1308 2024

SOEP Survey Papers
Series G – General Issues and teaching Materials

Persistent Identifier (PIDs) für Variablen in SOEP-Core

Knut Wenzig und Dominique Hansen

DIW SOEP

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

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

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

The SOEP Survey Papers are available at http://www.diw.de/soepsurveypapers

Please cite this paper as follows:
Knut Wenzig und Dominique Hansen, 2024. Persistent Identifier (PIDs) für Variablen in SOEP-Core. SOEP Survey Papers 1308: Series G – General Issues and teaching Materials. Berlin: DIW Berlin/SOEP

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany

soeppapers@diw.de

The German Socio Economic Panel at DIW Berlin

Persistent Identifier (PIDs) für Variablen in SOEP-Core

Knut Wenzig und Dominique Hansen

2024

Wir möchten Denise Rolle, Jan Goebel und Jana Nebelin für ihre Anmerkungen und Kritik danken, die wesentlich zur Verbesserung dieses Manuskripts beigetragen haben.

Inhaltsverzeichnis

1 Motivation

Persistente Identifikatoren spielen im Forschungsdatenmanagement eine zunehmende Rolle, weil sie digitale Objekte referenzierbar und – einen Resolvingdienst vorausgesetzt – zugänglich machen.

Im Release v38.1 veröffentlicht das SOEP über 100.000 Variablen.[1] Für jede Variable soll ein persistenter Identifikator registriert werden.

Diese Arbeiten wurden im Rahmen eines vom KonsortSWD geförderten Kurzprojekts durchgeführt. Zum Projektende war allerdings eine PID-Registrierung in einem Produktivbetrieb noch nicht verfügbar. Deswegen wurden zwar Registrierungen mit einem Testsystem vorgenommen, aber noch keine PIDs registriert, die produktiv genutzt werden können.

2 Die zu registrierenden digitalen Objekte und die vorhandene Infrastruktur

Ein Datenrelease des SOEP-Core, dem seit 1984 laufenden Haushaltspanel mit jährlicher Befragung, umfasst mittlerweile über 500 Datensätze mit über 100.000 Variablen. In jedem Release, beispielsweise mit *v38.1* bezeichnet[2], werden alle seit 1984 erhobenen Informationen zusammengefasst veröffentlicht.

Jedes Release gibt es in 6 Editionen[3] mit unterschiedlichen Zugangsvoraussetzungen: *onsite*, *remote*, *area-types*, *planning-regions*, *eu*, *international* und *teaching*. Bestimmte Informationen sind also nur während eines Forschungsaufenthalts am DIW Berlin verfügbar, andere auch für den Einsatz in der Lehre.

Für jede Edition eines Releases werden 6 Datenpakete für unterschiedliche Kombinationen von Softwarepaketen und Sprachen erzeugt: Stata (zweisprachige Dateien mit deutsch als

1 `https://doi.org/10.5684/soep.core.v38.1o`

2 Wie das kleine *v* schon nahelegt, wird im allgemeinen Sprachgebrauch in diesem Zusammenhang auch von *Version* gesprochen. Innerhalb des SOEP-Metadatensystems gibt es tatsächlich auch ein Feld *version*. Allerdings hat dieser Begriff in einem technischen Kontext zu viele unterschiedliche Bedeutungen, weswegen zur Verhinderung von Missverständnissen in diesem Bericht von *Release* gesprochen wird.

3 `https://companion.soep.de/Data%20Structure%20of%20SOEPcore/Editions.html`

Defaultsprache), SPSS (deutsch und englisch), R (deutsch und englisch) sowie CSV (inklusive zweisprachige Variablen- und Werte-Labels in zusätzlichen CSV-Tabellen).

Jede PID für eine Variable muss auf eine Landing-Page verweisen können. Hier kommt das Portal `paneldata.org` ins Spiel, das derzeit weitgehend releaseagnostisch ist. Im Portal wird nur das jeweils aktuellste, in der Regel jährlich erscheinende, Release dokumentiert.

Für das SOEP wird derzeit für jede Edition eines jeden Releases eine DOI registriert.

In so fern stellt sich die Frage, welche Variablen in konkreten Datensätzen eines Releases registriert werden sollen:

- Jede Variable jedes Datensatzes jeder Edition?
- Jede Variable jedes Datensatzes für jedes Softwarepaket und jede Sprache?
- Jede Variable jedes Datensatzes jedes Releases?

