1535 2025

SOEP Survey Papers
Series D - Variable Descriptions and Coding

SOEP-Core v40 – Activity Biography in the Datasets "pbiospe" and "artkalen"

Jascha Dräger and Tim Eder

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**

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

Please cite this paper as follows:

Jascha Dräger and Tim Eder. 2025. SOEP-Core v40 – Activity Biography in the Datasets "pbiospe" and "artkalen". SOEP Survey Papers 1535: Series D. Berlin: DIW Berlin/SOEP.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
© 2025 by SOEP

ISSN: 2193-5580 (online)

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

Contact: soeppapers@diw.de

Activity Biography in the Datasets "pbiospe" and "artkalen"

by Jascha Dräger and Tim Eder

(based on earlier work by Rainer Pischner, Henning Lohmann, Marco Giesselmann, Mila Staneva, Paul Schmelzer, and Maik Hamjediers)

Pbiospe and *artkalen* encompass activities over the life course and distinguish between spells in education, employment (full- and part-time employment as well as minor employment and registered unemployment), retirement, housekeeping, parental leave and others (see Table 3 for an overview). Please note that these spells do not capture transitions between educational institutions or jobs and employers; spells reflect a continuous status for instance in full-time employment regardless of potential job changes. In the yearly Individual Questionnaire, the respondents are also asked to report job changes between the previous and current years and this information can be added to *artkalen* using the *pgen* data files.[1]

The spell file *artkalen* is collected from Individual Questionnaires as a calendar-matrix of months of the previous year and respective statuses of 15 categories[2] (for example Question 141 in 2023.[3] The information from each annual Individual Questionnaire is then attached to the information of previous surveys in the *pkal* Data File. Then, Spells are generated for all consecutive months with the same category reported. The generated spell file starts with the year before the entrance into the sample and ends with a respondent's last observation.

The spell file *pbiospe* is based on the information on activity status over the life course, which is collected as a matrix from every respondent answering the Biography Questionnaire (for example Question 202 in 2023). The observations start at the age of 15 and end at the current age (up to age 65). Information on activity status covers only the period up to the time the biography is collected. To update the ongoing occupational career in *pbiospe*, information from the yearly Individual Questionnaire is also used by aggregating the recorded spells from *artkalen* into yearly values.

Content of the Datasets

Table 1 contains a list of all the variables in the datasets. The variables *begin* and *end* indicate the beginning and the end of a spell. These variables are age entries in *pbiospe* and month entries in *artkalen* starting with January 1983 as the first month. Pbiospe does also include reference to calendar years, e.g. *beginy* and *endy*. The *spellnr* is a serial identifier of spells of each activity status of a given person. The variable *spelltyp* contains information on the activity

[1] To add this information, it is necessary to split each spell at the time point of an interview. Afterwards, reports of job changes can be merged at each respective time point.

[2] For more detailed Information on the categories in the Biography and Calendar Questionnaire can be found in Table 3

[3] For persons who were temporarily unavailable for interviewing, it is sometimes possible to fill in the gaps in their occupational status. If these persons fill out the additional questionnaire for temporary dropouts later on, we can use the information collected there (see files $PLUECKE).

status during the spell, e.g., employed full-time or unemployed (For detailed information see Table 5)

Variables	Information in:	
	pbiospe	***artkalen***
cid	Original Household Number	Original Household Number
pid	Never Changing Person ID	Never Changing Person ID
spellnr	Serial Number of the Spell per Person	Serial Number of the Spell per Person
spelltyp	Type of Spell	Type of Spell
begin	Age Spell Begins	Month Spell Begins
end	Age Spell Ends	Month Spell End
beginy	Year Spell Ends	
endy	Year Spell Ends	
zensor	Censor Variable	Censor Variable
source		Originating Dataset
other_flag		Flags spells generated from open-ended responses for "other" spells
duplicate_flag		Flags copied spells
spellinf	Spell Construction Information	
erhebj	Survey Year Biography Data	
kalyear	First Observation in Calendar Data	
beginbX*	Age Spell Begins, Xth Initial Biography Spell	
endbX*	Age Spell Ends, Xth Initial Biography Spell	
beginybX*	Year Spell Begins, Xth Initial Biography Spell	
endybX*	Year Spell Ends, Xth Initial Biography Spell	
begink X*	Age Spell Begins, Xth Initial Calendar Spell	
endkX*	Age Spell Ends, Xth Initial Calendar Spell	
beginykX*	Year Spell Begins, Xth Initial Calendar Spell	
endykX*	Year Spell Ends, Xth Initial Calendar Spell	

