Enhancing FAIR compliance: A controlled vocabulary for mapping Social Sciences survey variables

Janete Saldanha Bach[1] and Claus-Peter Klas[2]

Abstract

In Social Sciences surveys, the dynamic relationship among survey instruments and study entities like questionnaires, variables, questions, and response formats evolve. When reusing variables, researchers may need to modify variable attributes such as labels or names, question-wording, or response scales. Therefore, explaining these relations across different waves and studies is necessary to track how variables relate to each other. Although standards like Data Documentation Initiative – Lifecycle (DDI-LC) and DataCite model these relationships, these frameworks fall short of capturing the complexity of variable relationships. The DDI Alliance Controlled Vocabulary for Commonality Type employs codes—such as 'identical,' 'some,' and 'none'—to outline shifts in entities like variables; however, this approach is insufficient for disambiguating these relationships since they do not differentiate the variable attributes subject to change. To bridge this gap, we introduce the GESIS Controlled Vocabulary (CV) for Variables in Social Sciences Research Data. This CV is specifically designed to enhance semantic interoperability across various organizations and systems. By establishing explicit relationships, it not only facilitates harmonization across different study waves but also enriches data reuse. This enhancement supports advanced search and browse functionalities. The CV, published via the CESSDA vocabulary manager, seeks to forge a semantically rich, interconnected knowledge graph specifically tailored for Social Science Research. This endeavour aligns with the FAIR data principles, aiming to foster a more integrated and accessible research landscape.

Keywords

Controlled vocabulary, Survey variables - Social Sciences, Knowledge graphs

1. Introduction and motivation

In Social Sciences, research outputs are increasingly characterized by interdependent entities. These entities encompass a wide range, including surveys, questions, response schemas, datasets, variables, and various data types like audio and video files produced during data collection. Among these, variables within quantitative Social Science datasets emerge as a particularly interesting entity. Common variables in Social Sciences surveys include demographic factors such as age, education level, income, marital status, and more.

This first approach is motivated in the context of the Consortium for the Social, Behavioural, Educational and Economic Sciences - KonsortSWD[3] of the German National Research Data Infrastructure - NFDI[4]. The KonsortSWD Task Area 5-Measure-1 project[5] provides a technical solution to meet the growing demand for data services within the KonsortSWD's research data ecosystem, as highlighted by Klas et al. (2022).

The KonsortSWD PID registration service aims to assign PIDs for individual variables in datasets to make data findability and accessibility on the level of inline data objects of studies more efficient. As a consequence, variables are the most relevant entity to map their relations across waves[6] and studies. Our initial step in the KonsortSWD project involves identifying and documenting relations between variables. We aim to store these relationships within metadata in the PID registration service. Once a variable is documented and assigned a PID, it can be automatically incorporated into relationship maps, such as knowledge graphs (KGs). Examples of large-scale graphs include the Research Graph for connecting research data repositories, as discussed by Aryani et al. (2018), the Open Research

Knowledge Graph (Stocker et al., 2018), and The OpenAIRE research graph data model (Manghi et al., 2019).
Given that attributes of variables in Social Sciences, such as labels, names, question wording, or response scales, are prone to change, it is essential to offer a transparent explanation of their relationships across different waves and studies. This clarity is necessary to comprehensively track the evolution and interconnections of variables. While existing methodologies employ standards like the Data Documentation Initiative – Lifecycle (DDI-LC) and DataCite to model these relationships, they often do not fully encompass the intricate nature of variable relationships in Social Sciences. The DDI Alliance Controlled Vocabulary for Commonality Type[7] , for example, uses codes such as 'identical,' 'some,' and 'none'— to classify changes in entities like variables; However, this system falls short in effectively distinguishing the specific attributes of variables that undergo changes.
To bridge this gap, we have developed the GESIS Controlled Vocabulary for Variables in Social Sciences Research Data (outlined in section 3.1). This Controlled Vocabulary (CV) is designed to augment semantic interoperability across various organizations and systems. It provides not just a concise textual identifier for each variable relationship but also includes detailed descriptions to elucidate the nature of these relationships. This paper introduces this CV, highlighting its capability to represent these connections with machine-actionable features that facilitate the construction of a KG for Social Sciences. Both the CV and the KG are tailored for the detailed granularity required in research data, specifically survey variables.

Motivation Scenario. Assigning Persistent Identifiers (PIDs) to the finer attributes of datasets enables individual elements to be referenced and retrieved, complete with the necessary metadata for both machine-actionable and human access. Utilizing PIDs for referencing research data and their detailed entities aligns with the FAIR [8] principles of data usage, enhancing data reuse and citation, as well as facilitating applications in KGs. However, the relationships between variables in datasets are notably complex.
These relationships encompass a variety of aspects, including but not limited to different versions of variables, derived formats in subsequent waves, variations in labels and naming, and alternative response schemas in questionnaires and surveys. Variables may be added or omitted from wave to wave, influenced by the evolving research questions and objectives of the study. Additionally, the types of values variables hold—such as numerals, free texts, or controlled vocabularies—contribute to their differentiation. These properties can also change within the same study's lifecycle. For example, a variable's label might be altered from one wave to another while its underlying concept remains consistent. Likewise, the values of variables are subject to updates in their cardinalities, categorization, or response schema and scale, often modified to adapt to study evolution requirements or new sociological approaches.
In disciplines like Social Sciences, Economics, and Behavioural Sciences, which explore areas like the social structure of populations, political attitudes, opinions on various societal aspects, and competencies of adults, such variable attributes are highly sensitive to shifts in the empirical reality of a changing world. Studies in these fields must account for this evolution to maintain relevance in researching society. In this context, updating variables becomes crucial for accurately measuring transformation and societal dynamics.
The Data Documentation Initiative (DDI)[9] standards, which are prevalent in social science research, employ a set of controlled vocabularies to aid systems in identifying, locating, and accessing data for research. Developed and maintained by the DDI[10], these metadata elements are integral for research data management. An example is the metadata element *BasedOnObjectType*[11], used in scenarios where a new object is created based on an existing one, or when the new object represents more than just a version change, yet there is a need to reference the original object.
This feature is particularly crucial for tracking variable relationships across different waves and studies, as it enables detailed mapping of a variable's evolution. The *'BasedOnObjectType'* element offers a versatile approach to further describe the object. It can encompass multiple aspects: (a) references to

