KonsortSWD Measure 5.1: PID Service for variables report

Claus-Peter Klas[1*], Matthäus Zloch[2], Janete Saldanha Bach[3], Erdal Baran[4], Peter Mutschke[5].

March 2022

Abstract (in German)

Die Referenzierung von Forschungsdaten und der ihnen innewohnenden detaillierten Entitäten unterstützt die Nutzung im FAIR-Sinne. In Measure 5.1 verbessern wir den Stand der Technik bei der Zitierung von Forschungsdaten, indem wir eine Infrastruktur zur Referenzierung detaillierter Attribute, hier zunächst Variablen, innerhalb solcher Daten entwickeln. Durch die Zuweisung von PIDs zu diesen Attributen können einzelne Elemente der Datendateien referenziert und mit den erforderlichen Metadaten für den maschinellen und menschlichen Zugriff abgerufen werden. Die PIDs ermöglichen nicht nur die Zitierbarkeit in wissenschaftlichen Arbeiten, sondern auch den Zugang zur Verarbeitung der enthaltenen Daten selbst, z.B. in Skriptsprachen wie R oder Phyton.

Dieser Bericht enthält detaillierte Anwendungsfälle, das notwendige Metadatenschema und die vorgesehene Architektur für eine allgemeine, wartbare und skalierbare Infrastruktur, die die Registrierung von PIDs für kleinteilige Attribute in sozialwissenschaftliche Forschungsdaten ermöglicht. Neben der Hauptfunktionalität legen wir Wert auf wiederverwendbare und verallgemeinerbare Komponenten als Blaupause für andere Projekte. Alle entwickelten Komponenten werden als Open-Source-Software veröffentlicht.

Abstract

Referencing research data and there inherit detailed entities supports FAIR usage. In measure 5.1 we enhance the state of the art of citing research data, by developing an infrastructure to reference detailed attributes, here initially variables, within such data. By assigning PIDs to these attributes, individual elements of the data files can be referenced and retrieved with the required metadata for machine-actionable and human access. The PIDs will not only enable citeability within scientific papers but also give access for processing the contained data itself, e.g., within script languages like R or phyton.

This report provides detailed use cases, the necessary metadata schema, and the envisioned architecture for a general, maintainable, and scalable infrastructure that enables the registration of PIDs for small-scale attributes in Social Science research data. Besides the main functionality, we emphasize reusable and generalized components as a blueprint for other projects. All developed components will be published as open-source software.

Keywords:

Persistent Identifiers - PIDs. Research data citation. Research data - Social Sciences. Variables - Social Sciences. Granularity level identification. Research data services - technical infrastructure.

1. Introduction

Researchers in the Social Sciences have experienced an increasing research data availability within data repositories. The FAIR[1] principles (Wilkinson *et al.*, 2016) have embedded scientific knowledge construction in visibility and intensified reproducibility approaches, allowing better open research data obtainable. In publishing and opening the data practices, there is, on the one hand, the researcher in a data producer role that can provide evidence for their results, enhancing transparency and compliance with openness requirements. On the other hand, there are data re-users who are researchers, professionals, or interested citizens that can benefit from accessing, gathering, analysing, scrutinizing, or reusing the research data for many purposes.

However, there is an essential step between data producers and users to connect their tie: the data citation. The data citation not only recognizes and credits the data producer but links various actors in the scientific landscape in an extensive network since these data connect with more elements such as an instrument, an institution, a funding agency, a country, or one or more digital objects. Nevertheless, demonstrating the source details of what data was precisely used and where it came from is not a simple task for data re-users. One of the most challenging reuse phases for researchers is understanding licenses and citing data properly (Estevão, 2019).

Data in the Social Sciences context has various levels of granularity. Datasets commonly contain questions, variables, variables values, indicators values or scales and cohorts. In this sense, data producers have a significant responsibility since they publish the data and are the foremost interested in having their data available due to funders or institutional obligations or any other Open Science perceived advantages and commitments. The study and the dataset are the most common granularity levels to be identified when authors publish their research outputs. To identify those digital objects persistently, the Persistent Identifiers (PIDs) are the foundation on the long-term reference of scientific publication. A Persistent Identifier (PID) is persistent, unique and globally resolvable identifier that is based on an openly specified PID Scheme (European Commission, 2020).

Due to these complex data citation requisites, PIDs have been the assignment for digital resources identification. There are many PIDs standards, such as Digital Object Identifier (DOI)[2], Handle[3], Uniform Resource Name (URN)[4]or Archival Resource Keys (ARK), even though their goal is the same: to provide a unique name to an entity, permanently and uniquely referencing it on the Internet.

PIDs as data identifiers have far advantages in terms of machine-actionable features. Such examples enable citation tracking and aggregating, scientific production combination,

[1] 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.
[2] https://www.doi.org/
[3] https://www.handle.net/
[4] https://www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml

empowering authoritative, promote digital connections among researchers, organizations, and research outputs. To measure the impact of entities, also tracking the PID resolution, should be used. The research infrastructure like manuscript submission, grant application systems, and data discovery services is increasingly reliant on these connections.

Research infrastructures services have widened the PIDs assignment for its content and commonly adopt PID for identifying the datasets. Still, these research objects carry a high variation level in volume and internal units. Hence, assigning a PID to a whole dataset, as done with DOI through da|ra, is not enough to unambiguously identify the piece of information used and ensure the appropriate data citation and, consequently, the accreditation of research data. From a data citation reused perspective, identifying the most specific possible data unit is the way to ensure citation accurateness.

In FAIR terms, identifiers for inline data objects are an important step towards machine-actionable data, improving the findability and re-usability of data at the level of variables. Consequently, we suggest assigning PIDs to a lower granularity level, i.e., on variables in datasets. Our approach considers the data producers' and data users' best interests. Assigning PIDs to the variables will make research data easier to find and cite, which benefits both roles: data producers and re-users. PIDs are the backbone of a FAIR data infrastructure because they allow for referencing and cross-referencing data and metadata objects.

When citing data at the level of variables this citation enhances transparency, disambiguates data, favours reproducibility, assures provenance, and fosters reusability in the best sense of the FAIR principles. We aim to support data producers in assigning the PIDs for the variable units they have produced in their research. Consequently, the service will also add value to data users, who can efficiently discover and cite interesting data preserved in the long term.

The PID service aims identify the object, it does not intent to explain the content of the identified object. However, in the sense of reuse, researchers are much more interested in the content of the variables. To this end, a more significant effort is necessary to understand the variable concept meaning and its values. Researchers need to analyse data documentation exhaustively to find relevant variables for their research (Bensmann *et al.,* 2020).

The researchers seek variables concepts to evaluate if they are related to their studies' interests. To this end, when researchers have access to the dataset, they need to go a long way to check the adherence of such data. They need to understand the variable content (gender, income, etc.) or the variable concept. To assess the variable relevancy, researchers need: (1) discover and access the dataset URL; (2) check the dataset version; (3) get how to open the dataset (it can mean manage new software); (4) download the dataset; (5) find and read data documentation; (6) seek for the variable in the documentation to assess the content adherence; (7) open the dataset and find the variable and its values; (8) choose the values and (9) apply statistics before reusing data. Using the PID (even though it will not identify its contents per se) points directly to variable data, which shortens these steps.