*Table 1: All Variables in the artkalen and pbiospe Datasets; *These variables list all the Spells that were combined into this Spell. As they are generated dynamically the number of variables varies.*

Both the Biography and the Individual Questionnaire allow for multiple activity statuses for a given year or month. No concept of main activity is used. A common combination is, for instance, “housewife/-husband” and “working part-time”. There are several other plausible combinations, but also combinations that are less plausible. However, a list of valid combinations of activity statuses defined according to legal or similar constructs would need to be based on very strong assumptions. In addition—in case of the yearly matrix in the Biography Questionnaire—activities are reported that took place in a calendar year in consecutive months, which makes it impossible to exclude combinations of activities.

Therefore, no data cleaning is performed at this stage. Consequently, the data may contain information on more than one activity for a given point in time.

Differentiation between Apprenticeship/Retraining and Minijob/Part-time spells in *artkalen*

For some activities, more detailed information was collected in later survey years.

Until 1998, participants where asked whether they did any kind of vocational training (*spelltyp* == 4), which includes both First Job Training, Apprenticeship and Continuing Education, Retraining. Since 1999, there are separate questions on whether participants did a First Job Training or Apprenticeship (*spelltyp* == 41) or whether they did Continuing Education or Retraining (*spelltyp* == 42). Thus, information on the more detailed activity is only available since January 1999 (month 193 since January 1983). To enable continuity, Spells with type 41 and 42 January 1999 are copied as type 4 Spells and expand the existing Spells if applicable. The spells that contain copies of more detailed spells are marked by the variable *duplicate_flag*.

Likewise, until 2003, respondents were asked whether they are part-time employed or whether they did a 'Mini job', but these two activities were not differentiated (*spelltyp* == 3). Since 2004, there are separate questions on whether participants were part-time employed (*spelltyp* == 31) or whether they did a 'Mini job' (*spelltyp* == 32). Thus, information on the more detailed activity is only available since January 2004 (month 252 since January 1983). Again, to enable continuity, spells of type 31 or 32 are coped as type 3 spells and expand the existing spells if applicable.

Table 2 gives and example of how this was done.

Before Duplication									
cid	pid	spelltyp	spellnr	begin	end	zensor	source	other_flag	duplicate_flag
868	8605	4	3	165	193	1	1	0	0
868	8605	41	4	193	206	1	1	0	0

After Duplication									
cid	pid	spelltyp	spellnr	begin	end	zensor	source	other_flag	duplicate_flag
868	8605	4	3	165	206	1	1	0	1
868	8605	41	4	193	206	1	1	0	0

Table 2: Process of Spell duplication for continuity

If you want to revert this change in *artkalen* you simply need to set the end variable for all Spells with *spelltyp* == 4 and *duplicate_flag* == 1 to month 192 if they exceed this month. For Spells with *spelltyp* == 3 and *duplicate_flag* == 1 you need to set the end variable to month 252 if they exceed it.

Aggregation of 'Other' Spells into usable spell types

Open Field Response	**Spelltyp in *artkalen***
"Short Work Hours"	2: Short Work Hours

“Maternity Leave”	7: Maternity Leave
“Not Working”	5: Unemployed
“Pre-Retirement”	6: Retirement
“1-Euro Job”	15: Minijob
“Social Aid”	5: Unemployed

Table 3: Aggregation of Spelltypes from Open-Field Responses

The calendar Questionnaire also includes an open field response option for the years from 1986 to 2020. Some responses from this field are used to generate useful Spell-Information. Table 3 depicts which information is used for which *spelltyp*es. Spells that are created in this way are flagged using the variable *other_flag*. There are 2923 Spells with aggregated ‘other’-information currently in *artkalen*, this adds up to 0.6 percent of Spells.

