.csv](#) maschinenlesbar vorliegenden Fremdschlüsselbeziehungen sind als Grundlage dieses Exports vorgesehen. Konzepte wie Persontypsteuerung lassen sich auch in DDI-Lifecycle 4 nicht formal ausdrücken und werden absehbar auch weiterhin als ergänzende CSV-Konvention gepflegt.

Anmerkung zur Konsistenz

Dieses System ist schrittweise gemeinsam mit den Daten gewachsen, die es beschreibt. Daher werden nicht alle Konventionen in allen Tabellen einheitlich angewendet, und eine strikte technische Validierung der Metadaten anhand des hier dokumentierten Schemas würde vermutlich Inkonsistenzen aufdecken. Soweit bekannte Abweichungen bestehen, sind diese in der jeweiligen codebook.csv dokumentiert. Die Schemadokumentation sollte daher als Beschreibung der beabsichtigten Struktur und der zugrunde liegenden Konventionen verstanden werden und nicht als formal durchgesetzter Vertrag.

Auch die Beschreibung des Systems ist ein dynamischer Prozess, sie wird unter <https://git.soep.de/mds/soep-meta> laufend aktualisiert.

Bekannte Eigenheiten und Designentscheidungen

Die Tabelle [logical_datasets.csv](#) trägt ihren Namen aus einer ursprünglich geplanten, später aufgegebenen Unterscheidung zwischen logischen und physischen Eigenschaften eines Datensatzes, wie sie etwa in DDI-Lifecycle angelegt ist. Eine Umbenennung in [datasets.csv](#) ist beabsichtigt, aber noch nicht umgesetzt; der paneldata.org-Export verwendet den Namen [datasets](#) bereits.

Die Editions-Zugangsstufen in [minedition_types.csv](#) sind aufsteigend geordnet von der niedrigsten zur höchsten Zugangsbeschränkung: [teaching](#) → [international](#) → [eu](#) → [area_types](#) → [planning_regions](#) → [remote](#) → [onsite](#) → [internal](#) → [arch](#). Die [minedition](#) einer Information bezeichnet die niedrigste dieser Stufen, auf der die Information noch verfügbar ist: Sie ist in dieser Edition und in allen restriktiveren Editionen enthalten, in offeneren dagegen nicht. Grundlage für die deskriptiven Statistiken auf paneldata.org ist die [eu](#)-Edition.

Der polymorphe Fremdschlüssel in [attachments.csv](#) ist in keiner gängigen relationalen Formalsprache direkt ausdrückbar. Er ist bewusst gewählt, um die Tabelle ohne Schemaduplizierung universell verwendbar zu halten.

Insbesondere [logical_variables.csv](#) und [variables_images.csv](#) sind versionsagnostisch: Sie referenzieren Variablen ohne Versionsangabe, womit die Verbindungen als über Versionen stabil angesehen werden. Für beide Tabellen ist die Ergänzung einer [version](#)-Spalte in einer künftigen Revision vorgesehen.

Die Spalten [template.id](#) und [old.template.id](#) in [questions.csv](#) enthalten Punkte, die in einigen Werkzeugen (darunter Mermaid-Diagramme) als Syntaxfehler behandelt werden. Der Sachverhalt ist als bekanntes Problem dokumentiert.

Die Spalte [instrument](#) in [questions_images.csv](#) weicht von der Spalte [questionnaire](#) der übrigen fragebogenbezogenen Tabellen ab; die Tabelle folgt hier bereits der Namenskonvention des paneldata.org-Exports, in dem [questionnaire](#) durchgängig zu [instrument](#) umbenannt wird.

Für [questionnaire_types.csv](#) und [minedition_types.csv](#) fehlt derzeit eine [study](#)-Spalte, obwohl die Werte SOEP-spezifisch sind. Eine Ergänzung ist für eine künftige Revision vorgesehen. [scale_types.csv](#) ist dagegen bewusst studienübergreifend modelliert, da die Skalentypen generische technische Formate beschreiben.