Following are the concept description of the Social Sciences variables and their relevance to the field and how they relate to other subjects, unveiling the variable's multidisciplinary attributes.

1.1. Variables in the Social Sciences quantitative approach

A variable is, in a broader sense, a unit essentially able or apt to vary, is subject to variation or changes, it is an entity whose quantity may assume any one of a set of values (Merriam-Webster, n.d.). The variables are also related to many domains, such as Mathematics, Sciences, Statistics, and Social Sciences. The statistical approach determines the correlation among variables in a dataset and identifies misfit values. From the statistic standpoint, a variable is defined as "a characteristic of a unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned. Among these categories, some examples are "income, age, weight, occupation, industry, disease, etc. (Organisation for Economic Co-operation and Development, 2007).

The quantity value may have a specified set of values assigned in the mathematical sense. However, there are also non-measurable characteristics, e.g., a range of options from which some categories are taken as attributed values (The International Statistical Institute, 2003). The variable level in the Social Sciences research data is a unit of quantitative data, commonly obtained through survey questionnaires or experiments and represented as a column in a tabular data sets format.

Addressing the Social Sciences interest, the variable concept is pivotal to quantitative analysis and macro-structural sociology. Since the term refers to variance instead of constancy, it relates to fundamental characteristics that vary (elements as gender, age, race, social class) and influence behaviour or attitudinal variables from a Social Sciences quantitative approach. When measuring those variables, commonly through questionnaires, the researchers analyse their importance in understanding human behaviour, attempting to identify cause and effect relationships among the values' elements (Drislane; Parkinson, n.d.).

Consider the relevancy and the leading function of variables within a given study, particularly in the data reuse procedure, when another interesting part will totally or partially rely on the variable's results to make recommendations, provide inferences, and produce outcomes through secondary data analysis practices. How should we not assure a precise, unique, traceable, unambiguous, long-lasting, and undouble means of identifying them and consequently qualifying to an accurate data citation? These advantages are some of what PIDs can guarantee when assigned to the variable level.

Data citation procedure should satisfy citation principles, such as the Data Citation FAIR Principles (Data Citation Synthesis Group, 2014) and align with Data Citation Best Practices (Task Group on Data Citation Standards and Practices, 2013) to ensure a trusted research process. Our proposal to assign PIDs to the variable level aligns with recommendations and policies for data citation in the Social Sciences domain. The Social Sciences & Humanities Open

Cloud (SSHOC)[5] recommendations for FAIR Data Citation (Larrousse; Gray, 2021) emphasizes the dataset's level of variation and recommends identifying the granularity level accordingly, using unique identifiers.

Through its Persistent Identifier (PID) policy (European Commission, 2020) the European Open Science Cloud (EOSC) advocates that the PID ecosystem should support multiple granularity levels and corresponding PIDs assignments to fulfil the community's need as a best practice. A flexible PID Service would respond to those needs, supporting the increase in the efforts and expenses of managing new services' functionalities and components while raising PIDs volume. Furthermore, another relevant PID Policy, jointly supported by the Consortium of European Social Science Data Archives (CESSDA) and the European Research Infrastructure Consortium (ERIC) Persistent Identifier Policy (Hausstein *et al*., 2017), also advises assigning globally unique PID and embedding it in the data file (e.g., variable).

The following section describes the shortcomings of citation of variables along use cases and our solution by providing a metadata schema and architecture to register PIDs for variables to overcome the shortcomings.

1.2. Organization of this report

This report summarizes our proposal to assign PIDs to a more granular data level, i.e., a variable within a dataset. In section 2, we detail the measure's goal in the sense of da|ra widening services and we discuss how expanding the use of PIDs in terms of the FAIR principles. We address the specific tasks of the measure and highlight how these solutions can benefit future applications.

Section 3 describes use cases on data citation and data access by providing examples of how tricky it can be to find and cite research data that are not accurately identifiable, illustrating the consequences of PIDs' absence in data citation. Besides, it outlines the methodological aspects of our proposed approach for citing variables.

Sections 4 and 5 relate to the technical infrastructure. Section 4 delivers the general functionality of the PID service, the metadata elements needed for a PID registration of variables, and the requirements and standards of system architecture fields. In Section 5, the general architecture of the PID registration service is defined, the major components are depicted within the registration process, and present a detailed description of the registration stages through workflows.

Concludes within section 6 the architecture and requirements to implementing the variable registration service, scheduled according to the milestones planning into phases for September 2022 and September 2023, respectively.

[5] The SSHOC project aims to build the SSH (Social Science and Humanities) part of the EOSC (European Open Science Cloud).

Section 7 points to future service evolution possibilities and the corresponding metadata requirements. It also indicates the essential practices like testing and clean code quality standards and adequate software licenses.

2. PID Services for Variables

The GESIS - Leibniz Institute for the Social Sciences and the ZBW - Leibniz Information Centre for Economics jointly manage da|ra to offer the DOI registration service for social science and economic data in Germany. Da|ra's current service provides PID assignments based on DOI standards at the datasets and datasets collections levels.

The PID service for variable is conceived within the KonsortSWD project. It is the objective of the measure 1 in the Task Area 5, which goal is to widen the functionality of da|ra[6] and the use of PIDs to the level of attributes, which the variable is the first data format prioritized at the service extension. Widening da|ra, which currently focuses on the Social Sciences and Economics, means to open up registration of research data for all participating disciplines – particularly neighbouring disciplines of the Social Sciences (e.g., the humanities). Expanding the use of PIDs implies using them for attributes, first variables data format, within datasets.

The widening service will register a PID for variables is based on Handle Technology, developed by the Corporation for National Research Initiatives (CNRI[7]) and administered by the DONA Foundation[8]. When registering so many PIDs, as the variable level requires, the DOI standard would be costly due to its fees. However, the service is intended to be compatible with any PID Standard and, therefore, also DOIs if any institutional user wants to adopt it.

By assigning PIDs to the variables, individual elements of the data files can be referenced and retrieved with the required metadata. This assignment is crucial when users access data through an application programming interface (API) or Jupyter Notebooks, or when researchers re-use questions from existing surveys for their own research (which is frequently the case). The specific tasks of the Measure will be 1) to develop technical solutions for fine-grained referencing and identification at the variables level, suitable for scaling. These solutions need to be integrated into the existing services (in close cooperation with TA 1 Community Participation). They will enable not only citation, but also the permanent identification of data objects through APIs for data exchange. 2) To meet the increasing demand for interoperability, data mappings (i.e., translation between machine-readable formats) should use the same naming as the Data Documentation Initiative (DDI) standard. It structures metadata at the level of variables in a machine-readable way. To simplify and standardise this process, da|ra will be

[6] https://www.da-ra.de
[7] https://www.cnri.reston.va.us/
[8] https://www.dona.net/index

upgraded to handle PIDs on variable level to process the relevant metadata standards of the communities.

The issue of semantic ambiguity is particularly problematic due to the unclear referencing of the datasets and the variables, the lack of versioning details, sometimes only referring to the concept without clarifying what variable is being reused. Nonetheless, in the case of each variable within a given dataset version having its own PID, the reused variables could be disambiguated. In addition, each registered variable will link to the dataset DOI, and there the version of the published study is clear.