Censoring; Coding of the Variable *zensor*

Right: Left:	Not censored	Censored missing	Censored before gap
Not censored	1	2	3
Censored missing	4	5	6
Censored after gap	7	8	9

Table 4: Overview of zensor-Variable; Note: (99) Gap Spells are marked with -2

Missing information on the beginning or end of a spell causes what is known as censoring problems. There are two types of missing data. First, data can be missing on periods outside the observation window (before the age of 15 and after the age of 65 in the case of *pbiospe* or before and after the panel participation in *artkalen*). Second, data can be missing on years within the observation window due to item non-response in particular years or due to temporary drop-outs (the latter applies to calendar information only). In this case, we speak of “gaps.” There are nine different patterns (see Table 4).

Aggregation of *artkalen* information into *pbiospe*

To enhance the data of the Biography Information *artkalen* data is aggregated into yearly values. *Spelltyp* in *pbiospe* is less granulated than the types in *artkalen*, so *spelltyp* is aggregated as depicted in Table 5. The category 3 “military / civilian service” denotes several activities, such as being imprisoned, serving in the military, doing former community service or any voluntary service (FSJ or Bundesfreiwilligendienst). If you are more interested in distinguishing these, it is necessary to trace down the information in the person data $P (e.g. variable pkal09a in bhp) and to merge them onto *artkalen* or *pbiospe*.

The *artkalen* Spells are first converted into the new spelltypes (Gaps are dropped) and transformed to yearly begin and end variables[4], before extending Spells if necessary to achieve

[4] Note that this means that also when only 1 month of activity is reported in artkalen the whole year will be recorded with the activity in pbiospe. Also, if this happens in consecutive years those Spells are expanded, so if a person has 1 month Retraining (14) Spells in March 2000, 2001, and 2002 respectively, this person is recorded with a Training (2) Spell in pbiospe that begins in 2000 and ends in 2002.

a congruent set of Spells. A question that arises when merging the data is how to handle overlapping pieces of information. The basic principle is to assign a value of a given status in a given year if the status is recorded in the calendar or in the biography information or both. An example might help to illustrate this: the calendar records full-time employment for the years 2005 and 2007 while the biography records full-time employment for the period from 2000 up to 2006. The merged data from *pbiospe* contains a spell that begins in 2000 and ends in 2007.

	pbiospe	***artkalen***
1	School/University	School, University (8)
2	Apprenticeship/Training	Vocational Training (4), First Job Training, Apprenticeship (41), Continuing Education, Retraining (42)
3	Military/Civilian service	Military, Community Service (FSJ, BufDi, Zivildienst) (9)
4	Full-time employed	Full-Time Employment (1)
5	Part-time employed	Part-Time Employment/marginal employment (3), Second Job (11), Part-Time Employment (31), Mini-Job (up to 556€ (in 2025)) (32)
6	Unemployed	Unemployed (5)
7	House-Husband/Wife	House-Husband/Wife (10)
8	Retired	Retired (6)
9	Other	Other (12) Maternity Leave (7)
10	Short Work Hours	Short Work Hours (2)
99	Gap	Gap (99)

Table 5: Aggregation of Spelltypes from artkalen to pbiospe

However, the initial information is restored by including additional variables, which allows for alternative ways of merging the data (see below). The variables *spellinf*, *erhebj*, and *kalyear* contain general information on the sources of the information captured in each spell. Table 6 shows that the majority of spells are based on biography information only (51.2 percent). One-third of all spells (33.03 percent) are not observed in the Biography Questionnaire but only in the calendar data. The remainder of spells contain information from biography as well as calendar data. Usually, these spells combine one period observed in the Biography Questionnaire with a period observed in the calendar. Only 0.5 percent of the spells combine more than one period in any of the two sources (*spellinf*=4, 5 or 6).

spellinf	N	%	% cum.
Biography only	303486	51.20	51.20
Calendar only	195790	33.03	84.24
1 biography, 1 calendar spell	90478	15.27	99.50
2+ biography, 1 calendar spell(s)	734	0.12	99.62
1 biography, 2+ calendar spell(s)	2143	0.36	99.99
2+ biography, 2+ calendar spell(s)	84	0.01	100

