Breaking Down Hurdles of Current Data Citation Practices. Use Cases and Benefits of Persistent Identifiers for Dataset Elements

Janete Saldanha Bach, Claus-Peter Klas, Peter Mutschke

GESIS – Leibniz Institute for the Social Sciences

The paper introduces a service to assign Persistent Identifiers (PIDs) on the level of the inline data objects of a dataset, such as survey variables in the Social Sciences, resulting from the consortium KonsortSWD of the German National Research Data Infrastructure (NFDI). This technical solution aims to make data findability and accessibility on the lower granularity level of studies more efficient. In the Social Sciences, for instance, PIDs are commonly available on the study level, which is insufficient to unambiguously identify the dataset elements used in a paper and ensure an accurate data citation. By assigning PIDs to the fine-grained level of attributes, individual dataset elements can be referenced and retrieved with the required metadata. Referencing research data and their inherited entities by PIDs supports FAIR data usage, i.e., research data can be Findable, Accessible, Interoperable and Reusable. Textual data citations without a unique identifier are non-standard practices that lead to considerable time-consuming problems in unambiguously identifying relevant elements of a dataset and reusing them. From the technical perspective, it also hinders automated access to data elements below study level. Our PID service simplifies FAIR data management and benefit both researchers and research data centres (RDCs), fostering credibility results and ensuring the sustainable reusability of data. RDCs directly benefit from PIDs as they enable citation tracking and impact measurement, linking articles using the same dataset elements. It empowers the RDC's authority by demonstrating a commitment to best practices, enhancing its reputation in the research community by adopting recommendations to support PIDs at multiple granularity levels, such as the European Open Science Cloud (EOSC) PID policy. Furthermore, it promotes digital connections among researchers, organisations, and research outputs. Explicit relations between those elements are possible and favour the formation of a network into a knowledge graph representation. Since PIDs are machine-actionable, they are the technical bridges to the FAIR principles as they increase the traceability of research results.

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18063

1 Introduction

Persistent identifiers (PIDs) are the backbone of FAIR data infrastructures as they enable a reliable data citation. FAIR stands for the Findability, Accessibility, Interoperability, and Reusability of research data (Wilkinson et al. 2016). However, in many cases PIDs are only available on study or dataset level but not on the level of the inline data objects that are usually used by researchers. In the Social Sciences, for instance, survey datasets usually contain hundreds of so-called variables but usually only a few of them are used in a research article, making it difficult to clearly identify and reference the variables used, as PIDs are only assigned on study level and the variables used are described in various, semantically often ambiguous textual forms. Moreover, researchers may cite the data provider or papers referencing the data instead of the data itself, or they place footnotes, image captions, or acknowledgments, rather than locating citations in the reference lists (Gregory et al. 2023). Moreover, researchers from various fields often do not follow any standard, such as the Data Citation Principles (Data Citation Synthesis Group 2014).

These inconsistent data citation practices create significant challenges in reliably identifying and reusing the data underlying a paper. Thus, current data citation practices often lack unique identifiers for relevant dataset elements, making it difficult for researchers to cite their data in a reliable way, for other researchers to reuse the data and for data providers to identify and annotate the important elements of datasets.

This paper introduces a PID service resulting from the consortium KonsortSWD[1] of the German National Research Data Infrastructure (NFDI[2]). The primary objective of this service is to enhance the reusability and findability of data by focusing on a more detailed level of datasets. The service assigns PIDs to specific, fine-grained dataset elements which represent the primary entities of research, such as survey variables in the Social Sciences, making it easier to reference and find relevant data objects. The PIDs are retrieved with the necessary metadata, facilitating both machine-actionable and human access. This metadata provides essential information about the data element, enabling users to understand its context and relevance. By incorporating PIDs and metadata, it ensures that users can effectively comprehend and utilize the retrieved information, in a more efficient and user-friendly way to manage and access data at a granular level. Hence, this detailed citation approach ensures data provenance, findability, and accessibility, fostering trust and promoting efficient data reuse.

The paper demonstrates how the service can simplify FAIR research data management at lower granularity levels, in section 2. Section 3 explains the PID registration service provider and the process of assigning PIDs for dataset elements. Section 4 details different social science for assigning PIDs to dataset elements below study level, and section 5

1 KonsortSWD (Consortium for the Social, Behavioural, Educational and Economic Sciences) is funded by the National Research Data Infrastructure (NFDI). KonsortSWD Homepage: `https://www.konsortswd.de`.

2 German National Research Data Infrastructure (NFDI) Homepage: `https://www.nfdi.de`.

concludes with the key contributions of the services for researchers and research data centres (RDCs).