The following section presents some uses cases, demonstrating how variables are cited in publications and what problems arise from this practice in a non-standard means.

3. Use cases for PIDs on entity level

We identified two major use cases namely *data citation* for publications and *data access* for **variables,** which are enabled with the general function for *registering variables* with a persistent identifier. This persistent identifier can be cited in a paper or, e.g., used to gain data access within a script. The data access use case will not be provided within this measure, but needs to be enabled by each service provider, who wants to support data access to their research data on a variable level. Within measure T5-M3 such an initial, general data access API is developed based on the Digital Object Interface Protocol (DOIP) standard (DONA Foundation, 2018). To better understand the use cases, they are specified in more detail here (by using personas and providing examples).

3.1. Use case 1: data citation

Doris has written a paper with the new accumulation of the European Values Study – EVS (**ZA7500**) (EVS, 2020). *Doris* needs about 20 of the total 466 variables contained in the data file for her analysis. In order to provide transparency of her research she spent a lot of time and effort in the past to create an appendix with the question formulations and response categories for all variables she has used in the studies that underly her research papers, because unique and stable links to variable documentation have not been available. Now, given the variables registry provided by KonsortSWD, *Doris* needs to include only the URIs for the variables she has used in her new paper. By this, readers are enabled to directly and in a predictably and unambiguously way access the documentation of the variables used in the study. Further information can be linked from there, such as the original language version of the question which generated the variable.

The cited use case illustrates how challenging it can be to reuse data without adequate granularity level identification. In addition, Doris' reuse of data documentation will not enhance reproducibility as it could do if PIDs were available and used to cite the data. The entirety of Doris' effort to ensure transparency must be described and detailed explained in her results documentation. Nonetheless, any scrutinized attempt could fail due to the misunderstanding on how and which data exactly was reused.

3.2. Use case 2: data access to variable level

Ellen is working on a longitudinal study with data from the GESIS Panel. She only needs some variables from the extensive data file. Retaining measures of income during the Corona pandemic and satisfaction with the federal government are relevant to Ellen's evaluation. Both attributes were collected multiple times but not necessarily in the same GESIS Panel wave. She obtains the DOIs of the variables from the study and variable level documentation, e.g., GESIS Search[9] or any other study or variable documentation she needs and imports the data from the search or study documentation, including the metadata, into her analysis environment.

Ellen must add the specification of which waves the reused data refers to when providing a citation to reused data due that the same variable name can be found in multiple physical datasets, from different cohorts and years. Otherwise, it will not ensure the authenticity of the research data or facilitate access through waves and could undermine the data verifiability. Since the same variables were collected from multiple waves, ambiguous data citation probability rises. Sometimes scholars forget versioning relevancy or do not provide the most complete citation possible. Such behaviour collides data reuse and reproducibility in the Social Sciences research. Registering the variable as a PID goes beyond disambiguating them; it will strengthen trust in the provenance of research data since human mistakes are reduced while tracking data and making them easier to find and cite.

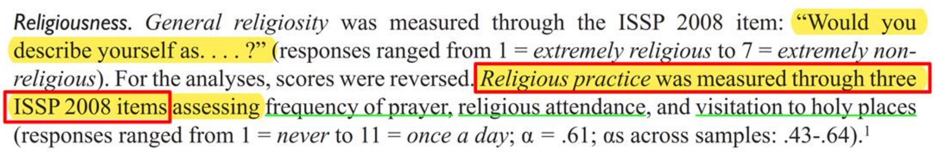
3.3. Problems of current data citation practices

As described in the use case on data citation, currently variables are cited “in text”, without a unique identifier, if at all. Usually only the study is cited. This makes it inefficient for the service provider to identify important variables and for the researcher to re-use variables. Following are some examples of reused data citations that address multiple problems in understanding data without a unique identifier.

3.3.1. Example 1: no citation

In the first example, the author reuses the ISSP 2008 data. Still, there is no appropriate citation on the variable(s). The only information refers to the three "items" (concepts) concerning the author reused data: (1) frequency of prayer, (2) religious attendance, and (3) visitation to holy places. These three concepts were aggregated in one measure - religious practice. Even though the author reproduces the question when addressing general religiosity ("would you describe yourself as"), there is no available transcript of any question regarding any one of those three items for the "religious practice" measure. The paper only mentions "three ISSP 2008 items", but it is unclear what those items are, and their variables related.

[9] https://search.gesis.org

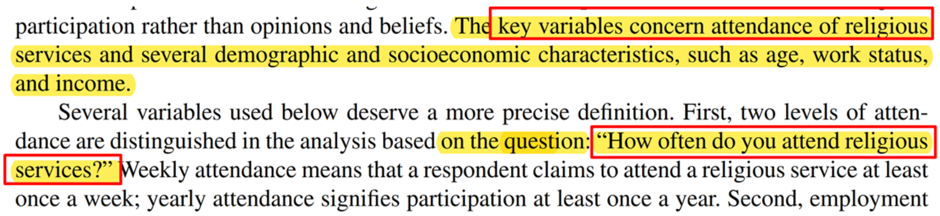


Religiousness. General religiosity was measured through the ISSP 2008 item: "Would you describe yourself as. . . . ?" (responses ranged from 1 = *extremely religious* to 7 = *extremely non-religious*). For the analyses, scores were reversed. *Religious practice* was measured through three ISSP 2008 items assessing frequency of prayer, religious attendance, and visitation to holy places (responses ranged from 1 = *never* to 11 = *once a day*; α = .61; αs across samples: .43-.64).[1]

Figure 1: Data citation example 1 – excerpt from Clobert *et al.* (2014).

3.3.2. Example 2: partially citing the questions

The author uses the question to identify the variable "attendance on religious services." Yet, the paper reused more variables as age, work status and income, but did not add citations for the others.



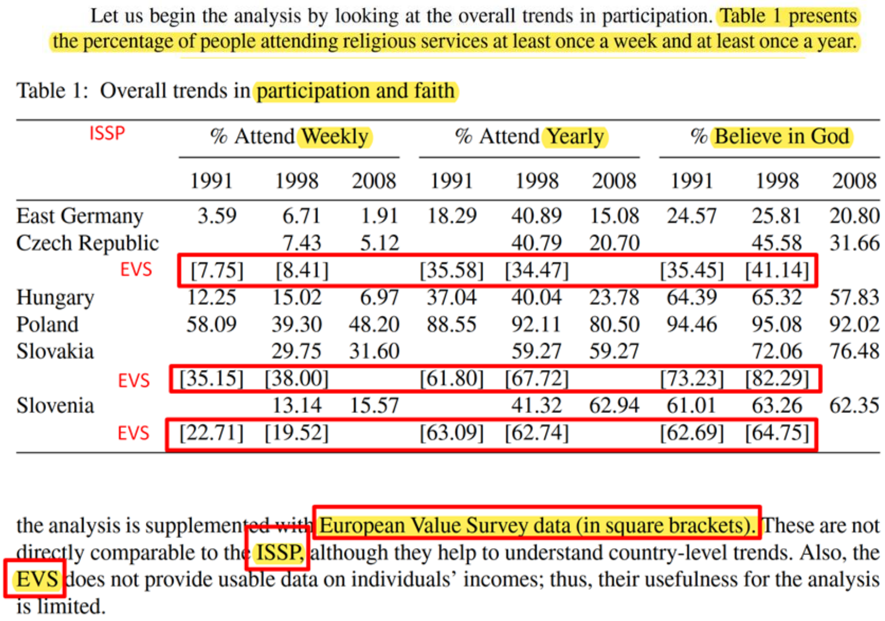
participation rather than opinions and beliefs. The key variables concern attendance of religious services and several demographic and socioeconomic characteristics, such as age, work status, and income.