any number of objects that serve as a foundation for the new object, (b) a description of how the content from the referenced object was incorporated or altered, and (c) a code for specific typing of the object in line with an external controlled vocabulary.
We have applied a created Controlled Vocabulary (CV) to specify the relationships of variables using the *'BasedOnObject'* element, enhancing the description with a comprehensive set of variables' attributes. Additionally, we have incorporated elements from DataCite, specifically the *'Relation_Type'* metadata field and its subfields, as defined by the DataCite Metadata Working Group (2021). We have also integrated properties from Schema.org[12], such as *'isBasedOn'*, *'isBasedOnUrl'*, and *'isPartOf'*, to further enrich the metadata and facilitate robust data management and traceability in social science research.
Recognizing the Data Documentation Initiative (DDI) as a pivotal standard in the social science community for documenting and managing research data, we have adopted the DDI standard as the foundation for our CV codes. The DDI standard encompasses the entire research data lifecycle and provides metadata elements for describing data sets and related objects such as questions, variables, and values in datasets (Thomas et al., 2014). In line with this, we propose to expand the descriptions of relationships starting with *BasedOnObjectType*[13] as an initial approach.
Since our goal is to track variables across different waves and studies, 'BasedOnObjectType' emerges as the most fitting relation to use, especially when creating an object that represents more than just a version change and requires maintaining a reference to the original object. A key feature of *'BasedOnObjectType'* is its versionable property ('BasedOnReference_Versionable'), which allows for any type of versionable object to be referenced and repeated across multiple base objects. This flexibility is significant because it allows unlimited repetition and applicability to any kind of object.
Enhancing the definitions and explanations of these explicit relations leads to improved semantic clarity across and between variables, subsequently enhancing data findability and promoting the reuse of research data. Standardized terms through controlled vocabularies enable machine-actionable functions, further augmenting KGs.
Our contributions, based on the existing modelling metadata to describe relation types among entities, are as follows:

1. Extend the descriptions to elucidate relations for enhanced semantics, facilitating comparability between variable relations across waves (refer to section 3 for details).
2. Develop a Controlled Vocabulary (CV) for variable relations in the Social Sciences, aimed at boosting semantic interoperability across organizations and systems (detailed in section 3).
3. Establish a comprehensive framework for identifying relational connections of variables, integrating diverse DDI elements such as survey questions, response schema, data papers, interactive resources (like codes or scripts using the variable), data management plans, or audio/video data (see section 4 for the complete list).

This paper is structured in 5 sections: following the introduction, in section 2 we provide the related literature, further explaining the complex relation between variables with fundamental requirements to support KGs and the associated metadata standards. Section 3 exemplifies the knowledge graph for variable relations and provides examples of extended descriptions. Section 4 discusses the needs to explicitly variables' connections with other entities. Section 5 concludes and indicates further efforts.

2. Related literature

Surveys are fundamental in Social Sciences research, serving as a primary method for investigating variables (Babbie, 1990). Widely recognized as a key research paradigm, surveys are instrumental in measuring people's perceptions, intentions, and behaviours (Ajzen and Fishbein, 2005). Variables, the entities that shapes Social Science data, vary among individuals and over time, reflecting a range of values (Kaur and Mittal, 2021). Attitudinal variables, encompassing beliefs, values, opinions, attitudes,

and perceptions on specific topics, are central to major European surveys like the International Social Survey Programme – ISSP (ISSP Research Group, 1992), the European Social Survey - ESS[14], the National Educational Panel Study – NEPS (Roßbach and NEPS, 2016) and The Socio-Economic Panel – SOEP (Liebig *et al.,* 2021), among others.
These surveys also frequently explore observed behaviours, frequency of actions, and intentions within target groups. Furthermore, cross-domain studies often utilize psychological variables (examining aspects such as personality traits, emotional states, motivation, self-esteem, and self-efficacy) (Bollen, 2002), and environmental variables (concerning physical or social environments, resource access, and social support) (Cox, 2015), interchangeably for Social Sciences objectives.
In datasets, variables are organized as tabular data, structured in columns[15] and rows[16] to facilitate manipulation and inference, aligning with the study's objectives. This arrangement is typical for variable data that has been collected, archived, and disseminated. Survey variables from these studies encompass a broad spectrum of topics, accumulating vast amounts of data from numerous individuals over extended periods. This leads to the generation of thousands of variable units, necessitating scalable data management solutions for large-scale KGs.
For example, the SOEP-Core[17], the main component of the German Socio-Economic Panel (SOEP). This extensive longitudinal study, ongoing since 1984, annually surveys the living conditions and attitudes of over 15,000 households, involving about 30,000 individuals. It represents the most comprehensive long-term study of social developments in Germany. Within this repository, there are 560 datasets, encompassing 21,280 questions across 309 instruments, and 101,574 variables[18].