2 PIDs simplify FAIR research data management at lower granularity levels

A PID is a persistent, unique, and globally resolvable identifier based on an openly specified PID Scheme, which allows for reliable and lasting reference to the associated research outputs (European Commission. Directorate General for Research and Innovation and Board. 2020). PIDs serve as the foundation for the long-term referencing of scientific publications, ensuring the consistent identification of digital objects. In the context of Social Sciences, for instance, research outputs have a range of granularity levels. The study and dataset represent the most common granularity levels identified with Persistent Identifiers (PIDs) when researchers publish their findings. However, datasets on survey data in the Social Sciences typically comprise questions, variables, variable values, indicator values or scales, and researchers are interested in the content of the variables. To this end, a more significant effort is necessary to understand the variable content meaning and its values. Figure 1 depicts granularity levels of research data PID are commonly used.

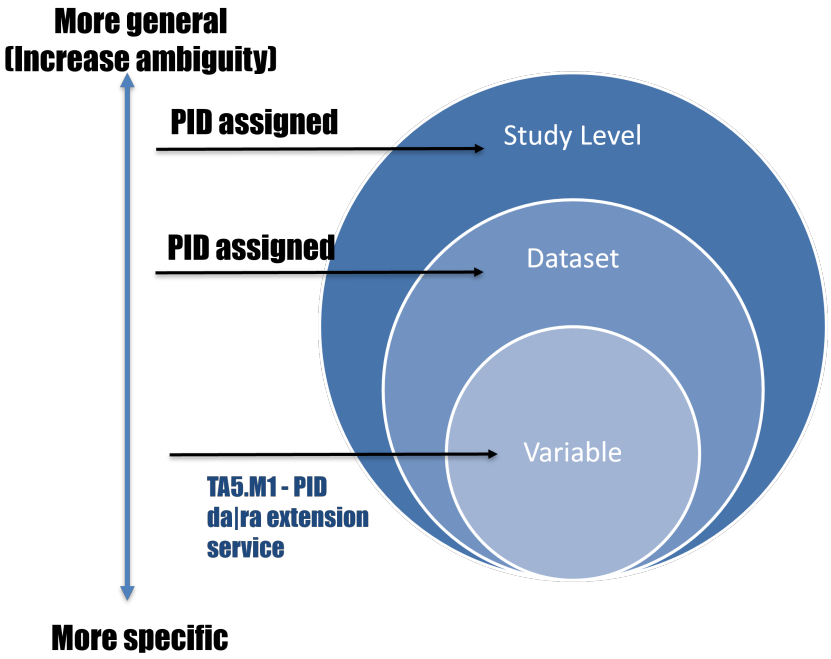

Figure 1: The Research data granularity levels (for the case of survey data in the Social Sciences).

In terms of data citation, once researchers locate and obtain a dataset of interest, they experience a long and complex process to extract the most relevant information and have to analyse data documentation exhaustively to find relevant dataset elements for their

research (Bensmann et al. 2020). In the following example we considered the case of a survey variable in the Social Science as the relevant dataset element in question. If a PID at the that level is unavailable, researchers also must:

1. Locate data citations in the paper: Researchers must identify data citations in the text of relevant studies, examining citations, quoted questions, or other hints such as websites or dataset provider institutions.
2. Identify and access the data source: After finding an interesting dataset, researchers should locate and access it on the data provider's website.
3. Verify dataset version: Researchers must ensure they are using the correct version, as some studies may cite earlier dataset versions.
4. Review data documentation: Users should examine data documentation and draw inferences.
5. Find matching dataset elements in the documentation: Users must identify variables that correspond to the referenced data in the paper.
6. Obtain dataset access: Researchers need to familiarize themselves with access rules (open, limited, or restricted/sensitive) and apply for access if necessary.
7. Learn how to open dataset files: Users must determine the appropriate software or driver for opening files, based on the file format.
8. Download the dataset: After meeting all requirements, users can download the dataset for further use.
9. Open the dataset and identify elements and their values: Users can manually locate dataset elements or use statistical software commands, depending on the file type (CSV, spreadsheets).
10. Apply statistical analysis: Users must understand dataset elements values and analyze the information accordingly.
11. Reuse variables: Users should generate new insights and alternative analyses using the same dataset.
12. Cite the data: Without a PID, users may cite the dataset name, provider name, or a report or study where the dataset elements were published, continuing the non-standard citation cycle.

Figure 2 illustrates the process of accessing and reusing dataset elements (here, survey variables) without PIDs. In contrast, assigning PIDs to identify dataset elements will simplify FAIR data management at lower level in three aspects:

1. boosting subsequent citation,
2. getting direct (meta)-data access, and
3. promoting data reuse.