Die Tabelle [versions.csv](#) ist als kontrolliertes Vokabular für Versionsbezeichner angelegt, wird derzeit aber noch nicht verwendet.

6 Zusammenfassung und Ausblick

Das SOEP-Metadatensystem verbindet einfache Werkzeuge (CSV, Git) mit dem Konzept der Selbstdokumentation zu einer wartbaren, erweiterungsfähigen Infrastruktur. Die soep-meta-Konvention macht das System für Maschinen und Menschen gleichermaßen lesbar, ohne eine separate Schemasprache einzuführen. Die 24 Tabellen in sieben Gruppen decken alle wesentlichen Aspekte der Datensatz- und Fragebogen-Dokumentation ab; die Verbindungstabellen stellen die strukturellen Brücken zwischen diesen Bereichen her.

Aktuell in Diskussion sind ein DDI-Lifecycle-4-Export-Skript sowie ein MCP-Server (*Model Context Protocol*), der das Metadatensystem für KI-gestützte Analysewerkzeuge erschließt. Beide bauen auf den bereits maschinenlesbaren Strukturen – insbesondere [dataset_links.csv](#) und der soep-meta-Konvention – auf und setzen keine weiteren Schemaänderungen voraus.

7 Danksagung und Hinweise

Ich danke Xiaoyao Han, Jana Nebelin und Claudia Saalbach für Beiträge zu den Metainformationen des Systems, Durchsicht des Beitrags und zusätzliche Hinweise. Bei der Erstellung des selbst dokumentierenden Metadatensystems auf Grundlage der vorhandenen SOEP-Metadaten und erster Textentwürfe war Claude AI eine große Hilfe. Alle verbleibenden Fehler liegen ausschließlich in der Verantwortung des Autors.

8 Anhang: Spaltenreferenz

Die folgenden Tabellen listen alle Spalten jeder Metadaten-Tabelle mit Datentyp, Primärschlüsselrolle (PK) und Kurzbeschreibung.

Kerntabelle

studies.csv: Studien

Spalte	Typ	Rolle	Beschreibung
<i>study</i>	str	PK	Die Spalte study enthält die ID der Studie und bildet den Primärschlüssel der Tabelle studies.csv.
<i>label</i>	str		Die Spalte label enthält einen englischen Anzeigenamen der Studie.
<i>description</i>	str		VERALTET: Dieses Feld wird nirgends verwendet.
<i>html_description</i>	str		VERALTET: Dieses Feld für eine Beschreibung im HTML-Format wird nirgends verwendet.

Datensatz-Dokumentation

logical_datasets.csv: Datensätze

Spalte	Typ	Rolle	Beschreibung
dataset	str	PK	Die Spalte dataset enthält den Namen des Datensatzes und bildet zusammen mit der Spalte study den Primärschlüssel der Tabelle logical_datasets.csv.
study	str	PK	Die Spalte study enthält mit der ID den Primärschlüssel für die Studie und wird in der Tabelle studies.csv eingeführt.
period	str		Kontrolliertes Vokabular aus periods.csv.
availability	str		VERALTET: Siehe minedition.
analysis_unit	str		Die Spalte analysis_unit enthält die Analyseeinheit. Kontrolliertes Vokabular aus analysis_units.csv.
conceptual_dataset	str		Die Spalte conceptual_dataset enthält den Datensatztyp. Kontrolliertes Vokabular aus conceptual_datasets.csv.
label	str		Freitext, englisch. Ein Anzeigename des Datensatzinhalts.
description	str		Freitext, englisch. Eine Kurzbeschreibung des Datensatzes.
units_de	str		Freitext, deutsch. Eine Verbalisierung des Primärschlüssels und weiterer inhaltlicher Eingrenzungen der Analyseeinheit.
units	str		Freitext, englisch. Eine Verbalisierung des Primärschlüssels und weiterer inhaltlicher Eingrenzungen der Analyseeinheit.
description_de	str		Freitext, deutsch. Eine Kurzbeschreibung des Datensatzes.
label_de	str		Freitext, deutsch. Ein Anzeigename des Datensatzinhalts.
foreign_keys	str		Die Spalte foreign_keys listet Spalten auf, die Fremdschlüssel sind – d.h. Primärschlüssel in anderen Datensätzen.
primary_keys	str		Die Spalte primary_keys listet die Spalten auf, die jede Zeile eindeutig identifizieren.
dollar_dataset	str		Die Spalte dollar_dataset enthält den Stammnamen des Datensatzes.
minedition	str		Die Spalte minedition enthält die Edition mit den geringsten Zugangsbeschränkungen, in der dieser Datensatz enthalten ist.
folder	str		Die Spalte folder steuert, in welchem Ordner der Datensatz abgelegt wird.

