PanelWhiz: Efficient Data Extraction using the German SOEP

John P. Haisken-DeNew (University of Bochum, Germany) and
Markus Hahn (Melbourne Institute, Australia)

Abstract: This paper outlines a panel data retrieval program written for Stata/SE or better, which allows easier accessing of the German SOEP household panel data set. Using a drop-down menu and mouse click system, the researcher selects variables from any and all available years of the panel. The data is automatically retrieved and merged to form a "long file", which can be directly used by the Stata panel estimators. The system implements modular data cleaning programs called "plugins". Yearly updates to the data retrievals can be made automatically. Projects can be stored in libraries allowing modular administration and appending.

Keywords: Panel data storage and retrieval
JEL: C81, C87, C23

1. Introduction

Applied social scientists have forever been faced with different data interfaces for different data sets. In most cases, an interface is not even available, forcing the researcher to address data files by name, and extract the information required by hand. However, the specific structure of a panel data can vary dramatically as described in Haisken-DeNew (2001), such that some data sets provide many files per year, differing by their population, or level of aggregation etc., creating many obstacles for researchers.

PanelWhiz is a collection of subroutines that allows researchers to use an intuitive "common" graphical interface for accessing many datasets directly within the statistical package Stata/SE, (http://www.stata.com) whereby the researcher does not select individual variables, but rather vectors of variables (items) with one mouse click. This allows for an efficient method of selecting information for a data set retrieval, especially if the data set contains many waves (years) of information. With one mouse-click, data can be automatically retrieved, with merging and matching done automatically.

With the PanelWhiz system, the user can open data files by clicking on a browse page. The idea behind the tool is that because of the intrinsically longitudinal nature of the data, one is typically not interested in retrieving a variable in a particular wave, but rather in retrieving the variable for several waves, i.e. an item-correspondence. For all data sets, a variable renaming algorithm (where necessary) is used to ensure time consistent variable names (See Haisken-DeNew, 2001 for more information on this). Thus, if one opens a data file and one finds a variable of interest, one clicks on the variable and information for the entire item correspondence is also collected and added to a "PanelWhiz project". Straightforwardly, the object is to collect items and save them into the data project, allowing an automatic data retrieval. The data are extracted in "long" format allowing easy further data cleaning or direct estimation using Stata's panel "xt" commands.

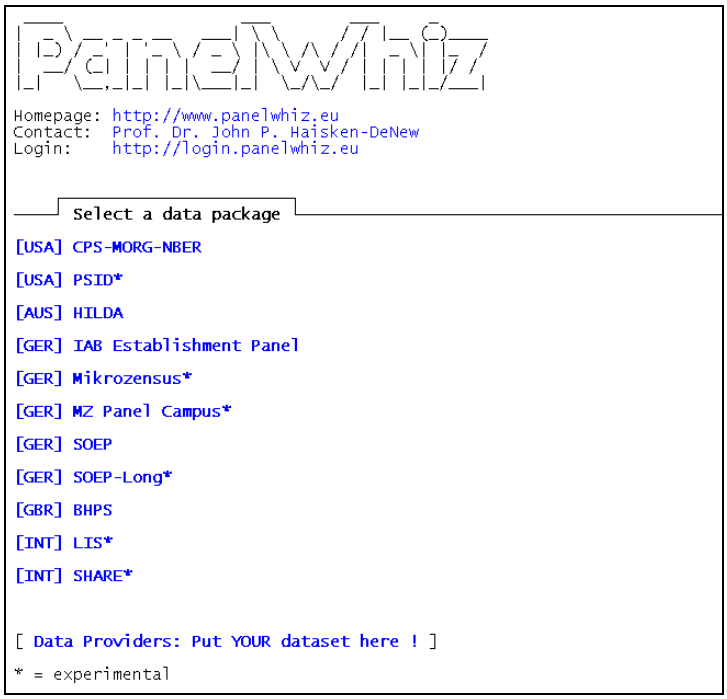
PanelWhiz, since appearing in 2005, now has several hundred registered users, using the common interface to access many different datasets other than the SOEP, such as the British BHPS, the German IAB Establishment Panel, the Australian HILDA, the American CPS, etc. Recently, support has been extended to the American PSID. This paper describes use of PanelWhiz for the German SOEP. However due to the generalized nature of PanelWhiz, the interface is almost completely identical for all other supported datasets.