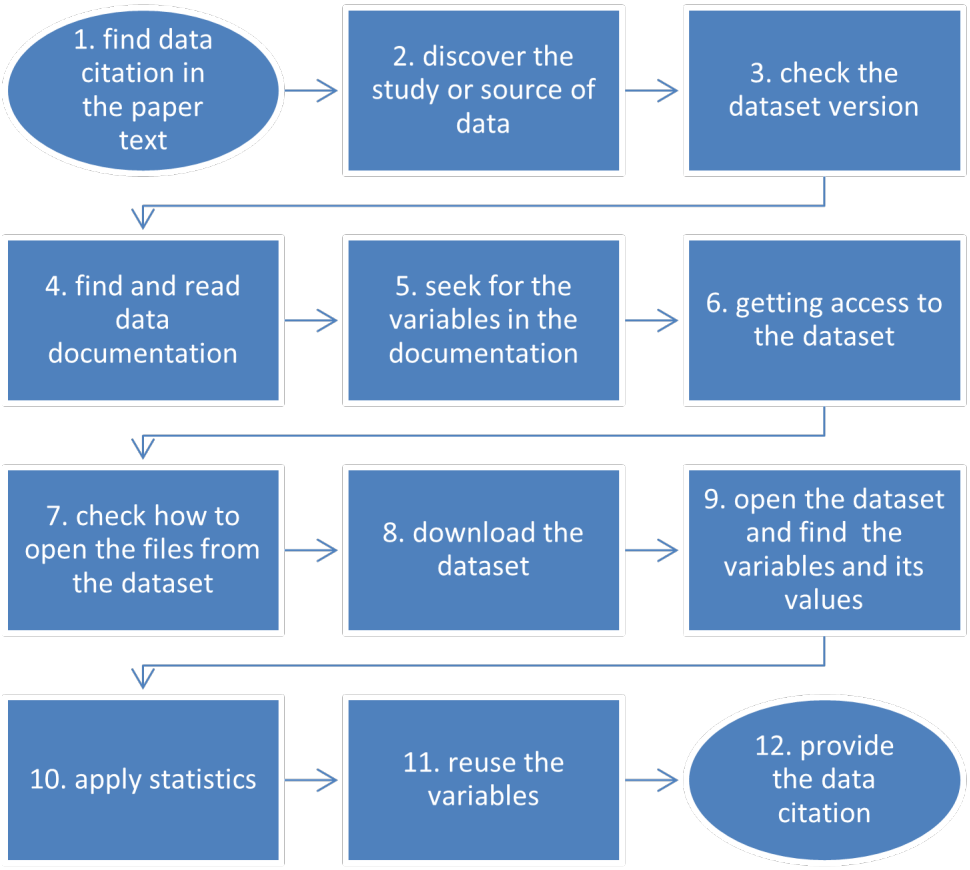

Figure 2: Steps to be taken for accessing and reusing dataset elements, here survey variables, without the availability of PIDs at dataset element level.

If one PID is registered for each element, it can (1) boost subsequent citation. In this case, automated scripts (do-files, R scripts, etc.) can be applied (2) get direct (meta)-data access, obtaining the selected element from the dataset automatically. The automatic access to the data in a dataset is enabled by just executing a script (i.e., using R, Stata, SPSS) that resolves a given PID and returns the data "behind" the PID in a proper format. One of the most common data formats in the Social Sciences are surveys and questionnaires results, a tabular dataset frequently stored in statistical packages files such as R or Stata, as well as spreadsheets or comma-separated values (.csv) files. Variables are distributed as rows (which contain objects) and columns (which contain properties) within datasets.

Some conditions are required to automatic access, which relies on the dataset's PID (typically a DOI) to retrieve the data. However, if multiple datasets are associated with the same DOI, this method will fail as the script cannot distinguish which dataset to access. And there are DOIs registered for dataset collections. For the automatic access

function to work effectively, only one dataset must be registered per PID, which is used to register the PID for a variable.

The dataset might have not restricted or closed access, and a REST API is available (Klas and Hopt 2022). Once the technical requisites are met, these automated scripts can technically give access to the dataset elements for direct usage without downloading the entire data file, either the complete dataset, but singled-out elements using PIDs (Klas, Saldanha Bach, and Mutschke 2023). The following workflow (Figure 3) depicts the process of accessing and reusing elements with PIDs. Researchers can take advantage of these machine-actionable features when a dataset element is identified with a PID. Getting data through automated access is faster and (3) promote data reuse, going through fewer steps if a PID is unavailable, compared to Figure 2.