2.1 Relations between variables

Beyond managing the enormous quantity of variables, many complex relations are also challenging to interpret through textual analysis. Relations within variables extend beyond simple variations and include aspects such as different versions, derived formats in new waves, variations in labels and naming, and alternative response schemas through questionnaires and surveys. The variability of a variable goes beyond just the response options provided by individual cases in a survey. Here are detailed examples of how variables' relationships manifest within different attributes:

a) Variables' Name: Often, a variable is related to other variables from different studies. For instance, a study on work-life balance may include a variable named 'work_life_bal', which correlates with the variable named 'job_sat' in a separate study. Despite different names, both variables aim to measure the same concept, such as job satisfaction.
b) Survey Question: Variables are frequently used in survey questions across different studies. For example, the question 'How satisfied are you with your job on a scale of 1 to 5?' utilizes the variable 'job_sat' to gather data. In another wave or study, the same variable 'job_sat' might be reused, but the question could be modified to suit a new response scale, like 'How satisfied are you with your job on a scale of 1 to 7?'
c) Scales: The way a variable's answers are represented can differ between studies. A variable like 'job_sat' might initially use a Likert scale with response options from 1 (Strongly dissatisfied) to 5 (Strongly satisfied). However, in another wave or a different study, the same variable might be measured with an extended Likert scale, ranging from 1 (Strongly dissatisfied) to 7 (Strongly satisfied). This broader range allows for capturing more nuanced responses, for example:

Likert Scale 5
1. Strongly dissatisfied
2. dissatisfied

Likert Scale 7
1. Strongly dissatisfied
2. Moderately dissatisfied

3. Neutral
4. satisfied
5. Strongly satisfied

3. Slightly dissatisfied
4. Neutral
5. Slightly satisfied
6. Moderately satisfied
7. Strongly satisfied

Panel studies often survey the same individuals or groups repeatedly, measuring the same variables across multiple waves to examine changes in opinions over time. However, the dimensions of these measures or other rules may also evolve. Variables across series are related not only in terms of their content but also in their quantity, and a one-to-one relationship is not always present. In some cases, multiple variables from a previous series are merged into a single variable in the next, or a single variable is divided into several. Researchers may be interested in determining if a specific variable is consistently present across all time points. In cross-sectional studies, conducted at a single point in time, variables may also differ depending on the sample or population studied. For instance, a study focusing solely on college students may include more diverse variables or measures than a study encompassing the general population. Transparency in documenting variables and any modifications across different waves or samples is critical, regardless of the study design. Variables may be modified due to several factors:

a) The research question or study goals: Variables are selected to answer specific research questions, which may evolve over time, necessitating the measurement of different variables in later study waves;
b) The sample or population: Variables can vary across studies or waves depending on the population studied. Different target groups, like college students, may have distinct variables compared to other demographic groups;
c) Measurement instruments or methods: Variations in survey questions can alter how variables are measured, resulting in differences across studies or waves;
d) The societal, political, or economic environment: Broader conditions can influence variables. For example, the SOEP [19] was expanded in 1990 to include East Germany post-reunification and in 2016 to incorporate a sample of refugees;
e) Data availability or quality: New data sources or improvements in data collection can lead to variations in the variables used across studies or waves.

Variable documentation is essential for providing transparency and provenance information about the lifecycle of variables in studies. However, deriving insights from these sources can be complex and time-consuming. To address this, controlled vocabularies, as developed by the DDI Alliance's Controlled Vocabularies Group (CVG), are vital for defining metadata element meanings, improving consistency, comparability, and efficiency of documentation, and enhancing information retrieval (Jaaskelainen, Moschner, and Wackerow, 2010).
The DDI Alliance's controlled vocabulary for Commonality Type, which aims to describe the degree of similarity between items, uses codes like 'identical,' 'some,' and 'none.' For instance, 'identical' indicates that all attributes of a variable are the same, while 'some' suggests similarity but not complete identity. However, 'some' does not specify which attributes differ, making it insufficient for disambiguating relationships between variables. A third code, 'none,' indicates an absence of comparability where it was expected. We will describe support standards to better document the reuse and adaptation of variables across waves and studies.

2.2 Modelling metadata standards

Social science research employs a variety of methods and standards to document studies and enable the tracking of variables across different studies or waves. Commonly used standards and best practices include codebooks, data dictionaries, metadata schema, and longitudinal tracking.

Codebooks provide detailed information about variables and data collected in a study. Best practices for creating codebooks involve offering clear descriptions of variables and their measurements, including coding instructions, recoding procedures, and maintaining consistent terminology and formatting. The Data Documentation Initiative (DDI) is the standard file format for codebooks, utilizing eXtensible Markup Language (XML) for metadata specification.
Data Dictionaries focus more on the structure and format of the data. They are crucial for ensuring consistent data collection and organization. Best practices here include providing clear definitions for variables, specifying variable names and labels, and adhering to standard data types and codes. Data dictionaries often use standard file formats like DDI and Statistical Data and Metadata Exchange (SDMX) [20].
Metadata Standards offer comprehensive information about study design, sampling methods, and data collection procedures. The Dublin Core Metadata Initiative[21] is a standard format for metadata for digital objects in general, while the DataCite Metadata Schema (DataCite Metadata Working Group, 2021) is specifically tailored for documenting research data publication and citation.
Longitudinal tracking is another technical feature available for researchers to follow individual respondents over time in longitudinal studies. These features help ensure that the same variable is measured for individuals across different study waves, and these variables lead to identifiable persons. This practice includes using unique identifiers for individuals, which requires a higher level of data security and privacy to comply with privacy laws such as the General Data Protection Regulation - GDPR (European Union, 2016) while using standardized protocols for tracking individuals over time. Standard file formats, such as the Longitudinal Data File (LDF) format, are commonly used for longitudinal data.

Considering the importance of variables and their relationships, it is vital to describe the associated metadata to register these relationships and enable machine-actionable features through Persistent Identifiers (PIDs) and controlled vocabulary terms. The metadata schema is designed to cater to the growing needs for interoperability, data mappings, and Knowledge Graphs. This solution includes a metadata schema for persistent identification and cross-linking of relationships.