variables.csv: Variablen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
dataset	str	PK	Die Spalte dataset enthält den Namen des Datensatzes.
version	str	PK	Die Spalte version enthält die Versionsnummer des Datensatzes.
variable	str	PK	Die Spalte variable enthält den Namen der Variable.
template_id	str		Die Spalte template_id enthält die ID des Variablen-templates, das mit dieser Variable verknüpft ist.
label	str		Die Spalte label enthält das englische Variablenlabel bzw. den Anzeigenamen.
label_de	str		Die Spalte label_de enthält das deutsche Variablen-label bzw. den Anzeigenamen.
concept	str		Die Spalte concept enthält die ID des Konzepts.
type	str		Die Spalte type enthält den Datentyp der Variable.
description	str		Die Spalte description enthält eine englische Beschreibung der Variable.
description_de	str		Die Spalte description_de enthält eine deutsche Beschreibung der Variable.
minedition	str		Die Spalte minedition gibt an, in welcher Edition mit der höchsten Zugangsbeschränkung diese Variable noch verfügbar ist.

variable_categories.csv: Variablenkategorien

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
dataset	str	PK	Die Spalte dataset enthält den Namen des Datensatzes.
version	str	PK	Die Spalte version enthält die Versionsnummer des Datensatzes.
variable	str	PK	Die Spalte variable enthält den Namen der Variable.
value	int	PK	Die Spalte value enthält den numerischen Wert, der mit einem Label versehen werden soll.
label	str		Die Spalte label enthält das englische Wertelabel, d.h. den Anzeigenamen der Antwortkategorie in englischer Sprache.
label_de	str		Die Spalte label_de enthält das deutsche Wertelabel, d.h. den Anzeigenamen der Antwortkategorie in deutscher Sprache.

datasets_versions.csv: Datensätze in Versionen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
dataset	str	PK	Die Spalte dataset enthält den Namen des Datensatzes und bildet zusammen mit study und version den Primärschlüssel der Tabelle datasets_versions.csv.
version	str	PK	Die Spalte version enthält die Versionsnummer und bildet zusammen mit study und dataset den Primärschlüssel der Tabelle datasets_versions.csv.