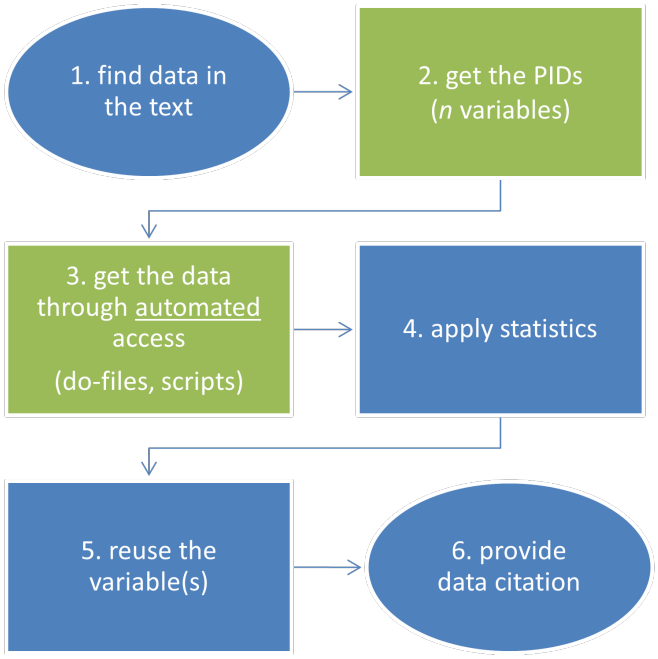

Figure 3: Accessing and reusing dataset elements with a PID.

3 PID registration

The PID service developed in the context of KonsortSWD is a technical solution aiming to make data findability and accessibility on the lower granularity level of studies, here survey variables in the Social Sciences, more efficient. To use the service, data holders (such as research data centers) must be registered in advance and authenticated within the PID registration service. Since a study may contain numerous dataset elements, an automated method for bulk PID registration is available. Using a script or integrated software within the documentation tool, all elements can be registered within the service. The request of PIDs is a task for data providers. In order to get as many PIDs as needed, the data provider must submit in the registration service a minimal set of metadata (Bach, Klas, and Mutschke 2023), including the suggested PID, landing page, original dataset PID

(commonly, a DOI), and other relevant metadata fields to identify each dataset element. The registration service then validates the metadata, confirms the registered study PID, and stores the metadata. Finally, the data provider includes the PID on each dataset element's landing page for citation purposes. As many variables exist within a study, an automated way to register PIDs as bulk is available. All variables can be registered through a script or integrated software in the infrastructure's documentation tool, which means the registration of many variables at once. To this end, any data provider can register an arbitrary number of variables (bulk registration) through one REST API endpoint using a REST client. This process automates the registration, avoiding much work for the data provider. See the first service report (Klas et al. 2022) for details.

The PID registration for lower-level elements assign Handle[3] PIDs, supported by the third-party Persistent Identifier Consortium for eResearch (ePIC)[4] API[5] registration service. The system is conceived to provide a general, maintainable, and scalable infrastructure that enables the registration of PIDs to the level of attributes.

4 Use cases in the Social Sciences

In the following we discuss use cases in the Social Sciences collected in the context of KonsortSWD, demonstrating the benefit of having PIDs on dataset element level. The PIDs registration service benefit research data centres (RDCs), fostering credibility results and ensuring the sustainable reusability of data. RDCs directly benefit from PIDs as they enable citation tracking and impact measurement, linking articles using the same dataset elements. It empowers the RDC's authority by demonstrating a commitment to best practices, enhancing its reputation in the research community by adopting recommendations to support PIDs at multiple granularity levels, such as the European Open Science Cloud (EOSC) PID policy (European Commission. Directorate General for Research and Innovation and Board. 2020). Furthermore, it promotes digital connections among researchers, organisations, and research outputs. Explicit relations between these elements are possible and favour the formation of a network into a knowledge graph representation. The agreed use cases are selected partners institutions participating in the KonsortSWD and play an essential role in shaping the service, testing the concept, and providing helpful feedback on the RDC daily activities associated with the Persistent Identifiers. The following use case descriptions are the Higher Education Analytical Data System (HEADS)[6] project from the German Center for Higher Education Research and

3 Handle System technology resolves PIDs such as Handles and DOIs. The Handle System was developed by Corporation for National Research Initiatives (CNRI) and is currently administered and maintained by the DONA Foundation. Handle.Net Registry (HNR) Homepage: `https://www.handle.net`.

4 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. ePic Homepage: `http://www.pidconsortium.net`.

5 ePic documentation Homepage: `https://doc.pidconsortium.eu/docs`.

6 The German Center for Higher Education Research and Science Research (DZHW) Homepage: `https://www.dzhw.eu/gmbh/index_html`.

Science Studies (DZHW); the GESIS Search[7] from the Leibniz Institute for the Social Sciences (GESIS), including the GESIS harmonisation tool: QuestionLink[8], the German Socio-Economic Panel (SOEP-Core) (Liebig et al. 2022), a longitudinal study from the German Institute for Economic Research (DIW), and the Qualiservice[9] qualitative data collection from the University of Bremen.

4.1 HEADS project from the DZHW

For the Higher Education Analytical Data System (HEADS) project at the German Center for Higher Education Research and Science Studies (DZHW) A standard data citation system is essential to make HEADS results widely usable and citable. The PID (Persistent Identifier) system is particularly well-suited for this purpose, as it assigns PIDs to (1) individual variables and (2) comprehensive information packages, which include a central reporting variable ("indicator") and related multivariate analyses conducted in HEADS. Both professionals and the interested public will benefit from using and citing data with PIDs. The dependent variable comprises several items from a larger theoretical construct. Each variable classified as "indicator" (see Figure 4) named ziwahr01 to ziwahr5 gets a PID.