Die Antworten sind nicht eindeutig, einzelne Entscheidungen haben ganz unterschiedliche Vor- und Nachteile:

- Soll beispielsweise mit Hilfe der PID auf die Werte einer Variable im Datensatz zugegriffen werden – wie es Saldanha Bach, Klas und Mutschke (2023a) beschreiben – ist es erforderlich zu wissen, wie eine Datei, die die Variable enthält, tatsächlich geöffnet werden kann. Es muss also bekannt sein, für welches Softwarepaket die Datendatei erstellt wurde, um die Datei mit der richtigen Importroutine öffnen zu können. Andererseits sollen sich die Inhalte der Spalten, also die eigentlichen Daten grundsätzlich nicht unterscheiden, wenn zwei Dateien betrachtet werden, die zwar für dasselbe Release und dieselbe Edition erzeugt wurden aber für unterschiedliche Softwarepakete.
- Ähnlich verhält es sich mit der Edition. Bestimmte Informationen sind nur in restriktiveren Umgebungen erhältlich, etwa in der EU und nicht in den USA oder nur vor Ort am DIW und nicht auf dem Rechner am Arbeitsplatz in einer Universität. Aber grundsätzlich sollen unabhängig vom Zugang die gleichen Ergebnisse berechnet werden können – soweit die Informationen eben verfügbar sind.
- Mit jeder Panelwelle wird ein Release im Vergleich zum Vorrelease zusätzliche Informationen enthalten: Entweder in Form von zusätzlichen Datensätzen (mit neuen Variablen) oder von zusätzlichen Zeilen in den sogenannten Long-Datensätzen. Der Großteil der Informationen verändert sich nicht. Sehr viele Ergebnisse, die mit der v37 ermittelt wurden, werden sich nicht ändern, wenn die v38 zur Datengrundlage wird.

Unzweifelhaft wäre es eine naheliegende Maximallösung, für jede Variable in einem Datensatz jeder Version, Edition und Softwareplattform und Sprache, eine PID zu registrieren. Statt etwa 100.000 PIDs müssten etwa 3,6 Millionen PIDs registriert werden. Der Aufwand, die Registrierungsinformationen zu warten, wäre immens – zumal es für jede PID eine Landing-Page geben muss, die aktuell nicht vorhanden sind. Nicht zuletzt muss es für jede Variable eine zugehörige DOI (vgl. Abschnitt 3) geben. Das wäre mit der aktuellen Registrierungspraxis des SOEP für DOIs nicht vereinbar.

Diese Maximallösung würde am Ende auch Unterschiede zwischen Variablen machen, wo keine Unterschiede angestrebt sind.

Die Frage der anzustrebenden Granularität hängt also immer auch vom konkreten Use Case ab, der realisiert werden soll. Allgemeingültige Antworten auf diese Fragen kann es nicht geben.

Am Ende wurden folgende Entscheidungen gefällt:

- Die PIDs für Variablen am SOEP sollen releasespezifisch sein. Dafür muss das Portal paneldata.org ertüchtigt werden, zwei gleichnamige Variablen, in einem gleichnamigen Datensatz aber unterschiedlicher Releases zu unterscheiden.
- Die PIDs für Variablen am SOEP sollen unabhängig von der Edition sein. Damit wird de facto nur für die Variablen der Edition *onsite* eine PID registriert. Im Portal paneldata.org ist auch nur eine Edition verfügbar.
- Es werden keine unterschiedlichen PIDs für Variablen in Datensätzen im Format einzelner Softwarepakete oder Sprachen registriert.

3 Die erforderlichen Metadaten

Saldanha Bach, Klas und Mutschke (2023b) haben im Projekt zur Registrierung von PIDs des KonsortSWD ein Metadatenschema entwickelt und veröffentlicht. Dort wird beschrieben, wie Metadaten im JSON-Format an eine REST-API gesendet werden, um sie zu validieren oder PIDs zu registrieren. Das Schema definiert folgende Felder: studyDOI, variableName, variableLabel, pidProposal, landingPage, resourceType, title, creators, publisher, publicationDate, availability, description.

Es fällt zunächst auf, dass das Schema keinen Datensatznamen enthält. Implizit wird an dieser Stelle wahrscheinlich davon ausgegangen, dass unter einer DOI jeweils nur ein Datensatz registriert wird.

Grundsätzlich sollen die Registrierungsinformationen aus dem Metadatensystem des SOEP kommen, das als eine sehr einfache Implementierung des Metadatenstandards DDI Lifecycle[4] verstanden werden kann. Zum Zwecke der Bearbeitung mit sozialwissenschaftlicher Standardsoftware liegen die SOEP-Metadaten als CSV-Tabellen vor. Diese werden zur Versionskontrolle in einem Git-Repository gehalten.[5] Hierbei gibt es u.a. eine Tabelle, die für jede Variable eine Zeile mit Informationen über diese Variable enthält.