Several variables used below deserve a more precise definition. First, two levels of attendance are distinguished in the analysis based on the question: "How often do you attend religious services?" Weekly attendance means that a respondent claims to attend a religious service at least once a week; yearly attendance signifies participation at least once a year. Second, employment

Figure 2: Data citation example 2 – excerpt from Minarik (2014).

3.3.3. Example 3: mixing data sources without citation

In the third example, the author uses two different datasets. The variables address attending religious services (variables from the ISSP dataset – and from EVS dataset (Attend Weekly, Attend Yearly and Believe in God). The author aggregates the three variables (ISSP plus EVS) under one concept he calls "participation and faith". The author differentiates the variables with a briefly information about using square brackets. The red marks were added to demonstrate the lack of sources details.



Let us begin the analysis by looking at the overall trends in participation. Table 1 presents the percentage of people attending religious services at least once a week and at least once a year.

Table 1: Overall trends in participation and faith

ISSP	% Attend Weekly			% Attend Yearly			% Believe in God		
	1991	1998	2008	1991	1998	2008	1991	1998	2008
East Germany	3.59	6.71	1.91	18.29	40.89	15.08	24.57	25.81	20.80
Czech Republic		7.43	5.12		40.79	20.70		45.58	31.66
EVS	[7.75]	[8.41]		[35.58]	[34.47]		[35.45]	[41.14]	
Hungary	12.25	15.02	6.97	37.04	40.04	23.78	64.39	65.32	57.83
Poland	58.09	39.30	48.20	88.55	92.11	80.50	94.46	95.08	92.02
Slovakia		29.75	31.60		59.27	59.27		72.06	76.48
EVS	[35.15]	[38.00]		[61.80]	[67.72]		[73.23]	[82.29]	
Slovenia		13.14	15.57		41.32	62.94	61.01	63.26	62.35
EVS	[22.71]	[19.52]		[63.09]	[62.74]		[62.69]	[64.75]	

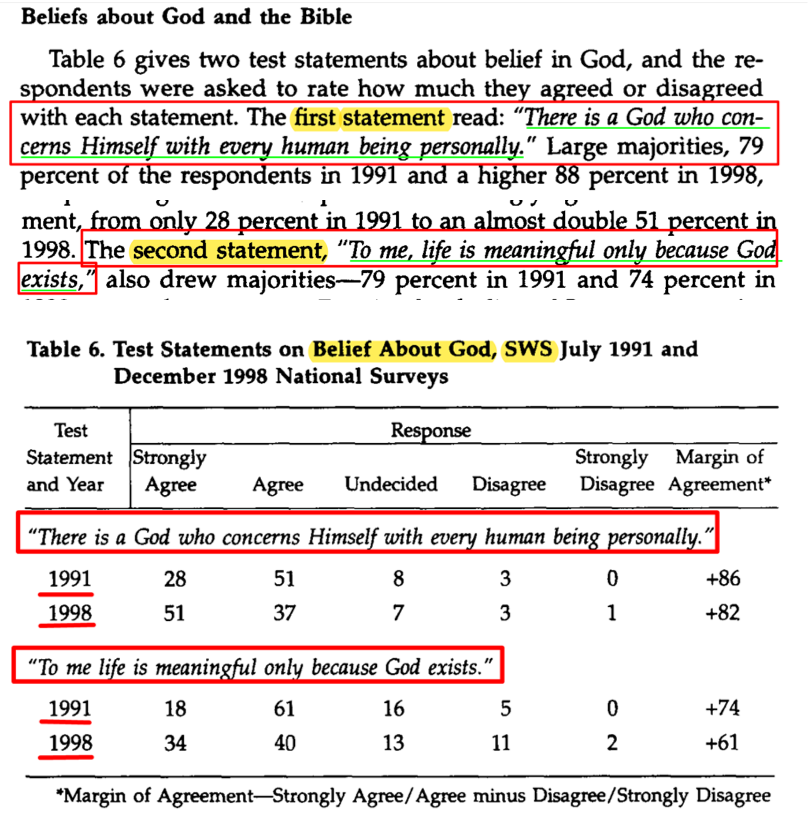
the analysis is supplemented with European Value Survey data (in square brackets). These are not directly comparable to the ISSP, although they help to understand country-level trends. Also, the EVS does not provide usable data on individuals' incomes; thus, their usefulness for the analysis is limited.

Figure 3: Data citation example 3 – excerpt from Minarik (2014).

3.3.4. Example 4: not referencing the question

This example highlights a common practice on data reusing that is more deeply concerned with transparency. This practice is when the author extracts variables from questions to compare or analyse them. In this case, the author excerpts one or more variables from a question and does not provide a suitable citation on which question was used. Since each variable is listed as an option within a range of answers from that question, looking manually into each question to find the variable is hugely time-consuming task.

The author of this example mentions general concepts to reuse questions and data from ISSP 1991-1998 waves. However, he only transcripts the questions' values when addressing the variables, skipping the question's description. There are many variables related to God within this dataset, which increase the ambiguity risk. In this example, finding the "statement" to check what variable it relates to, and the dataset's variable data is a double challenging because the paper analyse two waves. When a unique identifier is not present for each variable within each wave, there is no possibility to understand this granularity level without (again) scrutinizing the data file and searching the text.



Beliefs about God and the Bible

Table 6 gives two test statements about belief in God, and the respondents were asked to rate how much they agreed or disagreed with each statement. The first statement read: *"There is a God who concerns Himself with every human being personally."* Large majorities, 79 percent of the respondents in 1991 and a higher 88 percent in 1998,

ment, from only 28 percent in 1991 to an almost double 51 percent in 1998. The second statement, *"To me, life is meaningful only because God exists,"* also drew majorities—79 percent in 1991 and 74 percent in

Table 6. Test Statements on Belief About God, SWS July 1991 and December 1998 National Surveys

Test Statement and Year	Response					
	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree	Margin of Agreement*
"There is a God who concerns Himself with every human being personally."						
1991	28	51	8	3	0	+86
1998	51	37	7	3	1	+82
"To me life is meaningful only because God exists."						
1991	18	61	16	5	0	+74
1998	34	40	13	11	2	+61

*Margin of Agreement—Strongly Agree/Agree minus Disagree/Strongly Disagree

Figure 4: Data citation example 4 – excerpt from Abad (2001).

There are many variables related to God within this dataset. Table 1 provides a better understanding the question and variable internal complexity. According to its concepts (Q22a and Q22c), question 22 (Q22) holds at least two variables (v44 and v46).