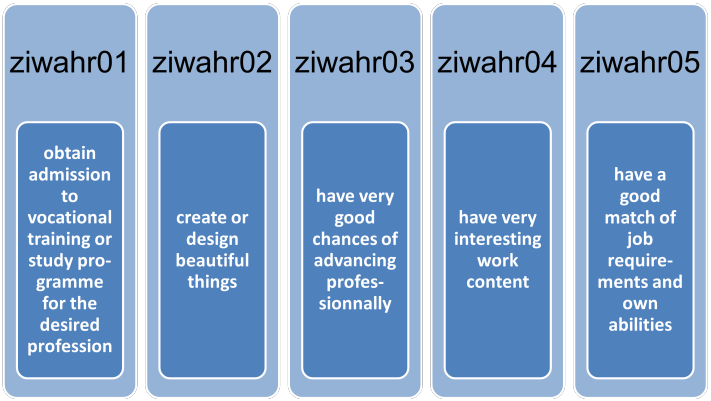

Figure 4: PIDs for each indicator variables.

In this use case, there are two ways to assign a PID: dependent variables can consist of multiple items, allowing for various intersections. Numerous independent variables enable differentiation of the dependent variable according to subgroups: gender, educational background, migration background, and school types. In this example, the dependent variable pertains to the target group's goals and achievements, operationalized as five "ziwahr" variables. These variables are analyzed concerning the independent variables of gender, educational background, migration background, and school types (see Figure 5).

7 GESIS – Leibniz Institute for the Social Sciences Homepage: `https://www.gesis.org/home`.

8 QuestionLink Homepage: `https://www.gesis.org/angebot/daten-aufbereiten-und-analysieren/question-link`.

9 Qualiservice - the data service center for qualitative social science research data is a data service center for archiving and sharing qualitative research data in the social sciences. Qualiservice Homepage `https://www.qualiservice.org`.

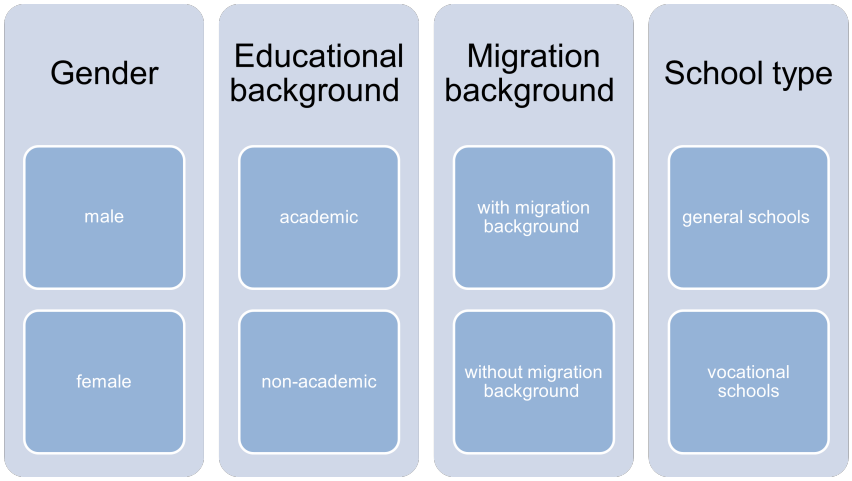

Figure 5: PIDs for each differentiation variables.

An information package is provided, displaying the values of the dependent variables (ziwahr01-05) according to the differentiation variables (gender, educational background, migration background, school types). For users, the critical aspect is the availability of a permanent location where the information, differentiated for subgroups or various differentiation variables, can be found and cited. It is immaterial whether users wish to use only the value of a subgroup, compare subgroups, or compare differences based on different differentiation variables. Consequently, a PID should be assigned to the information package to accommodate these various purposes (Figure 6).

As the PID is always connected to a "Landing Page", the data holder can define the level of data to be assigned a PID and describe it on the landing page.

4.2 GESIS Search

The GESIS Search[10] is a platform providing an integrated search across more than 6,500 national and international quantitative social science studies (mainly survey studies), more than 500,000 variables from those studies as well as instruments & tools and open access publications. The GESIS Search also provides links between diverse types of entities. However, PIDs are not assigned to variables so far. Applying PIDs to variables focuses on enhancing precise citation for secondary data analysis, as well as improving data discoverability and accessibility through automated access. With the accumulation of large datasets across studies and waves[11], dataset versions change over time. PIDs offer a more accurate way to distinguish repeated variables across the years, while ensuring direct access through automated features.