2.3 Metadata for variable relations

In the KonsortSWD project, the identification of variable relations is primarily conducted through the PID registration service. This process requires detailed metadata both at the study or dataset level and, importantly, at the individual variable level. This metadata is crucial for registering a variable and obtaining its Persistent Identifier (PID) (Saldanha Bach, Klas, and Mutschke, 2023). A key component of this metadata schema is dedicated to capturing the relationships between variables.
One of the central metadata fields in this schema is 'Related_Item', which describes the resource type, in this case, the variable. Each 'Related_Item' must be accompanied by an identifier, provided through the 'Related_Item_Identifier' field. This identifier is preferably a PID or a code from a controlled vocabulary.
Following the identification of the 'Related_Item', the type of its identifier ('Related_Item_Identifier_Type') must be specified. This is important due to the varied syntax used by different PID systems. Another critical field in this metadata schema is 'Relation_Type', designed to define the nature of the relationship between two variables, labelled as A and B. This foundational metadata schema, which mirrors the 'RelatedItem' field from DataCite (DataCite Metadata Working Group, 2021), consists of the following fields and subfields:

- *Related_Item*
    - *Related_Item_Identifier*
        - *Related_Item_Identifier_Type*
    - *Relation_Type*

The DDI variable cascade (DDI Training Group, 2021) categorizes comparability among variables into three layers: represented variable, conceptual variable, and instance variable. However, our current focus within DDI modelling[22] is limited to the variables within the context of a dataset[23], not extending to these abstraction[24] layers. This limitation stems from the service requirement, as our metadata acquisition is solely for variables that are being registered for PIDs. The extended description within the 'BasedOnObjectType' DDI is detailed in the subsequent section.

3. Enhancing description quality for Knowledge Graph Relationships

A Knowledge Graph (KG) is an advanced data model that encapsulates knowledge in a graph format, where entities and their interrelations are described in a way that machines can interpret. This model is constructed using a suite of established W3C standards, including the Resource Description Framework (RDF), JSON for data interchange, the Simple Knowledge Organization System (SKOS) for organizing knowledge, and the Web Ontology Language (OWL) for defining and categorizing web content. These standards are complemented by shared vocabularies and application programming interfaces (APIs), which facilitate the integration of data from diverse domains and sources.
A key feature of KGs is their consistent use of Persistent Identifiers (PIDs). These identifiers play a crucial role in ensuring that entities and their relationships are not only identifiable but also linkable across various data sources and applications. This capability is fundamental for integrating heterogeneous data sources into a cohesive and interconnected knowledge base. It supports semantic search and question answering, allowing users to query and retrieve knowledge using natural language queries. Beyond using KG in the high-tech industry, such as Google[25], Microsoft[26] and Amazon[27], KGs are also widely applied in scientific domains. Large projects such as The Open Academic Graph[28], used for research for scholarly publications, the Linked Open Data Cloud[29], interlinked datasets from various domains, and the Global Biodiversity Information Facility (GBIF)[30], a KG of biodiversity data are some examples.

In the Social Sciences, KGs have become instrumental for various research areas, including understanding societal and political debates, investigating fake news and misinformation (Gangopadhyay et al., 2023), and mining knowledge about opinions and interactions from X data, formerly known as Twitter (Fafalios et al., 2018). The GESIS Research Graph project[31] also features a prototype graph that interlinks publications, research data, projects, and people. These initiatives are part of the broader effort to build a KG infrastructure that links social science research data and resources across GESIS[32].
The core strength of KGs lies in their ability to connect, manage, and elucidate complex relationships, a feature that extends to variables in social science research. KGs are adept at capturing and representing the multidirectional connections between entities. By depicting variables as nodes and their relationships as edges, researchers gain a clearer understanding of how variables are associated and interact, facilitating the extraction of insights and predictions from data.
Adopting recognized standards and APIs in KG construction ensures that the information is interoperable and machine-interpretable. This approach fosters the development of intelligent services and applications capable of automating the analysis and processing of variable data. Thus, KGs emerge as vital tools in managing, analyzing, and sharing intricate information in social science research.

In this sense, KG design benefits from extended and detailed variable descriptions. Accurate descriptions improve data quality, leading to more reliable connections. Transparent documentation of variable allows other researchers to understand the relation scope. Detailed descriptions facilitate the integration of data from multiple sources and studies, enhancing data discoverability and making it easier for researchers to identify and use the data they need.

Datasets in repositories often consist of multiple files and sub-collections (Wehrle and Rechert, 2019; Bugaje and Chowdhury, 2017), making fine-grained levels, such as individual variables, crucial for research data management.

These detailed connections enhance data reuse by enriching the decision-making process for researchers who often select specific variables rather than entire datasets. A primary need for social scientists reusing data is to swiftly comprehend the meanings and values of variables within a dataset (Sun and Khoo, 2018). However, challenges arise from terminology polysemy, where similar variable concepts may have different names or variables with the same name may represent different concepts. This issue necessitates intellectual effort to understand variables across studies and waves. An extensive research university library's experience with data reuse, including issues of replicability and reproducibility, underscores the importance of providing descriptive information about variables' coverage across datasets and specifying variable data definitions (Scoulas, 2020). For example, Table 1 illustrates the relationship between variables A and B across different waves, where variable B in wave 2 is *BasedOn* the variable A from wave 1, although with a different name.

Table 1

Variables relations: differences across waves: variable name

Variables relations	Variables	Study Program	Wave	Variable name	Variable label
IsBasedOn.hasDifferentVarName	*B -> A*	*is equal*	*Is different*	*Is different*	*is equal*
Variable 1	Var_1	study#100	Wave1	Age	Age in years
Variable 2	Var_2	study#100	Wave2	Age_group	Age in years

Note: ***IsBasedOn*** (DDI-LC)

Figure 1 depicts the variables' relations across waves regarding different variable name. Variable B *is based on* Variable A because it was generated latter in a more recent wave. Although variable B is based on A, B has a different variable name (Age_group) than the original Variable A (Age).