Die Abbildung der Metadaten von SOEP-Core auf die für die PID-Registrierung erforderlichen Metadaten wird in Tabelle 1 dargestellt.

Zur Validierung und Registrierung dieser Daten wurde ein Python-Paket entwickelt[6]. Das Paket bietet, nach einer Installation, ein Kommandozeileninterface (CLI) an. Über dieses können Daten im Metadatenformat von SOEP-Core ausgewählt werden, die dann umgewandelt und konsekutiv an die Validierungs- und Regisrtrierungsschnittstelle gesendet werden.

Was das Metadatenprogramm des Registrierungsdienstes betrifft, erscheint es in Bezug auf den Namen des Datensatzes unvollständig. Solange unklar ist, wozu die Metadaten überhaupt genutzt werden (etwa in einem Suchportal) sollten die Metadaten allerdings auf das unmittelbar erforderliche Maß begrenzt werden, insbesondere wenn eine Aktualisierung der Informationen möglich ist.

4 `https://ddialliance.org/Specification/DDI-Lifecycle/3.3/`
5 Die Metadaten von SOEP-Core sind im Ordner *metadata* des Projekts `https://github.com/paneldata/soep-core` verfügbar.
6 `https://git.soep.de/dhansen/pid_4_soep`

Tabelle 1: Herkunft und Inhalt der für die PID einer Variable verwendeten Metadaten

Merkmal	Beschreibung
studyDOI	die DOI der Edition onsite. *Beispiel: 10.5684/soep.core.v38.1o*
variableName	Variablenname, Spalte variable in Tabelle variables.csv. *Beispiel: hid*
variableLabel	englisches Variablenlabel, Spalte label in Tabelle variables.csv, ggfs. Spalte label_de mit führendem [de]. *Beispiel: [de] Befragungsjahr*
pidProposal	{PID-Präfix}/{user}.{studyDOIsuffix}.{dataset}.{variable} Wenn der Punkt eindeutig Namensbestandteile voneinander abgrenzen soll, dürfte er nicht Bestandteil der einzelnen Partikel sein, wie das hier beim studyDOIsuffix der Fall ist. *Beispiel: 21.T11998/soep.soep.core.v38.1o.ah.hid*
landingPage	https://paneldata.org/api/pid/?study={study}&version={version} &dataset={dataset}&variable={variable} Die Adresse der Landing-Page soll also für das SOEP aus den bekannten Registrierungsinformationen gebildet werden. Damit muss diese URL nicht zusätzlich in den Metadaten abgelegt werden. *Beispiel: https://paneldata.org/api/pid/?study=soep-core&version=v38.1&dataset=ah&variable=hid*
ressourceType	Variable (fix)
title	{variable} - {variableLabel}. *Beispiel: welle - Survey Year*
creators	SOEP Group (fix). Nachdem die DOI-Registrierungsinformationen bereits Personen-Namen enthalten, wird hier nur eine institutionelle Autorin angegeben
publisher	SOEP Group (fix).
availability	Die Metadaten des SOEP enthalten bereits eine Information, die bei der Erzeugung der unterschiedlichen Editionen genutzt wird. Diese wird in das kontrollierte Vokabular rekodiert, die hier erwartet wird.
description	Die Metadaten des SOEP enthalten bereits eine solche Beschreibung. Sollte nur die deutsche Variante vorliegen, sollte wie beim Variablenlabel wieder ein entsprechendes Präfix angefügt werden. Ggfs. könnte hier künftig auch ein Text einer zugehörigen Frage eingefügt werden.