VARIABLE	LABEL	QUESTION
v37	Closest to Rs belief about God?	Q.18 Please indicate which statement below comes closest to expressing what you believe about God.
v38	Best describes your beliefs abt God?	Q.19 Which best describes your beliefs about God?
v44	God concerns Himself with humans?	Q.22 Do you agree or disagree with the following ...**Q.22a There is a God who concerns Himself with every human being personally**.
v46	Life meaningful because God exists?	Q.22 Do you agree or disagree with the following ...**Q.22c To me, life is meaningful only because God exists.**
v71	Faith healers have God-given powers	Q.38 Now please think about something different. Please tick one box on each line below to show whether you think each statement is true or false. Q.38c Some faith healers really do have God-given healing powers.

Table 1: Variables related to the same concept in the dataset

Those are examples of how difficult it is to locate data without a specific citation. If there was PID assigned to each reused variable, the authors could apply it to cite the data properly. Without the PID, authors use various means such as access points to identify where they collected the data (e.g., questions, concepts, or even only the dataset general name). Those examples only highlight the current data citation practices when a PID is unavailable.

Our proposal addresses providing the most specific citation possible regarding variable use, by assigning a PID to each variable within a dataset. This more detailed citation approach builds trust in data provenance and fosters data findability and accessibility through the functionalities embedded in the PIDs services.

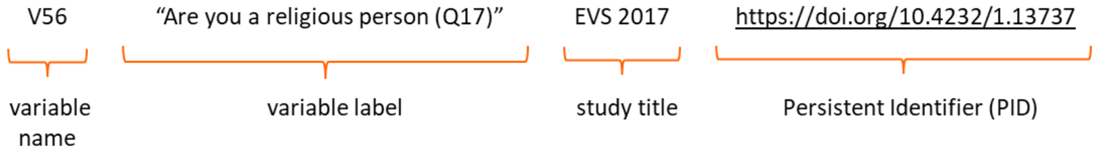
3.4. Proposed approach for citing variables

With the registering of variables, we suggest a specific citation scheme based on four metadata, as the following example: (1) the variable name, (2) the variable label, (3) the study title, and (4) the PID. In the example, the PID is a DOI standard. It aims to demonstrate the syntax of the PID and show how it points to the GESIS URL landing page[10].

[10] https://search.gesis.org/variables/exploredata-ZA7505_VarF034.

V56, "Are you a religious person (Q17)" in EVS 2017; https://doi.org/10.4232/1.13737.[11]

The citation scheme example is explained as follow:


V56
"Are you a religious person (Q17)"
EVS 2017
https://doi.org/10.4232/1.13737
variable name
variable label
study title
Persistent Identifier (PID)

The same PID can be used to provide *data access*. As an example, access in R can be given to the data of a variable by using the same PID as cited. The R code should run over the variable's PID to obtain access to the data:

```
variable <- ead.csv("https://access.gesis.org/data?pid=10.4232/1.13737&type=variable&fileType=csv")
```

A metadata scheme is required to comply with the minimum metadata necessary to provide a registration service to obtain PIDs.

4. Functionality and metadata schema

The major objective of the Measure 5.1 is to provide a PID service for variables. In the following we describe the general functionality of the PID service as well as the metadata scheme and the system architecture needed for registering a variable and assigning a PID to it.

4.1. Functionality of the PID service

The PID service for variables aims to register a variable and assign a PID to it. An infrastructure providing research data wants to enable the citation and possibly data access to a variable of a survey data set. Before requiring a PID, the data provider must be registered and authenticated logged in the registration service. As there can be many variables within a study, so an automated way to register PIDs as a bulk should be available. Through a script or integrated software in the infrastructure's documentation tool, all variables can be registered. The script registers the variables via the variable registry within da|ra. The infrastructure delivers a PID suggestion, a landing page, the original dataset DOI and some metadata to identify a single variable. The variable registry validates the metadata, confirms the registered PID and stores the metadata. The infrastructure adds the PID to each variable landing page for citation reasons.

As the service' users, the data providers can expect all necessary functionality and reliable service since a stable PID infrastructure supports the third-party (ePIC API) registration service (detailed in the section 5.2). Also, the registration service complies with high-quality

[11] EVS/WVS (2021). Joint EVS/WVS 2017-2021 Dataset (Joint EVS/WVS). GESIS Data Archive, Cologne. ZA7505 Data file Version 2.0.0, https://doi.org/10.4232/1.13737.

requirements and policies (Persistent Identifiers for eResearch, 2020), under which PIDs are maintained and an SLA[12] that standardizes the ePIC services, guaranteeing its continuation.

The registration service requires a set of minimum metadata to get a PID. A variable must be related to a study, a resource type, e.g., a dataset, and identifying variables information such as name and label is mandatory to obtain a PID from the registration service. Following, we introduce a metadata schema and a variable content sample to specify our metadata proposal.

4.2. Metadata Schema for variables

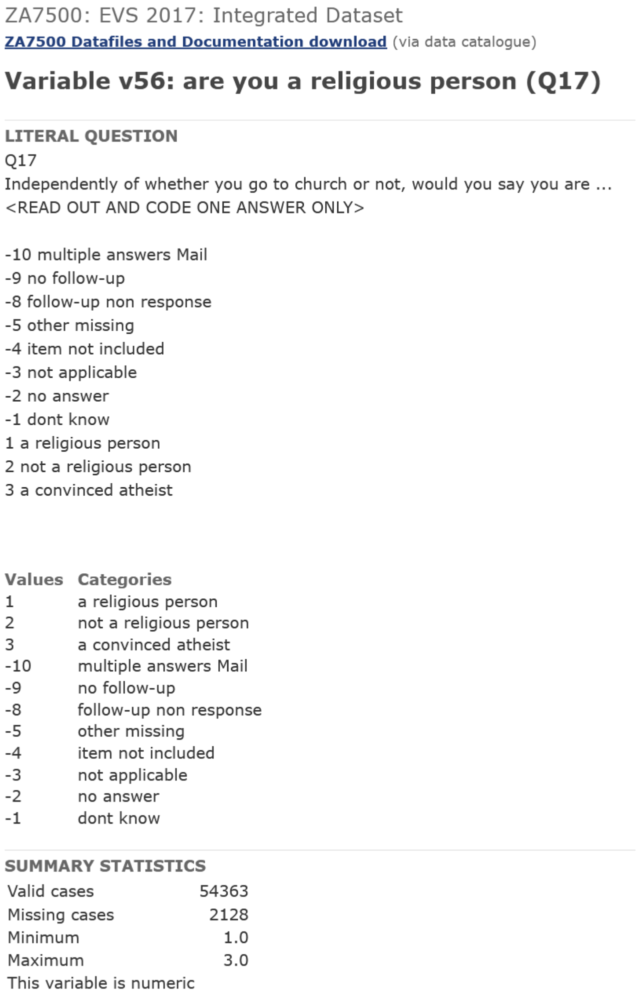
As a concept, a variable refers to some characteristic of an observed object. Variables are units of measure within a given survey, an interview script or an observation study. These measures can be a set of values to which a numerical mensuration or a category from a classification list can be set. There might be variables in answers to questions, administrative sources, or derived variables (e.g., age group derived from birth year) (World Wide Web Consortium, 2022).

Variables encompass re-usable parts like occupation classification, gender, age, and any type of organized and scaled information classes. Variable also embeds descriptions of its concepts and cohorts, linking their names to such content. Each variable in each study needs to be labelled, referring to its content definition and designating data type characteristics. The variables' concepts can also be linked to other variables. There is a complex interdependence between one variable and questions. The same variable can be related to many questions, and one question can be associated with many variables. Variables relate in various means across fields and domains, and for this study out interest is the Social Sciences approach.