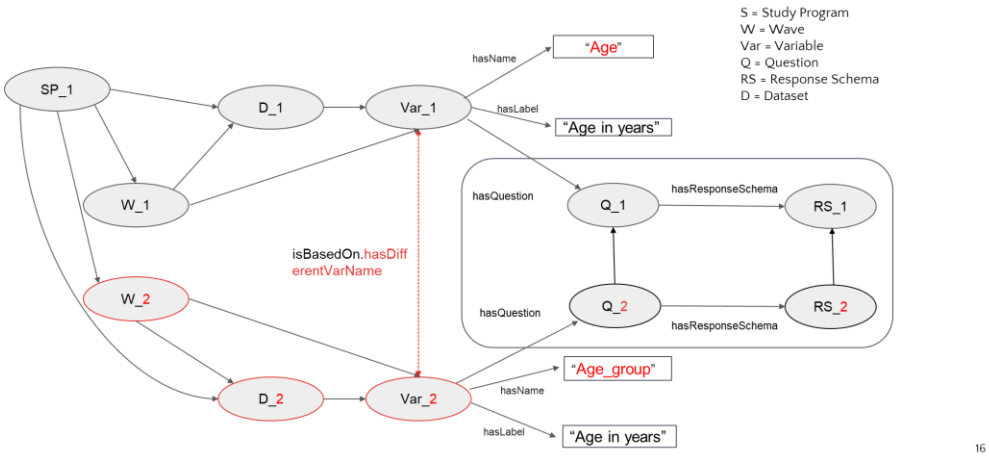

Figure 2: Representation of the relations in a Knowledge Graph: different variable name
Label: 1 = Equal; 2 = Different. The dotted lines represent relations between entities.

Figure 2 depicts one example of variables A and B relation where a variable B is BasedOn a Variable A, but it has a different question wording. In this case the word *'daß'* (the German word which means 'that') uses the characterr *ß* (called ***Eszett***). The *Eszett* letter is used only in German and can be typographically replaced with the double-s digraph 'ss.' In a more recent wave, the same word is replaced by the form 'dass,' adopting double-s. Figure 2 depicts the KG representation of variables' relations across waves regarding question-wording.

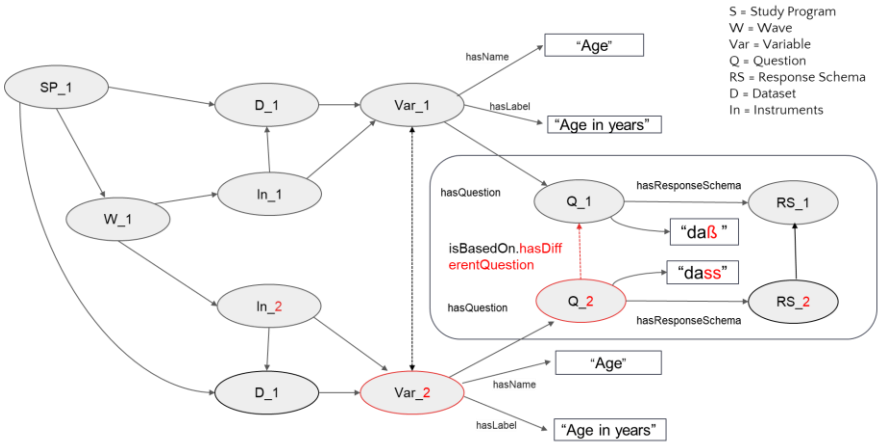

Figure 3: Representation of the relations in a Knowledge Graph: different question-wording
Label: 1 = Equal; 2 = Different. The dotted lines represent relations between entities.

Figure 3 depicts one example of variables A and B relation where a variable B is BasedOn a Variable A, but it has a different Response Schema. Likert Scale from variables A and B differs from 5 to 7 in each wave, respectively. Figure 3 depicts the variables' relations across waves regarding different Response Schema.

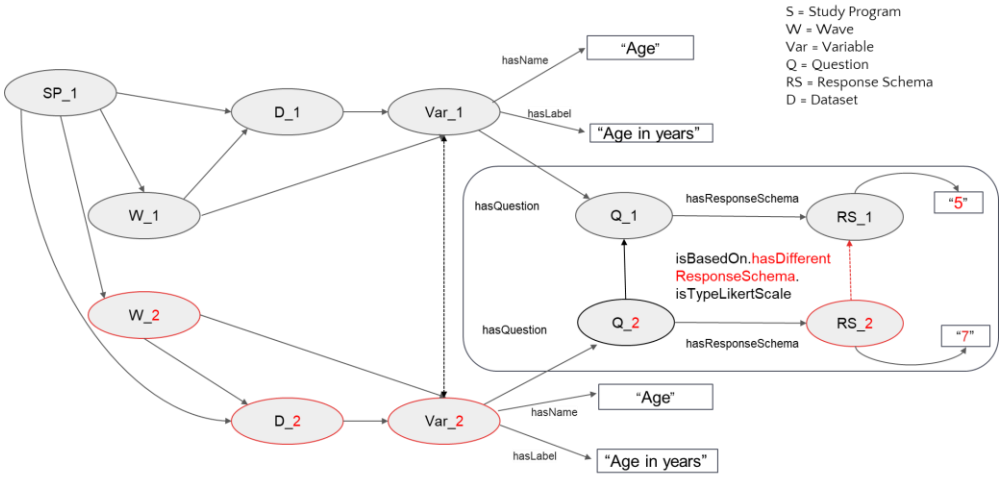

Figure 4: Representation of the relations in a Knowledge Graph: different response schema
Label: S = Study Program | W = Wave | Var = Variable | Q = Question | RS = Response Schema | D = Dataset