codebook.csv: Codebook

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
dataset	str	PK	Die Spalte dataset enthält den Namen des Datensatzes und wird in der tabelle logical_datasets.csv eingeführt.
version	str	PK	Die Spalte version enthält die Versionsnummer des Datensatzes.
section	str	PK	Die Spalte section enthält den englischen Titel des Abschnitts, in dem die Variable oder der Abschnittstext erscheint.
section_de	str		Die Spalte section_de enthält den deutschen Titel des Abschnitts.
variable	str	PK	Die Spalte variable enthält den Namen der Variable.
variabletext	str		VERALTET: Die Spalte variabletext enthält einen erläuternden Text, der im Codebook neben oder unter der Variable angezeigt wird.
group	str		VERALTET: Die Spalte group enthält eine ID für eine Untergruppe innerhalb eines Abschnitts.
grouplabel	str		VERALTET: Die Spalte grouplabel enthält die Bezeichnung der Untergruppe.
sectext	str		Die Spalte sectext enthält einen einleitenden englischen Text für den Abschnitt.
sectext_de	str		Die Spalte sectext_de enthält einen einleitenden deutschen Text für den Abschnitt.
waves	str		Die Spalte waves enthält die Wellenbezeichnungen, in denen die Variable erhoben wurde.
years	str		Die Spalte years enthält die Erhebungsjahre, in denen die Variable verfügbar ist.
references	str		Die Spalte references enthält Quellenangaben oder weiterführende Hinweise zur Variable.
contact	str		Die Spalte contact enthält eine Kontaktperson oder -stelle, die für Rückfragen zur Variable zuständig ist.
only_labeled_vals	str		Die Spalte only_labeled_vals steuert, ob in einer Häufigkeitstabelle im Codebook nur Werte angezeigt werden, die ein Wertelabel haben, oder alle Werte.

variables_images.csv: Abbildungen für Variablen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
dataset	str	PK	Die Spalte dataset enthält den Namen des Datensatzes und bildet zusammen mit study und variable den Primärschlüssel der Tabelle variables_images.csv.
variable	str	PK	Die Spalte variable enthält den Namen der Variable und bildet zusammen mit study und dataset den Primärschlüssel der Tabelle variables_images.csv.
url	str		Die Spalte url enthält die URL eines Bildes, das mit der Variable assoziiert ist (englische Version).
url_de	str		Die Spalte url_de enthält die URL eines Bildes, das mit der Variable assoziiert ist (deutsche Version).

Fragebogen-Dokumentation

questionnaires.csv: Fragebögen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
questionnaire	str	PK	Die Spalte questionnaire enthält die ID des Fragebogens und bildet zusammen mit der Spalte study den Primärschlüssel der Tabelle questionnaires.csv.
int_order	int		–
value	int		Die Spalte value enthält eine numerische ID des Fragebogens.
publish	int		Die Spalte publish steuert den Export des Instruments nach paneldata.org.
label	str		Die Spalte label enthält einen englischen Anzeigenamen des Fragebogens.
label_de	str		Die Spalte label_de enthält einen deutschen Anzeigenamen des Fragebogens.
qtyp	str		Die Spalte qtyp enthält ein Kürzel für den Typ des Fragebogens.
samplegroup	str		–
samplegroup_ext	str		–
samplegroup_ext_de	str		–
mode	str		Die Spalte mode enthält den Erhebungsmodus.
period	str		Die Spalte period enthält den Erhebungszeitraum des Fragebogens.
author	str		Die Spalte author enthält die Angabe der Autor*innen.
description	str		Die Spalte description enthält die englische Beschreibung des Fragebogens.
description_de	str		Die Spalte description_de enthält die deutsche Beschreibung des Fragebogens.
analysis_unit	str		Die Spalte analysis_unit enthält die Analyseeinheit. Kontrolliertes Vokabular aus analysis_units.csv.