Similar characteristics can group these measures, e.g., a data type such as numerical or textual values, or by any concept regarding the values, as ordered or random coded response domain. Also, the semantic approach can be used to the variable's content knowledge organization (Bensmann *et al.,* 2020).

A typical variable (depicted in Figure 5) consists of a variable name, here “v56”, a variable label, “are you a religious person (Q17)”. The label itself is usually a short version of the asked question. A variable is embedded within a Social Science questionnaire and overall, a study or survey. To register a variable and get a PID, some minimum metadata are necessary, such as name, label, and the corresponding study DOI, along with the standard metadata fields, e.g., published or creator. Further documentation, such as the question, answers, or descriptive fields within DDI (Data Documentation Initiative, 2020) are not necessary to obtain a PID.

[12] Service level agreement (SLA) registers the service level of a service provider to its customers and identifies their required expected level of service.



ZA7500: EVS 2017: Integrated Dataset

ZA7500 Datafiles and Documentation download (via data catalogue)

Variable v56: are you a religious person (Q17)

LITERAL QUESTION

Q17
Independently of whether you go to church or not, would you say you are ...
<READ OUT AND CODE ONE ANSWER ONLY>

-10 multiple answers Mail
-9 no follow-up
-8 follow-up non response
-5 other missing
-4 item not included
-3 not applicable
-2 no answer
-1 dont know
1 a religious person
2 not a religious person
3 a convinced atheist

Values	Categories
1	a religious person
2	not a religious person
3	a convinced atheist
-10	multiple answers Mail
-9	no follow-up
-8	follow-up non response
-5	other missing
-4	item not included
-3	not applicable
-2	no answer
-1	dont know

SUMMARY STATISTICS

Valid cases	54363
Missing cases	2128
Minimum	1.0
Maximum	3.0

This variable is numeric

Figure 5: Variable description in ZACAT (GESIS)

Table 2 describes the proposed metadata fields for registering a variable in a generic PID standard. All fields are mandatory, unless stated otherwise. Some fields are mandatory depending on the PID standard. Assuming that the service implementation can employ the DOI standard to assign the PIDs, some metadata fields required for DOI are likewise flagged in the metadata schema. Then fields are listed as required for assigning any PID as well as required for the DOI standard.

FIELD NAME	DESCRIPTION	FIELDS FLAGGED	EXAMPLE VALUE
Study DOI	Occurrence: 1. A DOI. The DOI identifies the study in which the variable to register appears.	Required for any PID Required for DOI Standard	10.7801/351
Variable Name	Occurrence: 1. A string. The name of the variable to register.	Required for any PID	HS021

FIELD NAME	DESCRIPTION	FIELDS FLAGGED	EXAMPLE VALUE
Variable Label	Occurrence: 1. A string. The label of the variable to register.	Required for any PID	Arrears on utility bills
PID Proposal	Occurrence: 0..1. A PID. The PID proposal of the variable to register. Generated automatically or provided by the user		
Landing Page	Occurrence. 1. A URL. The landing page of the variable to register, via which a user may obtain detailed information about the variable to register.	Required for any PID Required for DOI standard	https://search.gesis.org/variables/exploredata-ZA6644_Varp6de
Resource type	Occurrence: 1. A string. The type of the resource.	Fixed. Internal. Required for any PID Required for DOI standard registration	Variable
Title	Occurrence: 1. A string. The title of the resource to register, which is generated by the registration service using the variable name and the variable label.	Fixed. Internal. Required for any PID Required for DOI standard	HS080: Do you have a colour TV?
Creators	Occurrence: 1-n. A string. The creators of the variable to register. The creators are the main researchers involved. Note, the creators may be different from that of the study. Multiple entries are possible. May be a corporate/institutional or personal name.	Required for any PID Required for DOI standard	Hugo Hugmann; Christoph Christ; GESIS

FIELD NAME	DESCRIPTION	FIELDS FLAGGED	EXAMPLE VALUE
Publisher	Occurrence: 1. A string.	Required for any PID Required for DOI standard	GESIS Data Archive
Publication date	Occurrence: 1. A date. Format: YYYY-MM-DD. The date of the publication of the study in which the variable to register appears.	Required for any PID Required for DOI standard	
Availability	Occurrence:1. A string from CV. The conditions governing the access to the resource.	Required for any PID Required for DOI standard	Download
Description/ Abstract	Occurrence. 1. A Text. Could be question text, answers, etc?		

Table 2: Proposed metadata fields for registering a variable

Labelling: Fields flagged with

- "Fixed" cannot be changed by the user.
- "Required for PID" in general for any PID.
- "Required for DOI" by DataCite to register a DOI standard.
- "Internal" are not required to be filled in by the user.

5. Architecture of the PID registration service

In the following, we will describe the general architecture of the PID registration service, and its individual components as shown in the architecture diagram (see Figure 6: Architecture overview) in more detail.

5.1. System's architecture

There are two major components within the system architecture. The aim is, that any service providers can register an arbitrary number of variables (bulk registration) through one REST API endpoint using a REST client. The first component provides the server-side REST API, whereas the second, backend component, the registration service, handles the correct and verified registration process.

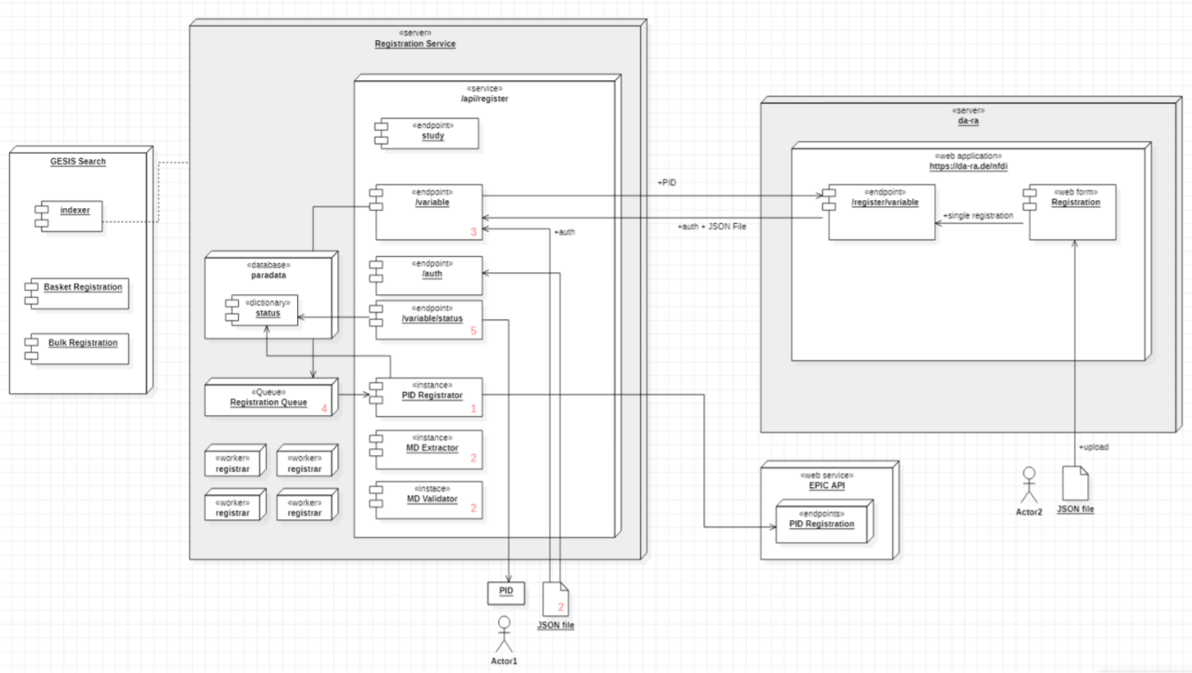


Figure 6: Architecture overview

Following the architecture and concepts are explained as well as functionalities to the variable registration requirements.