The ease of discovering and visualizing dataset variable relations through KGs significantly enhances their comparability across different waves of a study. This functionality aids in understanding how variables interact within and between diverse datasets of several types. In longitudinal studies, tracking changes in variables across different waves becomes more manageable, facilitating the often costly and time-consuming harmonization process among datasets. KGs, with their search and browse functionalities, also augment data discoverability and findability (Wu et al., 2019). They enable (inter)disciplinary data reuse by visually depicting how variables are distributed across multi-wave studies and identifying which variables have been consistently used over time.
Controlled vocabulary with extended descriptions of relations simplifies the task of finding connections between variables within the same study or across different datasets. We provide concise textual identifications for each *relation_type*, supplemented by a CV and thorough explanations of these relationships. This approach extends beyond merely naming and labelling variables. It also facilitates the discovery of relationships inherited within the DDI structure and other potential entities, such as Data Papers and additional resources (refer to section 4).
Employing these proposed relationships and the resulting controlled vocabulary leads to the creation of a semantically rich, common framework for Social Science research. These connections can be effectively represented in a KG across various institutions, in line with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles for variables. This method not only enhances the understanding of variable relations but also promotes the efficient and informed use of research data in the Social Sciences.

3.1 Controlled vocabulary (CV) for variables relations in the Social Sciences

The CESSDA (Consortium of European Social Science Data Archives) vocabulary manager plays a pivotal role in documenting and clarifying the relationships within social science research data. It provides extended descriptions and controlled vocabulary terms that describe links across various waves and studies, in conjunction with questions and other related entities. A study within this framework can encompass multiple waves, and each wave may include multiple surveys. Each survey is comprised of numerous questions, and each question can relate to one or more variables. These variables are defined by a variable name (or a Variable ID, typically a code), a variable label (which describes the variable), a response schema, and potentially, terms from a controlled vocabulary.

Our focus is understanding the relationship of a variable from its own standpoint, specifically how it can be related to these different attributes. The CV adheres to the DDI property *'BasedOnReference_Versionable*.' This property allows for references to any number of objects that form the basis of the variable, a *'BasedOnRationalDescription'* detailing how the content of the referenced object was incorporated or altered, and a *'BasedOnRationalCode'* for specific typing of the *'BasedOnReference'* in accordance with an external controlled vocabulary. This CV[33] is published at the CESSDA CV manager. Our initial contribution to this domain includes six relation types, each thoroughly detailed within the CV, listed below:

1. IsBasedOn.hasDifferent**WaveVariable**
2. IsBasedOn.hasDifferent**SurveyVariable**
3. IsBasedOn.hasDifferent**VarNameVariable**
4. IsBasedOn.hasDifferent**VarLabelVariable**
5. IsBasedOn.hasDifferent**QuestionVariable**
6. IsBasedOn.hasDifferent**ResponseSchema**

To summarize the relations, Table 2 provides examples of changes in the variable's attributes.

Table 2: Variables relations extended descriptions

Related_Item	***Proposed Relation_Type***	***Examples***	Vn = Variable Name	Vl = Variable Label	Q = Question	RS = Response schema	Sy = Survey	W = Wave	S = Study
Variable in WAVES	IsBasedOn.hasDifferentWave	study#100-Wave1-Variable:v5.-> study#100-Wave2-Variable:v5	=	=	=	=	=	≠	=
Variable in SURVEYS	IsBasedOn.hasDifferentSurvey	study#100-Wave1-SurveyA-Variable:v5.-> study#100-Wave1-SurveyB-Variable:v5					≠	=	=
Variable NAME	IsBasedOn.hasDifferentVarName	study#100-Wave1-Variable:v5-"job_sat".-> study#100-Wave2-Variable:v7-"job_sat".	≠					≠	=
Variable LABEL	IsBasedOn.hasDifferentVarLabel	study#100-Wave1-Variable:v5-"job_sat".-> study#100-Wave2-Variable:v5-"work_sat"		≠				≠	=
Variable QUESTION wording	IsBasedOn.hasDifferentQuestion	study#100-Wave1-Questionabc-Variable:v5. -> study#100-Wave2-Questionxyz-Variable:v7.			≠			≠	=
Variable RESPONSE SCHEMA (response values)	IsBasedOn.DifferentResponseSchema.isTypeLikertScale	study#100-Wave1-Qabc-Variable:v7-Likert4points -> study#100-Wave2-Qabc-Variable:v7-Likert5points .				≠		≠	=
			LABEL	=	Equal				
				≠	Different				

* Based on DDI term *IsBasedOn* and on the Controlled Vocabulary for Variables relations for Social Sciences research data.
Label: S (Study); W (Wave); Sy (Survey); Q (Question); VN (Variable Name); VL (Variable Label); RS (Response schema).
1 = Equal; 2 = Different

The following section address relations between variables and other entities within different studies.

4. Relations inherited within the DDI framing

To explain the multifaceted interactions of variables with different entities, we have pinpointed specific types of connections that form a network of elements. This exploration helps identify which elements correlate most effectively with variables. For instance, a variable is inherently linked to a survey question. Our aim is to demonstrate how a variable can be connected to a range of entities beyond its original study context. Take, for example, a hypothetical variable named 'job_sat'. We can visualize its relationships with various entities through the following scenarios:

a) Research or Data Papers: A variable may be cited or featured in academic papers. For instance, a paper exploring job satisfaction might reference 'job_sat' as a key factor in understanding employee well-being;
b) Landing Page: Websites can offer detailed metadata about a variable. An example is the webpage 'www.example.com/job-satisfaction', which could provide comprehensive metadata and descriptions of 'job_sat'

c) Interactive Resources: Scripts or codes often utilize variables for data analysis. For example, a Python script could use 'job_sat' to process survey data, create visual representations, or analyse trends in job satisfaction
d) Data Management Plan: Such plans might include anticipated use of variables. A workplace wellness study's plan could specify using 'job_sat' for data collection and analysis;
e) Audio/Video Data: Variables can be incorporated into multimedia formats. A video presentation on study outcomes, for instance, might include discussions and visualizations of 'job_sat', highlighting its impact on employee happiness.