10 The GESIS Search is a Search Portal provided by GESIS to find information about social science research data and publications. GESIS Search Homepage: `https://search.gesis.org`.

11 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.

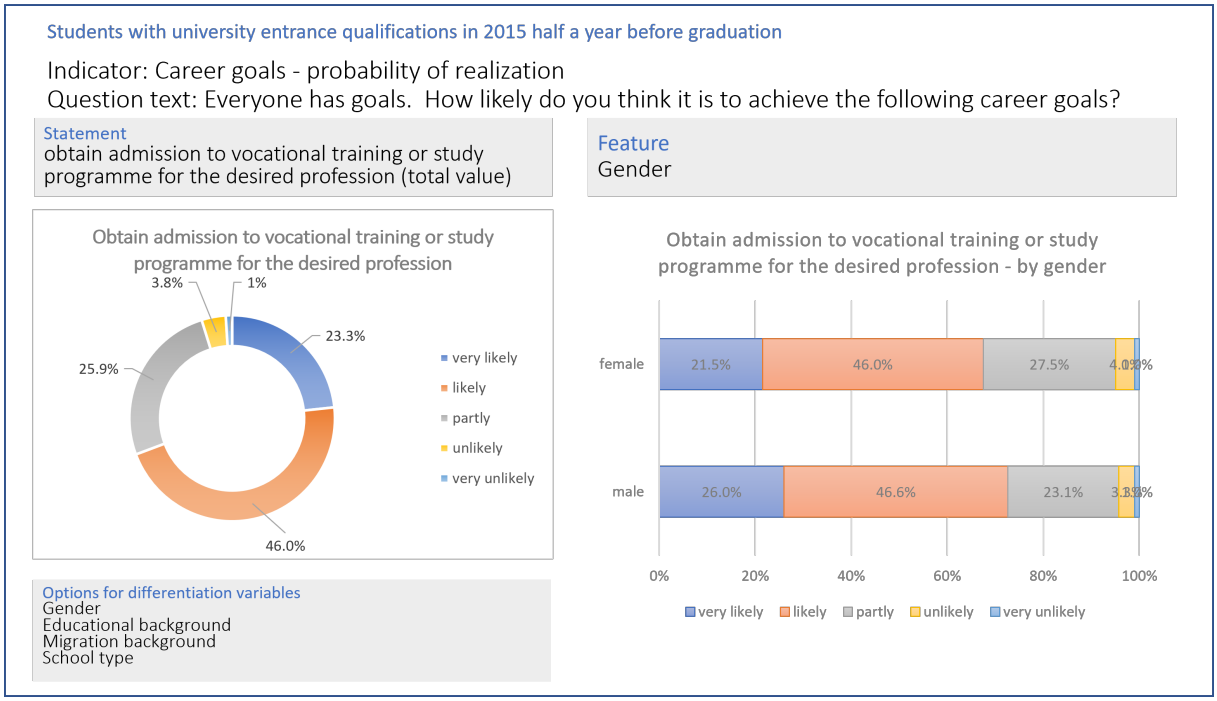

Figure 6: PIDs for information packages.

The direct access feature (automated access, i.e., by a computer program) is particularly relevant for GESIS for harmonisation tools within their code packages. These tools offer scripts and do-files that calculate response scales and provide harmonized measures and equivalent measures for similar variables across different studies, rather than providing data directly. Researchers are responsible for accessing datasets themselves from data providers. By assigning unique identifiers to each variable and embedding the variable's PIDs in these codes, it would simplify the use of numerous harmonized variables on the same topic from distinct sources. GESIS leads several projects that utilize such harmonisation tools.

4.2.1 GESIS harmonisation tool: QuestionLink:

QuestionLink is a tool that harmonizes sixty-eight political interest variables from seven measurement instruments: ALLBUScompact, GLES, GPANEL, ISSP (1990), ESS, NEPS, and SOEP. It helps researchers find and pool German data on political interest constructs over time and across large survey programs. However, accessing relevant data and applying the correct recording script require researchers to identify and retrieve the specific variables in source datasets. PIDs at the variable level would enhance QuestionLink and simplify its use.

For the selected instruments used sixty-eight political interest variables for harmonisation purposes, some surveys have the same variable name for multiple years, while others have different names per wave. Since question formulations and response labels may slightly differ across data collections while registering the same concepts, there is an elevated risk

Table 1: QuestionLink Example 1: Variable name similarity in the ALLBUS survey.

Survey	Instrument	Wave	Year range	Variable name	DOI
ALLBUS	ALLBUS B 10pt	cumulation	1982 – 1988	pa02	10.4232/1.13775
ALLBUS	ALLBUS A 5pt	cumulation	1980 – 2018	pa02a	10.4232/1.13775

of ambiguous citation when differentiating variables used in the same survey over the years.