questions.csv: Fragen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
questionnaire	str	PK	Die Spalte questionnaire enthält die ID des Fragebogens und wird in der Tabelle questionnaires.csv eingeführt.
question	str	PK	Die Spalte question enthält die Fragenummer, die nicht leer sein darf und keine Leerzeichen enthalten sollte.
item	str	PK	Die Spalte item enthält eine ID des Items innerhalb der Frage.
template.id	str		Die Spalte template.id enthält die ID des Fragetemplates.
text_de	str		Die Spalte text_de enthält den deutschen Fragetext.
instruction_de	str		Die Spalte instruction_de enthält die deutsche Intervieweranweisung.
text	str		Die Spalte text enthält den englischen Fragetext.
instruction	str		Die Spalte instruction enthält die englische Intervieweranweisung.
scale	str		Die Spalte scale enthält die Skalierung des Frageitems.
answer_list	str		Die Spalte answer_list enthält die ID der Antwortliste.
range	str		Die Spalte range enthält den Wertebereich der bei scale=int erlaubten Antworten.
concept	str		Die Spalte concept enthält die ID des Schlagworts.
filter	str		Die Spalte filter enthält den Eingangsfiler – eine Bedingung, die steuert, ob die Frage vorgelegt wird.
goto	str		Die Spalte goto enthält den Ausgangsfiler – eine Anweisung, welche Frage auf eine Frage folgt, i.d.R. in Abhängigkeit von der gegebenen Antwort.
description_de	str		Die Spalte description_de enthält eine deutsche Beschreibung.
description	str		Die Spalte description enthält eine englische Beschreibung.
internalnotes	str		Die Spalte internalnotes enthält eine interne Notiz.
scripnotes	str		Die Spalte scripnotes enthält Programmierhinweise vom SOEP an infas.
number	str		Die Spalte number enthält eine zusätzliche Fragenummer, oft die alte Fragenummer.
old.template.id	str		Die Spalte old.template.id enthält die ID des Fragetemplates, auf dem diese Frage basiert.
dataset	str		Die Spalte dataset enthält einen Datensatznamen als Vorschlag.
var_fix	str		Die Spalte var_fix enthält einen Variablennamen als Vorschlag.
infasid	str		Die Spalte infasid enthält eine von infas vergebene ID.

answers.csv: Antworten

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
questionnaire	str	PK	Die Spalte questionnaire enthält die ID des Fragebogens und wird in der Tabelle questionnaires.csv eingeführt.
answer_list	str	PK	Die Spalte answer_list enthält die ID der Antwortliste.
value	int	PK	Die Spalte value enthält den numerischen Wert einer Antwortoption.
label_de	str		Die Spalte label_de enthält die deutsche Antwortmöglichkeit, so wie sie der Befragungsperson präsentiert wurde.
label	str		Die Spalte label enthält die englische Übersetzung der Antwortmöglichkeit.
infasid	str		Die Spalte infasid enthält eine von infas vergebene ID.

questions_images.csv: Abbildungen für Fragen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
instrument	str	PK	Die Spalte instrument identifiziert den Fragebogen, zu dem das Bild gehört.
question	str	PK	Die Spalte question enthält die Fragennummer oder den Bezeichner aus der questions.csv.
url	str	PK	Die Spalte url enthält die URL eines Bildes, das mit der Frage assoziiert ist (englische Version).
url_de	str		Die Spalte url_de enthält die URL eines Bildes, das mit der Frage assoziiert ist (deutsche Version).
label	str		Die Spalte label enthält einen kurzen englischen Anzeigenamen des Bildes (z. B. Screenshot).
label_de	str		Die Spalte label_de enthält einen kurzen deutschen Anzeigenamen des Bildes (z. B. Screenshot).

Verbindungstabellen

logical_variables.csv: Verbindung Frageitem–Variable

Spalte	Typ	Rolle	Beschreibung
study	str		Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
dataset	str		Die Spalte dataset enthält den Namen des Datensatzes und wird in der Tabelle logical_datasets.csv eingeführt.
variable	str		Die Spalte variable enthält den Namen der Variable und wird in der Tabelle variables.csv eingeführt.
questionnaire	str		Die Spalte questionnaire enthält die ID des Fragebogens, dessen Frageitems mit der Variable verknüpft werden sollen.
question	str		Die Spalte question enthält die Fragenummer des Frageitems, das mit der Variable verknüpft werden soll.
item	str		Die Spalte item enthält die Item-ID des Frageitems, das mit der Variable verknüpft werden soll.
concept	str		VERALTET: Diese Spalte wird nicht genutzt.