By expanding the KG to encompass these diverse entities, using controlled list values from resource types descriptions, we maintain the KG's interoperability across different domains. Linking entities to their associated variables provides a comprehensive overview of their interdependent connections. This approach significantly enhances the data's findability, accessibility, interoperability, and potential for future reuse

5. Conclusion

To effectively register relationships between variables within studies, standard metadata fields are indispensable. These standards are crucial for ensuring interoperability among different systems and enabling automation features. Accurately representing possible *relation_types*, and formally documenting them, significantly enhances meta searching and meta browsing capabilities. This makes it easier to find and access relevant data. The key requirements are to uniquely identify each variable with a Persistent Identifier (PID) and to clearly define its relationships using controlled vocabulary terms. Such approaches are instrumental in fostering machine-actionable data features, thereby strengthening the findability of data and enhancing the reusability of data at the variable level.
Data users can benefit from the ability to correspond variables, exploring their consistency or comparability over time, across different waves and studies. With machine-readable and actionable features, complex recommendation systems can be developed. These systems can display relationships between variables and other entries in relationship maps, such as those represented in KGs. While the PID Registration Service's primary function is not to provide KG visualization, its inclusion of the *'Related_Item'* field and corresponding subfields in its metadata schema lays the groundwork for documenting variables in a way that enhances KG applications.
We propose an extended description for the *'Relation_Type'* description and a controlled vocabulary terminology based on the DDI term *'IsBasedOn'*. This approach enables researchers and other interested parties to easily locate the most relevant and usable variables for their research needs. For data holders, this method facilitates the maximization of value-added services through the increasing interconnection of research output entities. Variables are not only linked to their inherent elements like questions, questionnaires, survey waves, and response scales but can also be inputs for interactive resources such as scripts or Do-files. There is also potential for registering and assigning PIDs to questions and response schemas from existing surveys for reuse purposes.
Documenting and defining all these relations accurately, with detailed relation descriptions, will enhance the controlled vocabulary for the Social Sciences. This in turn will foster the reuse of the CESSDA Controlled Vocabulary tool among institutions, leveraging these interconnected relationships for broader research and analysis purposes.

References

Ajzen, I. and Fishbein, M. (2005), 'The influence of attitudes on behavior', in Albarracin, D., Johnson, B. T. and Zanna, M.P. (Eds), *Handbook of Attitudes and Attitude Change*, Lawrence Erlbaum Associates, Mahwah, NJ.

Aryani, A. *et al.* (2018) 'A Research Graph dataset for connecting research data repositories using RD-Switchboard', *Scientific Data*, 5(1), p. 180099. Available at: https://doi.org/10.1038/sdata.2018.99.

Babbie, E.R. (1990). *Survey Research Methods*, Wadsworth Publishing, Belmont, CA.

Bollen, K.A. (2002) 'Latent Variables in Psychology and the Social Sciences', *Annual Review of Psychology*, 53(1), pp. 605–634. Available at: https://doi.org/10.1146/annurev.psych.53.100901.135239

Bugaje, M. and Chowdhury, G. (2017) 'Is Data Retrieval Different from Text Retrieval? An Exploratory Study', in S. Choemprayong, F. Crestani, and S.J. Cunningham (eds) *Digital Libraries: Data, Information, and Knowledge for Digital Lives*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 97–103. Available at: https://doi.org/10.1007/978-3-319-70232-2_8

Cox, M. (2015) 'A basic guide for empirical environmental social science', *Ecology and Society*, 20(1), p. art63. Available at: https://doi.org/10.5751/ES-07400-200163

DataCite Metadata Working Group (2021) 'DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4', p. 82 pages. Available at: https://doi.org/10.14454/3W3Z-SA82.

DDI Training Group (2021) 'Variables and the Variable Cascade'. Available at: https://doi.org/10.5281/ZENODO.5180568.

European Union. (2016). 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)'. *Official Journal of the European Union*. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj

Fafalios, P; Iosifidis, V.; Ntoutsi, E. and Dietze, S. TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets. In *15th Extended Semantic Web Conference* (ESWC'18), Heraklion, Crete, Greece, June 3-7, 2018.
https://doi.org/10.48550/arXiv.1810.10308

Gangopadhyay, S., Boland, K., Dessí, D., Dietze, S., Fafalios, P., Tchechmedjiev, A., ... & Jabeen, H. (2023, May). Truth or dare: Investigating claims truthfulness with claimskg. In *Second International Workshop on Linked Data-driven Resilience Research* (D2R2'23) co-located with ESWC 2023, May 28th, 2023, Hersonissos, Greece. Available at: https://ceur-ws.org/Vol-3401/paper7.pdf

ISSP Research Group (1992) 'International Social Survey Programme: Role of Government II - ISSP 1990 International Social Survey Programme: Role of Government II - ISSP 1990'. GESIS Data Archive. Available at: https://doi.org/10.4232/1.1950

Jaaskelainen, T., Moschner, M. and Wackerow, J. (2010) 'Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability', *IASSIST Quarterly*, 33(1), p. 34. Available at: https://doi.org/10.29173/iq649

Kaur, Loveleen and Mittal, Ritu. (2021). 'Variables in Social Science Research'. *Indian Res. J. Ext. Edu*. 21 (2&3), April & July, 2021. URL: https://www.researchgate.net/profile/Ritu-Mittal-2/publication/351080413_Variables_in_Social_Science_Research/links/6083aa49907dcf667bbda5cf/Variables-in-Social-Science-Research.pdf

Klas, C.-P. *et al.* (2022) *KonsortSWD Measure 5.1: PID Service for variables report*. Zenodo. Available at: https://doi.org/10.5281/ZENODO.6397367.