Table 1 highlights the similarity in variable names within the ALLBUS instruments across different year ranges. The PID (DOI) identifies only the dataset, not the variable, and their names are similar.

Assigning PIDs for pre-harmonisation variables in the QuestionLink tool uniquely identifies individual variables per survey, wave, and year, preventing ambiguity due to repetitive variable names, which can be confusing and misleading.

4.3 SOEP-Core from DIW

The use case SOEP-Core[12] from the German Institute for Economic Research (DIW) encompasses various sub-samples and questionnaires related to households and individuals' members from Germans living in the former eastern and western German states, but also foreign citizens, and immigrants residing in Germany. Questions focus on finances, utilities, and general living conditions, while personal questionnaires explore work life, leisure activities, political interests, new family members, children's education, and youth-specific topics. The SOEP-Core consists of 101.574 variables, available from 560 data collections, distributed in 21.280 questions, and 309 instruments. The DIW office in Berlin documents the SOEP-Core complexity information extensively, covering topics, survey design, data editions, and distribution files. Panel data's interface[13] offers data and variable details, including variable landing pages. For example, the variable "*Interest in Politics*" (see Figure 7), with a landing page displaying variable values and timeline relations.

The SOEP-Core features a complex data structure with numerous datasets and variables across long-term investigations. Assigning a PID to identify these variables would lead the institute to utilize machine-actionable features to track and monitor the scientific output of specific variables. As the study covers various themes, PIDs enable tracking variable usage by subject and target types, such as household or individual-related information

12 German Socio-Economic Panel Study (SOEP-Core) Homepage: `https://paneldata.org/soep-core`.

13 Variable bip/bip_171: Interest in Politics from the Panel data: `https://paneldata.org/soep-core/datasets/bip/bip_171`.

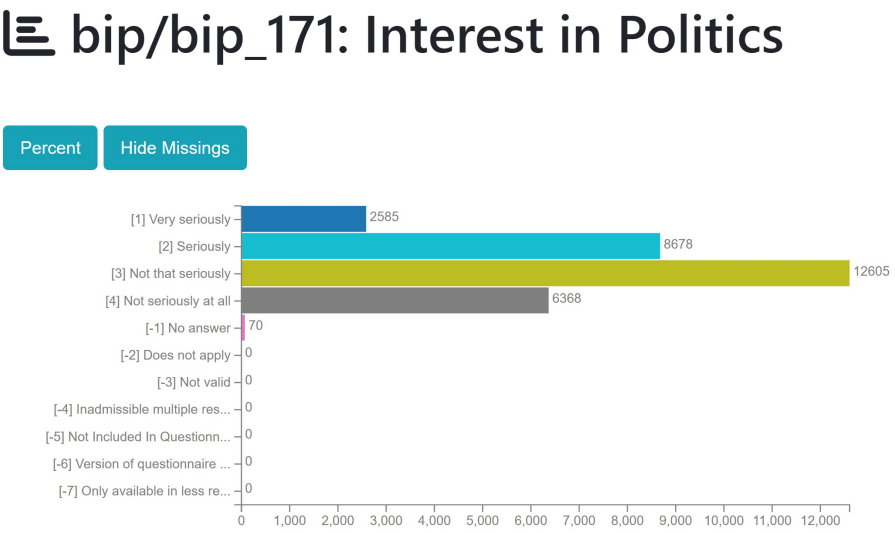

Figure 7: Variable graph: bip/bip_171: *Interest in Politics*.

in academic publications. This detailed tracking supports better evaluation of dataset usage, leading to improved decision-making regarding data services. Also, search and connect information from other datasets using the same variable under different labels. This approach is especially relevant for variables used in the harmonisation process within the same study or across different studies, such as the *Political Interest* variables used in the QuestionLink tool. Many variables are documented already and have a unique landing page. Registering this page as the variable's PID landing page would expedite the PID registration process, automatically linking to related datasets and enhancing findability.

4.4 Qualiservice

The use case Qualiservice is a data service centre from the University of Bremen. Its research data primarily includes qualitative interview transcripts and contextual data, which document the primary data collected during the research process. Qualiservice structures and standardizes data through metadata, registering elements according to the DDI 3.2 standard. PIDs (DOIs) are now assigned to identify dataset elements rather than study levels. A data collection, or dataset, can encompass various data types (interviews, observations), file versions (video, text, audio), or be organized by specific survey methods. PIDs for elements are assigned at the file level to distinguish between similar data types and file names, offering a direct method for citing, identifying, and accessing the target dataset element.