generations.csv: Verbindung zwischen Variablen

Spalte	Typ	Rolle	Beschreibung
input_study	str		Die Spalte input_study enthält die ID der Input-Studie.
input_dataset	str		Die Spalte input_dataset enthält den Namen des Input-Datensatzes.
input_version	str		Die Spalte input_version enthält die Distributionsversion der Input-Variablen.
input_variable	str		Die Spalte input_variable enthält den Namen der Input-Variable.
output_variable	str		Die Spalte output_variable enthält den Namen der Output-Variable.
output_dataset	str		Die Spalte output_dataset enthält den Namen des Output-Datensatzes.
output_version	str		Die Spalte output_version enthält die Distributionsversion der Output-Variablen.
output_study	str		Die Spalte output_study enthält die ID der Output-Studie.

attachments.csv: Anhänge

Spalte	Typ	Rolle	Beschreibung
type	str	PK	Die Spalte type gibt an, auf welchen Objekttyp sich der Anhang bezieht.
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
dataset	str	PK	Die Spalte dataset enthält den Namen des Datensatzes.
variable	str	PK	Die Spalte variable enthält den Namen der Variable.
instrument	str	PK	Die Spalte instrument enthält die ID des Fragebogens.
question	str	PK	Die Spalte question enthält die Fragenummer.
url	str	PK	Die Spalte url enthält die URL des Anhangs.
url_text	str		Die Spalte url_text enthält den anzuzeigenden Text des Links.

dataset_links.csv: Verbindung zwischen Datensätzen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
source_dataset	str	PK	Die Spalte source_dataset identifiziert die Quellseite ("Eine"-Seite oder linke Seite) der Beziehung.
target_dataset	str	PK	Die Spalte target_dataset identifiziert die Zielseite ("Viele"-Seite oder rechte Seite) der Beziehung.
source_key	str		Die Spalte source_key enthält eine leerzeichenge trennte Liste von Spaltennamen aus source_dataset, die den Join-Schlüssel auf der Quellseite bilden.
target_key	str		Die Spalte target_key enthält eine leerzeichenge trennte Liste von Spaltennamen aus target_dataset, die den Join-Schlüssel auf der Zielseite bilden.
label	str	PK	Die Spalte label enthält einen kurzen (sprachneutralen) Anzeigenamen der Beziehung, die als Kantenbezeichnung in ERD-Diagrammen verwendet wird.
source_card	str		Die Spalte source_card enthält das Mermaid-erDiagram-Endpunktsymbol für die Quellseite der Beziehung.
target_card	str		Die Spalte target_card enthält das Mermaid-erDiagram-Endpunktsymbol für die Zielseite der Beziehung.
description_de	str		Die Spalte description_de enthält einen deutschen Freitext-Hinweis für Ausnahmefälle, die durch die anderen Spalten nicht vollständig abgebildet werden.
description	str		Die Spalte description enthält einen englischen Freitext-Hinweis für Ausnahmefälle, die durch die anderen Spalten nicht vollständig abgebildet werden.

Themen und Konzepte

topics.csv: Topics

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
name	str	PK	Die Spalte name enthält die ID des Themas.
parent	str		Die Spalte parent enthält den Bezeichner des übergeordneten Themas.
label	str		Die Spalte label enthält den englischen Anzeigenamen des Themas.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen des Themas.
parent_old	str		VERALTET: Die Spalte parent_old enthält die ID des übergeordneten Themas im früheren Themensystem.
topic_old	str		VERALTET: Die Spalte topic_old enthält die ID des Themas im früheren Themensystem.

concepts.csv: Konzepte

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
name	str	PK	Die Spalte name enthält die ID des Konzepts und bildet zusammen mit study den Primärschlüssel der Tabelle concepts.csv.
topic	str		Die Spalte topic enthält den Bezeichner des Themas, dem das Konzept zugeordnet ist.
label	str		Die Spalte label enthält den englischen Anzeigenamen des Konzepts.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen des Konzepts.
topic_old	str		VERALTET: Die Spalte topic_old enthält den Bezeichner des Themas im früheren Themensystem.