5.2. General description of the architecture and concepts

The diagram shows the "Registration Service", the central component of the architecture. It is designed to be a standalone service that can perform any registration processes via its main endpoint /api/register. The endpoint /variable is the main endpoint that handles requests for variable registration purposes.

Functionalities belonging together are bundled into components of the service, for example, authentication (*auth*-component), metadata extraction (*MD Extractor* and *MD Validator*), and registration (*PID Registrator*). Since variable registration requires metadata to comply with a pre-defined schema, MD Extractor & Validator are specific *instances* of a validator and an extractor. Abstractly implementing this enables to be open for future enhancements, for example, adding other endpoints to register different resource types (Datasets, Studies), and to switch between implementations on a per request basis (prospect).

The same applies to *PID Registrator*. It is a specific registration instance, which uses a particular third-party registration service and its defined endpoints to register PIDs. We chose the Persistent Identifier Consortium for eResearch[13] (ePIC) API[14] for that because it already features

[13] ePIC is an international consortium provides a reliable Handle-based PID infrastructure for research data. ePIC has currently nine members and it is open for any center that stores scientific/research data. See also: http://www.pidconsortium.net/.

[14] https://doc.pidconsortium.eu/docs/

all necessary functionality and is well supported. In contrast to the variable-specific *MD Extractor* and *MD Validator* instances, requiring resources respecting a particular schema, the ePIC API is not tight to a specific resource type to obtain a PID.

In general, the architecture of the Registration Service is designed to be asynchronous. For that, we employ a *Registration Queue*, where single variable registration jobs are put. Employing a queue has several advantages:

1. The queue decouples the workload of our registration service from the actual PID registration process of the third-party service (ePIC API). Thus, it makes it asynchronous and more scalable;
2. It enables "bulk registration", which means the registration of many variables at once. The size of the queue (i.e., how many requests can be held by the queue) as well as the number of workers (i.e., parallel registrations) should be a matter of configuration. Although "bulk registration" will not have high priority in the first implementation phase, we designed the architecture like this as prospect. We will initialize the queue with only one worker in the beginning;
3. The queue is designed to be persistent; thus, tasks (i.e., registration requests) that are put on the queue can be restored and will not get lost in case of a service failure.

The actual implementation of the queue is still a matter of choice. Generally, the queue can be implemented as an external service, e.g., with Apache Kafka[15] or Redis[16]or a part of the registration service itself, while it is implemented via a Spring Boot message queue.

Utilizing a registration queue requires storing the status of a request in an extra database. Along with the status of the variable registration, we store all other metadata of the submitted variable there, for example, the time of registration, the user performing the request, and more. The user may query the status via the */*api*/*register/variable/status endpoint.

5.3. Description of the Workflows

In the following, we will describe the workflow, starting from a user, here a data provider or a Research Data Center (RDC), which wants to obtain a persistent identifier (PID) for their variables. We will refer to the steps shown in Figure 7: Workflow diagram, which represents the variable registration service process.

[15] https://kafka.apache.org/
[16] https://redis.io/

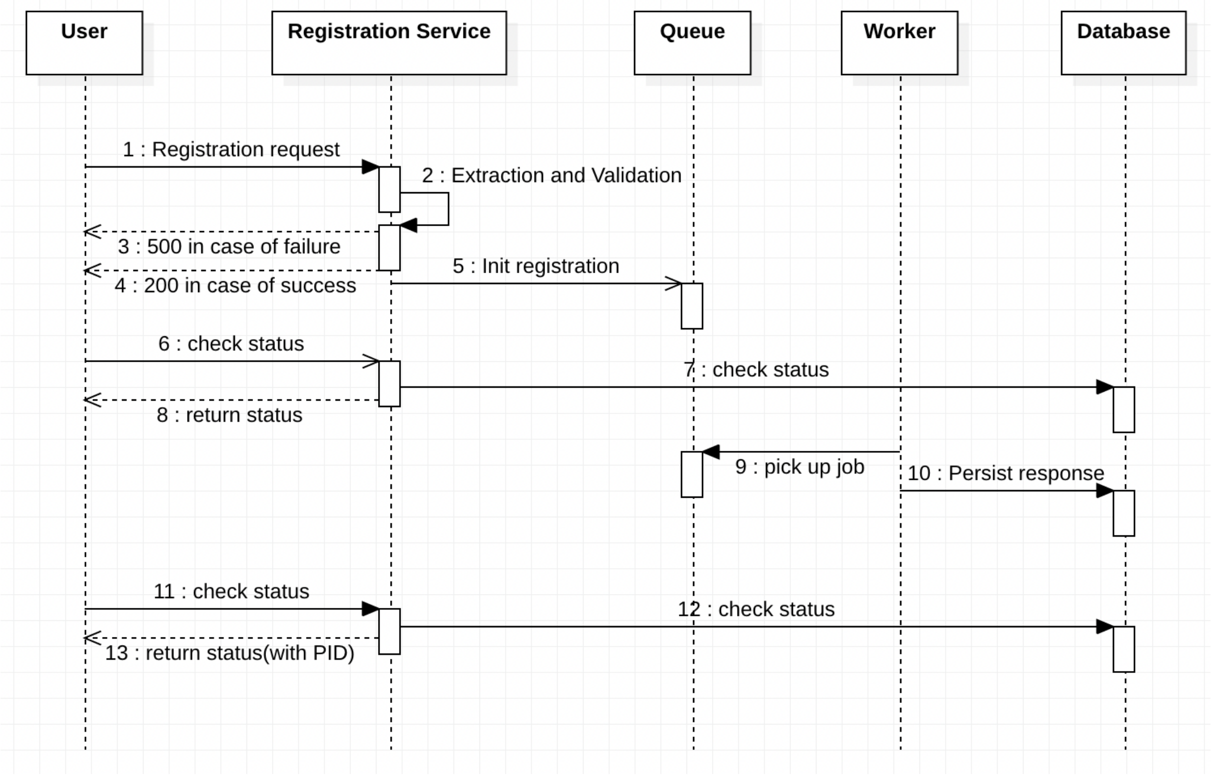


Figure 7: Workflow diagram

Access to the service is provided via token-based authorization (see *auth*-component in Figure 6: Architecture overview), which the user must obtain first. A valid token can be re-used for a given amount of time. To get a fresh token, the user needs to authenticate via the /api/register/auth endpoint beforehand. The general user management will be done via the central GESIS Login[17].