Considering the complex data structure encompassing a wide range of data files and formats, assigning a PID to each dataset element will simplify FAIR management of Qualiservice data. Relevant use cases include that Qualiservice's data structure varies from tabular data due to attributes like text, videos, and descriptive data. It demonstrates the PID service flexibility, including also qualitative data. PIDs can be assigned to identify

the dataset element that is considered most relevant to the data provider. Data files grouped in sets, such as datasets or collections, can maintain PIDs to identify related data from a specific study. Also, assigning a PID at the file level can be beneficial for disambiguating similar data files, given their data types and file naming similarities. This unique identifier gives end-users a straightforward method for citing, identifying, and accessing the target file. Since PIDs are machine-actionable elements, Qualiservice can leverage automatic features when assigning PIDs at the file level, such as linking related files within the same study or from different studies and directly accessing them.

Each use case has its unique aspects, as illustrated in previous examples. However, common advantages and benefits for RDCs and their users are consolidated in the conclusion section.

5 Conclusions

PIDs for lower granularity levels enhance FAIR data management, enabling Research Data Centers (RDCs) to benefit from the machine-actionable features of PIDs. By efficiently promoting data findability and accessibility at lower granularity levels, RDCs can make more informed decisions regarding services based on data utility, streamline data governance activities, and potentially reveal relationships between dataset elements across studies and datasets. This information lays the groundwork for knowledge graph visualization and fosters digital connections among researchers, organizations, and research outputs. Additionally, PIDs simplify harmonisation processes, which are often costly and time-consuming.

Regarding FAIRness, PIDs for lower granularity levels offer numerous benefits for researchers and data providers. They simplify FAIR data usage by providing unique identifiers for data elements below the study level, such as survey variables. PIDs enable referencing and retrieval of individual elements and metadata retrieval for data elements below the study level. They also help disambiguate data citations, promote safe and accurate data citations, and enhance recognition of produced data. Furthermore, PIDs foster credibility in research results and ensure the sustainable reusability of data while reducing documentation complexity. Additionally, PIDs offer feasible identification for various data types, including non-rectangular data attributes such as text, videos, and descriptive data.

Adopting PIDs to reference research data and their associated entities promotes FAIR data usage, as it significantly improves data findability, allows for more straightforward and, under certain conditions, automated access to data. Moreover, it enhances interoperability on a large scale by connecting dataset elements and other individual components, encourages data reuse, and simplifies the reproducibility of research. These benefits contribute to a more effective and efficient research ecosystem that fosters collaboration and knowledge sharing.

Future work. Dataset elements are interdependent and connected across studies and research outputs. In our approach, relationships between elements such as variables and

other attributes can be established, which we intend to incorporate in a knowledge graph representation of Social Science survey studies. PIDs, being machine-actionable, serve as the technical bridges that adhere to the FAIR principles, thus enhance the traceability of research results. By creating these connections and fostering a more comprehensive network, we can effectively improve the organization, accessibility, and overall understanding of research outcomes in these disciplines.

Acknowledgements

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

References

Bach, Janete Saldanha, Claus-Peter Klas, and Peter Mutschke. 2023. *KonsortSWD Measure 5.1: metadata schema extended report.* DOI: https://doi.org/10.5281/ZENODO.7588902.

Bensmann, Felix, Andrea Papenmeier, Dagmar Kern, Benjamin Zapilko, and Stefan Dietze. 2020. "Semantic Annotation, Representation and Linking of Survey Data". In *Semantic Systems. In the Era of Knowledge Graphs,* 53–69. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-59833-4_4.

Data Citation Synthesis Group. 2014. "Joint Declaration of Data Citation Principles". DOI: https://doi.org/10.25490/A97F-EGYK.

European Commission. Directorate General for Research and Innovation and EOSC Executive Board. 2020. *A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC).* Publications Office. DOI: https://doi.org/10.2777/926037.

Gregory, Kathleen, Anton Boudreau Ninkov, Chantal Ripp, Emma Roblin, Isabella Peters, and Stefanie Haustein. 2023. *Tracing data: A survey investigating disciplinary differences in data citation.* DOI: https://doi.org/10.5281/ZENODO.7555266.

Klas, Claus-Peter, and Oliver Hopt. 2022. "DDI Variable Documentation and data access using R". DOI: https://doi.org/10.5281/zenodo.7408629.

Klas, Claus-Peter, Janete Saldanha Bach, and Peter Mutschke. 2023. "GESIS Use case Variable publication and citation & Fine granular access to research data". DOI: https://doi.org/10.5281/ZENODO.7750031.

Klas, Claus-Peter, Matthäus Zloch, Janete Saldanha Bach, Erdal Baran, and Peter Mutschke. 2022. *KonsortSWD Measure 5.1: PID Service for variables report.* DOI: https://doi.org/10.5281/ZENODO.6397367.

Liebig, Stefan, Jan Goebel, Markus Grabka, Carsten Schröder, Sabine Zinn, Charlotte Bartels, Andreas Franken, et al. 2022. *Sozio-oekonomisches Panel, Daten der Jahre 1984-2020 (SOEP-Core, v37, Onsite Edition).* DOI: https://doi.org/10.5684/soep.core.v37o.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.