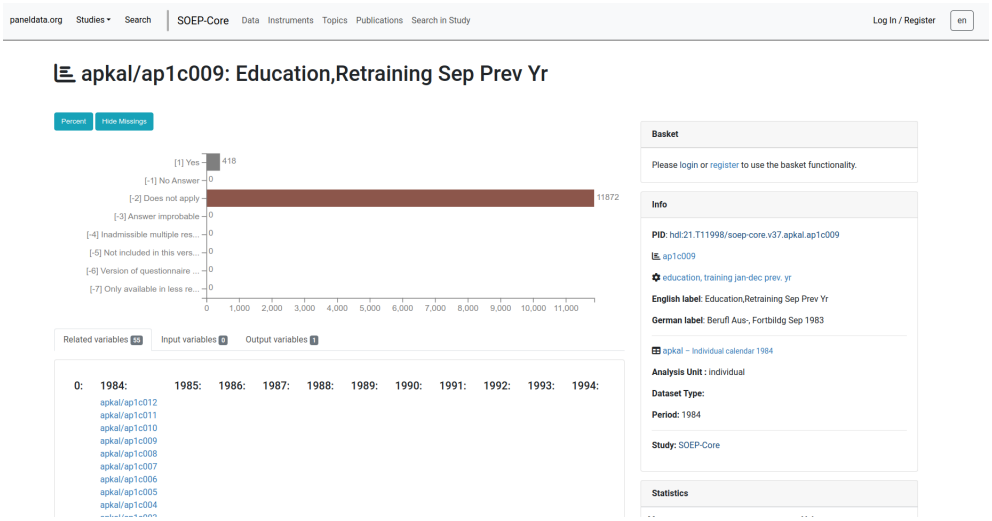

Abbildung 1: Die geplante Anzeige der PID auf einer Variablenseite in paneldata.org

Abbildung 2: Die geplante Anzeige, falls paneldata.org aktuell eine andere Version anzeigt, als die Version, für die die PID der Variablen registriert wurde.

4 Erforderliche Anpassungen am Metadatenportal paneldata.org

Die geplante Anzeige der PID für eine Variable soll auf der Variablenseite erfolgen, vgl. Abbildung 1. Hilfreich wäre ein Icon, wie es sich bei DOIs, ORCIDs oder RORs etabliert hat. Solange ein solches Icon nicht verfügbar ist, soll der Identifikator nach der Abkürzung PID in der Infobox ausgegeben werden.

Wie bereits erwähnt, ist paneldata.org releaseagnostisch und zeigt lediglich Informationen des aktuellen Releases an. Die PIDs sollen jedoch für jedes Datenrelease registriert werden. Für eine aufgerufene Landing-Page kann es daher zu drei Situationen kommen:

- PID des aktuellen Releases wird aufgerufen: Es erfolgt eine direkte Weiterleitung auf die Seite der Variablen, für die die PID registriert wurde.
- Die PID gehört zu einem alten Release und eine Variable mit gleichem Namen existiert noch in einem gleichnamigen Datensatz: Es werden minimale Information angezeigt und ein Link zur aktuellen Variable angeboten. (s. Abbildung 2)
- Die PID gehört zu einem alten Release und eine Variable mit gleichem Namen existiert nicht mehr in diesem Datensatz: Eine minimale Information wird angezeigt. Sollte es den Datensatz noch geben, wird dieser verlinkt.

5 Empfehlungen für die Implementierung des PID-Dienstes

- Auch wenn ein PID-Registrierungsdienst keine inhaltlichen Vorgaben zur Granularität der Inhalte machen sollte, wäre eine Handreichung zu diesem Thema sicherlich hilfreich.
- Die erforderlichen Metadaten sollten auf ein Minimum begrenzt werden, um die Nutzung möglichst zu erleichtern. Gleichzeitig sollte deutlich gemacht werden, wofür die angeforderten Metadaten tatsächlich genutzt werden. Das würde bei der Aufbereitung der Metadaten für die Registrierung sicherlich helfen.
- Im Schema wird einer Variable eine DOI (Feld studyDOI) zugeordnet. Damit wird eine weitgehende Vorgabe bezüglich der Verwendung von PIDs auf höherer Ebene (Studie, ggfs. Datensatz) gemacht.
- Zunächst ist eine Variable Teil einer Datentabelle, eines Datensatzes. Wenn nun die zugehörige studyDOI mehrere Datensätze bezeichnen kann, ist unklar, wo sich die Variable innerhalb des so bezeichneten digitalen Objekts befindet bzw. wie bei Namenskollisionen (zwei Variablen mit dem gleichen Namen in zwei Datensätzen, die unter einer DOI registriert werden) verfahren werden soll.
- Ggfs. wäre es sinnvoll, auch einzelne Datensätze registrieren zu können, soweit sie keine spezifische DOI haben.
- Für die Eigenschaft *availability* sollte ein kontrolliertes Vokabular zum Einsatz kommen.

6 Referenzen

- In der Zenodo-Community[7] des PID-Projekts sind Spezifikation und Software des Projekts verfügbar.
- Die API für die Registrierung ist unter `https://labs.da-ra.de/nfdi/` erreichbar.

Literatur

Saldanha Bach, Janete, Claus-Peter Klas und Peter Mutschke (2023a). *Breaking down hurdles of current data citation practices: use cases and benefits of persistent identifiers for dataset elements*. DOI: 10.5281/zenodo.8306007.

– (2023b). *KonsortSWD Measure 5.1: metadata schema extended report*. DOI: 10.5281/zenodo.7588902.

7 `https://zenodo.org/communities/konsortswd-ta5-m1`