To register a variable, the user needs to send a request to the /api/register/variable endpoint (step 1, *Registration Service*). The request contains a JSON file with the required fields for variable registration alongside the authentication token mentioned before. The required metadata fields are listed in the section 4.2 Metadata Schema for variables. The response returns a status code 200, if the service is available (step 3), and the registration request was successful, 500 and an error message otherwise (step 4). The response does not yet return the PID registered, since our service is designed to be asynchronous. The final PID can be obtained by requesting the status via the /api/register/variable/status endpoint (step 6 and 11).

After receiving the initial request, the component extracts, and transforms (MD Extractor) all data into an internal data structure, based on which a structural and content-based validation (MD Validator) takes place (step 2). Structural validation checks whether all required fields are present (see Section Metadata). Content-based validation checks, for example, whether the provided study DOI exists or whether the fields are in the right format (e.g., publication date).

[17] https://login.gesis.org

If validation fails, the user receives a negative response code with an error message and the registration is not forwarded (step 3).

After validation the registration request is placed on to the registration queue (step 4). A worker may pick up the registration request (step 9). Here, the term "worker" is used as a wrapper-term and refers to the component that takes care of the actual PID registration (see section PID Registrator in Figure 6: Architecture overview). This step finalizes the process of registering a PID.

6. Conclusion

This report concludes the milestone 2 within measure 5.1: PID Service for variables. In this report we present use cases, the description of the general architecture, based on da|ra as blueprint, and the metadata schema for variables. This report will be the main reference for the implementation of the variable registration service. The architecture itself and the metadata schema were in addition modelled in such a way, that further entities such as questions from questionnaires could be registered.

It is planned to implement each of the following services according to the milestone planning within milestone M3 as prototypical implementation alongside an evaluation of the complete PID concept until September 2022 and move the implementation in milestone 4 into a productive and sustainable infrastructure by the end of September 2023. All services are modelled as generic software components, where applicable. In addition, all software components will be implemented with testing and clean code quality standards by GESIS and will be published under the Apache license.

7. Future Work

The service is flexible and allows other data formats, including new attribute levels at the further stages, such as surveys' questions, audio and video data fragments, a selected part of an image, or any digital object elements that register data throughout the various knowledge subject fields. In this sense, further applications of the PID service at other attributes than variables require metadata fields to be foreseen accordingly. This extension of da|ra can be handled as a blueprint, including other subject-specific metadata standards (e.g., from experimental research, text-based research, and big data research). Some metadata fields examples that should be envisioned are any digital objects components, inheritance, multilingual and related fields as concept classification and concept classification versions. Regarding best practices, besides the extensive software documentation for open-source publication, this service will provide the technical documentation on assigning PIDs. All code produced within this measure will be made available as open-source software.

8. References

Abad, Ricardo G. (2001). Religion in the Philippines. *Philippine Studies*, v. 49, n. 3, p. 337–367.

Bensmann, F., Papenmeier, A., Kern, D., Zapilko, B., & Dietze, S. (2020). Semantic Annotation, Representation and Linking of Survey Data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12378 LNCS*, 53–69. https://doi.org/10.1007/978-3-030-59833-4_4.

Clobert, M., Saroglou, V., Hwang, K. K., & Soong, W. L. (2014). East Asian Religious Tolerance—A Myth or a Reality? Empirical Investigations of Religious Prejudice in East Asian Societies. *Journal of Cross-Cultural Psychology*, *45*(10), 1515–1533. https://doi.org/10.1177/0022022114546641.

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles*. Martone M. (ed.) San Diego CA: FORCE11. https://doi.org/10.25490/a97f-egyk.

Data Documentation Initiative. (2020). *DDI Lifecycle 3.3:* Variable. https://ddialliance.github.io/ddimodel-web/DDI-L-3.3/item-types/Variable/.

DONA Foundation. (2018). *Digital Object Interface Protocol Specification*. https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf.

Drislane, Robert; Parkinson, Gary. (n.d.). *Social Science dictionary*. Athabasca University, Canada. http://bitbucket.icaap.org/dict.pl?alpha=V.

Estevão, Janete Saldanha Bach. (2019). *Information Literacy for data reuse in Social Sciences in Virtual Research Environments*: proposal of requirements and competences. 2019. Thesis (PhD in Technology and Society) - Universidade Tecnológica Federal do Paraná, Curitiba, 2019. http://repositorio.utfpr.edu.br/jspui/handle/1/4675.

European Commission. (2020). *A Persistent Identifier (PID) policy for the European Open Science Cloud*: Report from the European Open Science Cloud FAIR and Architecture Working Groups. EOSC Executive Board (ed.). October, 2020. 20p. doi: 10.2777/926037.

EVS (2020). European Values Study 2017: Integrated Dataset (EVS 2017). *GESIS Data Archive, Cologne. ZA7500 Data file Version 4.0.0, https://doi.org/10.4232/1.13560.*

Hausstein, Brigitte, Borschewski, Kerrin, Jerlehag, Birger, van Horik, René, & van der Vaart, Lilian. (2017). *CESSDA ERIC Persistent Identifier Policy (1.0)*. Zenodo. https://doi.org/10.5281/zenodo.3611317.

Merriam-Webster. (n.d.). *Variable*. In Merriam-Webster.com dictionary. Retrieved November 15, 2021. https://www.merriam-webster.com/dictionary/variable.

Minarik, P. (2014). Employment, wages, and religious revivals in postcommunist countries. *Journal for the Scientific Study of Religion*, 53(2), 296–315. https://doi.org/10.1111/jssr.12113.

Nicolas Larrousse; Edward J. Gray. (2021). *Recommendations for FAIR Data Citation in the Social Sciences and Humanities*. Zenodo. https://doi.org/10.5281/zenodo.5361718.

Organisation for Economic Co-operation and Development (OECD). (2007). *Glossary of statistical terms*. p. 836. https://ec.europa.eu/eurostat/ramon/coded_files/OECD_glossary_stat_terms.pdf.

Persistent Identifiers for eResearch. (2020). *ePIC Quality of Service and Policies*. http://www.pidconsortium.net/?page_id=904.

Task Group on Data Citation Standards and Practices, C.-I., (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, 12, pp. CIDCR1–CIDCR7. DOI: http://doi.org/10.2481/dsj.OSOM13-043.

The International Statistical Institute. (2003). *The Oxford Dictionary of Statistical Terms*. Yadolah Dodge [ed]. Oxford University Press, 2003.

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

World Wide Web Consortium (W3C). (2022). *DDI-RDF Discovery Vocabulary*: A vocabulary for publishing metadata about data sets (research and survey data) into the Web of Linked Data. Variable and Variable Definition. https://rdf-vocabulary.ddialliance.org/discovery.html#variable-and-variable-definition.

Kontakt:
Dr. Claus-Peter Klas
GESIS – Leibniz Institute for the Social Sciences
Knowledge Technologies for the Social Sciences - KTS
Unter Sachsenhausen 6-8, D-50667 Cologne
www.gesis.org
claus-peter.klas@gesis.org
Tel.: +49 (0)221 - 47694 – 520

KonsortSWD is funded within the framework of the NFDI by the German Research Foundation (DFG) - project number: 442494171.

doi: 10.5281/zenodo.6397367