Table 6: Construction of Spells in pbiospe

The variables *beginbx-endykx* document the initial information from the two different sources and are probably not of interest to most users. However, based on these variables, users can fully separate the Biography data from the aggregated *artkalen* data. This is advisable if you want to use the more detailed *artkalen* information and combine it with the yearly information from *pbiospe* for earlier years only. The variable names indicate the "source" of the original information utilized (B: Biography -Questionnaire or K: calendar information from the yearly survey). As an example, we discuss one of the spells that combines information on more than one period from any of the two sources. The spell number 3 of person 9205 starts in 1983 and ends in 1994 (*spelltyp*=4: full-time employment). As the variable *spellinf* (=5) shows, this a spell that combines one period from the biography data with two periods from the calendar data. According to the biography data, the person worked full-time from 1983 (*beginyb1*) until 1992 (*endyb1*). There is overlapping information from the calendar data available from 1986 onwards (*kalyear*). According to these data, the person worked full-time from 1986 (*beginyk1*) to 1990 (*endyk1*) and from 1993 (*beginyk2*) to 1994 (*endyk2*). During the years 1991 and 1992, no full-time employment is recorded in the calendar data, which contradicts the information from the biography data.

pid	spellnr	spelltyp	beginy	endy	spellinf	erhebj	kalyear	beginyb1	endyb1	beginyk1	endyk1	beginyk2	endyk2
9205	3	4	1983	1994	5	1998	1986	1983	1992	1986	1990	1993	1994

Table 7: Example Observation in pbiospe

In *pbiospe*, no attempt is made to "resolve" such contradictions, as this would require rather strong assumptions. More important, such assumptions would differ according to the research question, which makes it even more difficult to provide a standard solution. Therefore, in such cases, we generate spells in the same manner as in less difficult cases, namely by combining the information from the calendar and the biography data. In the given example, this results in a full-time employment spell that starts in 1983 and ends in 1994. As mentioned above, there are very few spells that combine information on two or more periods (*spellinf*=4, 5, 6, less than 0.5 percent of all spells). There are even fewer such spells where the period of overlap is as long as in this example, where the biography data was collected many years after the persons joined the survey (*erhebj*=1998, *kalyear*=1986). However, users who are interested in combining biography and calendar data in a different manner can use the variables *beginbx-endykx* to fully separate the two types of data and to recombine the data on the basis of different rules of aggregation.

Overlap between pbiospe and artkalen

As stated above, the calendar information is used to update the biography information. However, there is also a certain overlap of the periods covered by the two types of data. This is shown in Table 8 It indicates, for persons included in *pbiospe*, the year in which the biography information was collected (variable *erhebj*). This year is usually also the last year for which biography information is available. The table also shows the first year recorded in the calendar data (variable *kalyear*). In most cases (53.8 percent), the earliest calendar information is available for the year before the biography interview. This is the case for

persons who answered the Biography Questionnaire in their first year as survey respondents. The calendar in the Individual Questionnaire refers to the year before the survey. There are, however, changes over time. In 1998, it was decided that first-time respondents from new samples would not be given the Biography Questionnaire in the first wave but in the second to reduce the entry threshold for these new respondents. Consequently, for the majority of persons in years after new samples were integrated (1999, 2001, 2003, 2007, 2010/11 – Samples E to I), the earliest calendar information is available two years before the biography information was collected. This was once again changed in 2011 and in 2013: respondents from samples J and K were given both questionnaires in their first year in SOEP and 2013 the sample M was introduced via a special Biography Questionnaire about individual migration histories without surveying the calendar data. The same applies to first-time respondents who are members of an old sample (e.g., persons who moved into a panel household) - they answer the Biography Questionnaire at the time of their first interview. The pattern is quite stable for most years before 1999. Notable exceptions is the year 1992. This is explained by the integration of East Germany into the SOEP in 1990 (Sample C). Most of the respondents in these samples answered just the Biography Questionnaire at the entrance into the SOEP. Another exception is the year 1987. In the years 1985 to 1987, the life course matrix was not part of any of the questionnaires. Therefore, the respective biography information was only available for persons who were interviewed in 1984. In 1988, biographic information was also collected for persons who became respondents in 1985, 1986, and 1987 (for all years *erhebj*=1987).