Manghi, P. *et al.* (2019) *The OpenAIRE Research Graph Data Model*. Zenodo. Available at: https://doi.org/10.5281/ZENODO.2643199.

Liebig, S. *et al.* (2021) 'Socio-Economic Panel, data from 1984-2019, (SOEP-Core, v36, EU Edition) Sozio-oekonomisches Panel, Daten der Jahre 1984-2019 (SOEP-Core, v36, EU Edition)'. SOEP Socio-Economic Panel Study. Available at: https://doi.org/10.5684/SOEP.CORE.V36EU.

Roßbach, H.-G. and NEPS, National Educational Panel Study, Bamberg (Germany) (2016) 'NEPS Starting Cohort 6: Adults (SC6 6.0.1)NEPS-Startkohorte 6: Erwachsene (SC6 6.0.1)'. NEPS National Education Panel Study. Available at: https://doi.org/10.5157/NEPS:SC6:6.0.1

Saldanha Bach, J., Klas, C.-P. and Mutschke, P. (2023) *KonsortSWD Measure 5.1: use cases description extended report*. Zenodo. Available at: https://doi.org/10.5281/ZENODO.7588944

Saldanha Bach, J., Klas, C.-P. and Mutschke, P. (2023) *KonsortSWD Measure 5.1: metadata schema extended report*. Zenodo. Available at: https://doi.org/10.5281/ZENODO.7588902

Scoulas, J.M. (2020) 'Learning from data reuse: successful and failed experiences in a large public research university library', *IASSIST Quarterly*, 44(1–2), pp. 1–15. Available at: https://doi.org/10.29173/iq966

Stocker, M. *et al.* (2018) 'Curating Scientific Information in Knowledge Infrastructures', *Data Science Journal*, 17, p. 21. Available at: https://doi.org/10.5334/dsj-2018-021.

Sun, G. and Khoo, C.S.G. (2018) 'A Framework to represent variables and values in Social Science research data sets to support data curation and reuse', in F. Ribeiro and M.E. Cerveira (eds) *Challenges and Opportunities for Knowledge Organization in the Digital Age*. Ergon Verlag, pp. 231–239. Available at: https://doi.org/10.5771/9783956504211-231

Thomas, W., *et al.* (2014). Data documentation initiative: technical specification Part I Version 3.2. URL: https://ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/HighLevelDocumentation/DDI_Part_I_TechnicalDocument.pdf

Wehrle, D. and Rechert, K. (2019) 'Are Research Datasets FAIR in the Long Run?', *International Journal of Digital Curation*, 13(1), pp. 294–305. Available at: https://doi.org/10.2218/ijdc.v13i1.659

Wu, M. *et al.* (2019) 'Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories', *Data Science Journal*, 18, p. 3. Available at: https://doi.org/10.5334/dsj-2019-003

Notes

[1] Dr. Janete Saldanha Bach is a Postdoc researcher at GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, Germany and can be reached by email: janete.saldanhabach@gesis.org.

[2] Dr. Claus-Peter Klas is the Team Leader Data & Service Engineering at GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, Germany.

[3] KonsortSWD (Consortium for the Social, Behavioural, Educational and Economic Sciences) is funded by the National Research Data Infrastructure (NFDI) https://www.konsortswd.de/
[4] German National Research Data Infrastructure (NFDI) Homepage: https://www.nfdi.de/
[5] Project Community at Zenodo https://zenodo.org/communities/konsortswd-ta5-m1
[6] Waves are different points in time when data is collected in a research study. Waves are typically associated with longitudinal studies, which involve the repeated observation of the same subjects over time.
[7] https://vocabularies.cessda.eu/vocabulary/CommonalityType?lang=en
[8] FAIR stands for Findable, Accessible, Interoperable and Reusable. It refers to the FAIR Data Principles developed by the FORCE 11 community, that recommend data should be shared according to these four concepts.
[9] The Data Documentation Initiative (DDI) is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle.
[10] https://www.ddialliance.org/about/about-the-alliance
[11] BasedOnObjectType available at https://ddialliance.github.io/ddimodel-web/DDI-L-3.3/composite-types/BasedOnObjectType/
[12] https://schema.org/Product
[13] https://ddialliance.github.io/ddimodel-web/DDI-L-3.3/composite-types/BasedOnObjectType/
[14] https://ess-search.nsd.no/CDW/ConceptVariables
[15] In a data file, a single vertical column, each being one byte in length. Fixed-format data files are traditionally described as being arranged in lines and columns. In a fixed format file, column locations describe the locations of variables. (URL: https://www.icpsr.umich.edu/web/ICPSR/cms/2042)
[16] In general, a 'line' in data file terminology refers to a physical unit of data that the computer reads and processes, one at a time. (URL: https://www.icpsr.umich.edu/web/ICPSR/cms/2042).
[17] URL: https://www.diw.de/soep
[18] URL: https://paneldata.org/soep-core/
[19] URL: https://doi.org/10.5684/soep.core.v37i
[20] https://sdmx.org/
[21] https://www.dublincore.org/
[22] https://ddi4.readthedocs.io/en/latest/userguides/variablecascade.html#example
[23] This layer is so-called *instance variables* in the DDI-CDI model and *variables* in the DDI-LC model
[24] Abstraction layers are the *represented variable* and *conceptual variable*
[25] https://developers.google.com/knowledge-graph
[26] https://www.microsoft.com/en-us/research/group/cognitive-services-research/knowledge-and-language/
[27] https://aws.amazon.com/neptune/knowledge-graphs-on-aws/
[28] https://www.microsoft.com/en-us/research/project/open-academic-graph/overview/
[29] https://lod-cloud.net/
[30] https://www.gbif.org/
[31] https://researchgraph.org/gesis-research-graph/
[32] https://www.gesis.org/en/research/applied-computer-science/knowledge-graph-infrastructure
[33] https://vocabularies.cessda.eu/vocabulary/Variables-Relations?lang=en