Kontrollierte Vokabulare

analysis_units.csv: Analyseeinheiten

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
name	str	PK	Die Spalte name enthält das Kürzel der Analyseeinheit und bildet zusammen mit study den Primärschlüssel der Tabelle analysis_units.csv.
label	str		Die Spalte label enthält den englischen Anzeigenamen der Analyseeinheit.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen der Analyseeinheit.
description	str		Die Spalte description enthält die englische Beschreibung der Analyseeinheit.
description_de	str		Die Spalte description_de enthält die deutsche Beschreibung der Analyseeinheit.

conceptual_datasets.csv: Datensatztypen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
name	str	PK	Die Spalte name enthält die ID des konzeptuellen Datensatztyps und bildet zusammen mit study den Primärschlüssel der Tabelle conceptual_datasets.csv.
label	str		Die Spalte label enthält den englischen Anzeigenamen des konzeptuellen Datensatztyps.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen des konzeptuellen Datensatztyps.
description	str		Die Spalte description enthält eine englische Beschreibung des konzeptuellen Datensatztyps.
description_de	str		Die Spalte description_de enthält eine deutsche Beschreibung des konzeptuellen Datensatztyps.

periods.csv: Zeiträume

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
name	str	PK	Die Spalte name enthält die ID des Erhebungszeitraums und bildet zusammen mit study den Primärschlüssel der Tabelle periods.csv.
label	str		Die Spalte label enthält den englischen Anzeigenamen des Erhebungszeitraums.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen des Erhebungszeitraums.
description	str		Die Spalte description enthält eine englische Beschreibung des Erhebungszeitraums.
description_de	str		Die Spalte description_de enthält eine deutsche Beschreibung des Erhebungszeitraums.

versions.csv: Versionen

Spalte	Typ	Rolle	Beschreibung
study	str	PK	Die Spalte study enthält die ID der Studie und wird in der Tabelle studies.csv eingeführt.
name	str	PK	Die Spalte name enthält die ID der Versions/des Releases und bildet zusammen mit study den Primärschlüssel der Tabelle versions.csv.
label	str		Die Spalte label enthält den englischen Anzeigenamen der Version/des Releases.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen der Version/des Releases.
description	str		Die Spalte description enthält eine englische Beschreibung der Version/des Releases.
description_de	str		Die Spalte description_de enthält eine deutsche Beschreibung der Version/des Releases.

questionnaire_types.csv: Fragebogentypen

Spalte	Typ	Rolle	Beschreibung
qtype	str	PK	Die Spalte qtype enthält die ID/das Kürzel des Fragebogentyps und bildet den Primärschlüssel der Tabelle questionnaire_types.csv.
label	str		Die Spalte label enthält den englischen Anzeigenamen des Fragebogentyps.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen des Fragebogentyps.

scale_types.csv: Skalentypen

Spalte	Typ	Rolle	Beschreibung
scale	str	PK	Die Spalte scale enthält die ID/das Kürzel des Skalentyps und bildet den Primärschlüssel der Tabelle scale_types.csv.
label	str		Die Spalte label enthält den englischen Anzeigenamen des Skalentyps.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen des Skalentyps.
description	str		Die Spalte description enthält eine englische Beschreibung des Skalentyps.
description_de	str		Die Spalte description_de enthält eine deutsche Beschreibung des Skalentyps.

minedition_types.csv: Zugangsstufentypen

Spalte	Typ	Rolle	Beschreibung
value	str	PK	Die Spalte value enthält die ID der Edition und bildet den Primärschlüssel der Tabelle minedition_types.csv.
label	str		Die Spalte label enthält den englischen Anzeigenamen der Edition.
label_de	str		Die Spalte label_de enthält den deutschen Anzeigenamen der Edition.
description	str		Die Spalte description enthält eine englische Beschreibung der Edition.
description_de	str		Die Spalte description_de enthält eine deutsche Beschreibung der Edition und gibt an, wie die Edition beim Shipout verwendet wird.