Year of Biography data collection	Difference to Kalyear*					
	same year or later	1 year ealier	2 years ealier	3 years ealier	4+ years ealier	Total
1984	12.80	0.00	0.00	0.00	0.00	12.61
1987	0.68	0.00	0.00	0.00	0.00	0.67
1988	0.23	0.00	0.00	0.00	0.00	0.22
1989	0.28	0.00	0.00	0.00	0.00	0.28
1990	0.27	0.00	0.00	0.00	0.00	0.27
1991	0.22	0.00	0.00	0.00	0.00	0.21
1992	5.88	0.00	0.00	0.00	0.00	5.79
1993	0.44	0.00	0.00	0.00	0.00	0.43
1994	1.25	0.27	0.00	0.00	0.00	1.23
1995	1.37	0.02	0.00	0.00	0.00	1.35
1996	0.59	0.00	0.00	0.00	0.00	0.59
1997	0.57	0.00	0.00	0.00	0.00	0.56
1998	0.52	0.11	0.00	0.00	7.59	0.52
1999	2.19	0.00	0.22	0.00	0.00	2.16
2000	0.31	0.00	0.00	0.00	0.00	0.31
2001	9.61	0.00	0.00	0.00	0.00	9.47
2002	0.68	0.00	0.00	0.00	0.00	0.67
2003	2.63	0.00	0.00	0.00	0.00	2.60
2004	0.57	0.00	0.00	0.00	0.00	0.56
2005	0.39	0.00	0.00	0.00	0.00	0.38
2006	0.29	0.00	0.00	0.00	0.00	0.28
2007	2.33	0.00	0.00	0.00	0.00	2.29
2008	0.38	0.00	0.00	0.00	0.00	0.37
2009	0.22	0.00	0.00	0.00	0.00	0.22
2010	10.67	0.00	0.00	0.00	0.00	10.52
2011	7.34	0.00	0.00	0.00	0.00	7.23
2012	2.69	0.00	0.00	0.00	0.00	2.65
2013	4.43	3.46	5.07	2.78	35.44	4.43
2014	0.48	0.27	0.00	0.00	5.70	0.47
2015	1.56	0.28	0.00	0.00	16.46	1.55
2016	3.39	47.43	41.04	39.35	23.42	4.00
2017	5.94	31.29	40.75	45.83	11.39	6.33
2018	1.70	5.00	5.07	9.72	0.00	1.75
2019	3.09	2.36	2.86	2.31	0.00	3.08
2020	1.84	9.02	4.99	0.00	0.00	1.93
2021	1.91	0.50	0.00	0.00	0.00	1.88
2022	8.69	0.00	0.00	0.00	0.00	8.57
2023	1.58	0.00	0.00	0.00	0.00	1.56
Total	100.00	100.00	100.00	100.00	100.00	100.00

*Table 8: Overlap between the data from Biography vs Calendar Questionnaire; *Year of Calendar data Collection*

Changes since last documentation (v35)

artkalen

- Generation modalities
    - We have switched from an incremental generation process that generates biographies for the new wave and appends them to the previously generated Spelldata from the previous versions, to generating the whole dataset with each version.
- Name changes
    - For consistency reasons with other datasets of SOEP following name updates were conducted:
        - *persnr -> pid*
        - *hhnr -> cid*
- New Variables
    - New Variables were introduced for the first time in v40
        - *duplicate_flag*: Flaggs duplicated Training Spells that were created for consistency reasons (see above).
        - *other_flag*: Flaggs Spells that were created from an open textfield response.
        - *source*: Names the originating dataset from which the information is sourced.
- Spells are now also created from answers to an open text field, if they fit certain categories (for more information see chapter “Aggregation of ‘Other’ Spells into usable spell types”).
- To increase granularity and continuity, spells of *spelltyp* 3 and 4 are now used for continuity along the whole timespan. *Spelltyp*es 31,32 and 41,42 respectively are used to distinguish between the granular differences between part-time employment(31)/Minijob(32) and apprenticeship(41)/retraining(42) (For more information see chapter “Differentiation between Apprenticeship/Retraining and Minijob/Part-time spells in *artkalen*”).

pbiospe

- Due to internal data restructuring, the generation process for the biography data in *pbiospe* now also aggregates from Spelldata of previous *pbiospe* versions.
- New Variables
    - The new Version of *pbiospe* includes more separate calendar and biography Spell variables. Since those are generated dynamically this can be traced back to the new generation method.
- New Spelltyp
    - Short Work Hours is now included as a separate spelltyp (10) in *pbiospe.*