2. Overview of PanelWhiz

2.1 Starting Up Stata with PanelWhiz

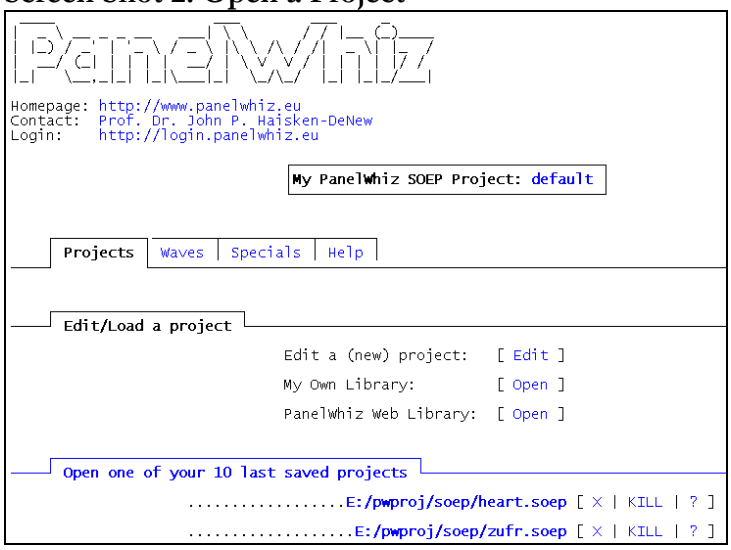
PanelWhiz is installed as a Stata Add-On and loads every time Stata is started. For example in following Screen Shot, one can select by mouse click the desired data set to be supported.

Screen Shot 1: Select Data Set



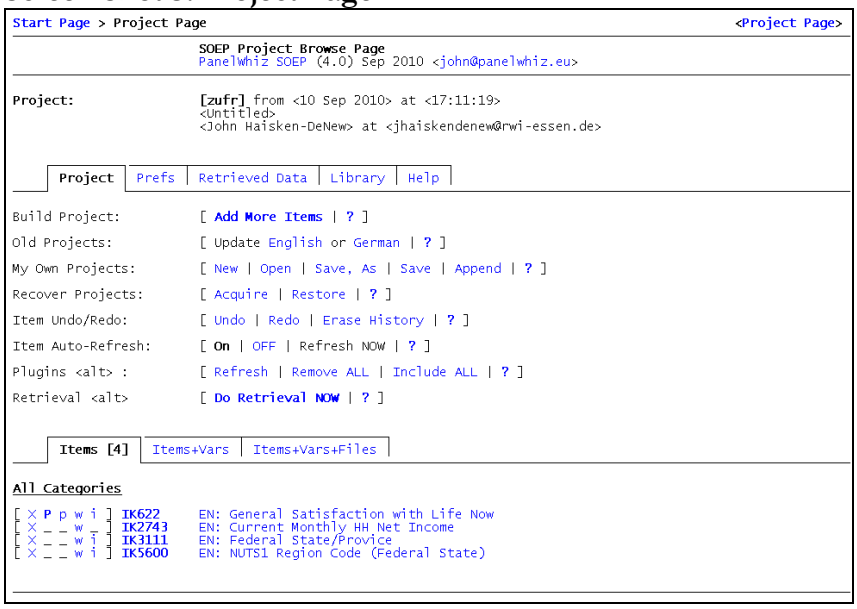
In this example, the SOEP has been selected. One can select an already existing project, or create a new one from scratch. In this example, we will examine an existing project **`zufr.soep`**. Because it has been already saved, PanelWhiz keep a note of the last 10 saved projects and allows easily loading by simply clicking on the blue link indicating the project name.

Screen Shot 2: Open a Project



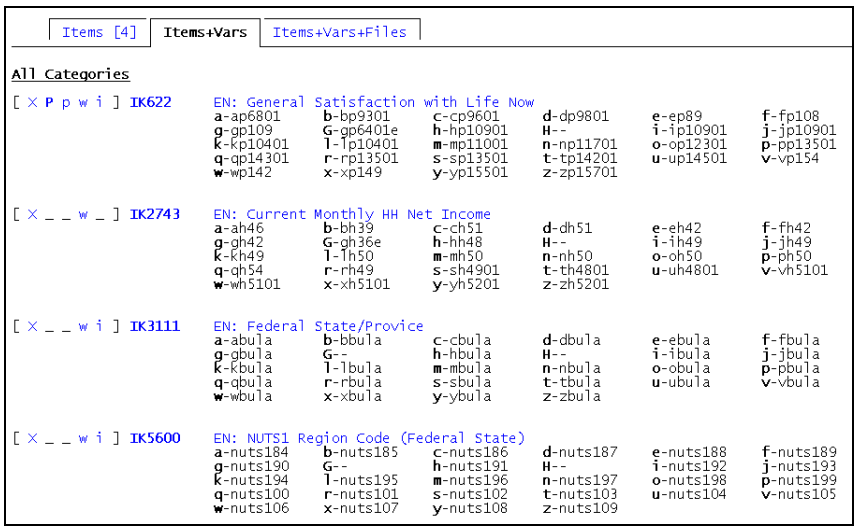
Here we have indeed opened the PanelWhiz project **`zufr.soep`**, and have a heads-up display indicating the contents of the project and the possible project commands. The top area displays the possible project commands and the bottom area the contents of the project. Here the project already contains 4 items (vectors of variables). The item labels are blue (and therefore clickable, linked to a keyword thesaurus).

Screen Shot 3: Project Page



The variables underlying the 4 items in the project are listed below. Each variable listed under the item displays the associated wave/year. In the SOEP, the "a" wave is 1984, the "b" wave is 1985 and so on. The "z" wave is 2009. Thus for the item ***IK622***, in 1984 the underlying variable is **`AP6801`** and in 2009 it is **`ZP15701`**.

Screen Shot 4: Start Page



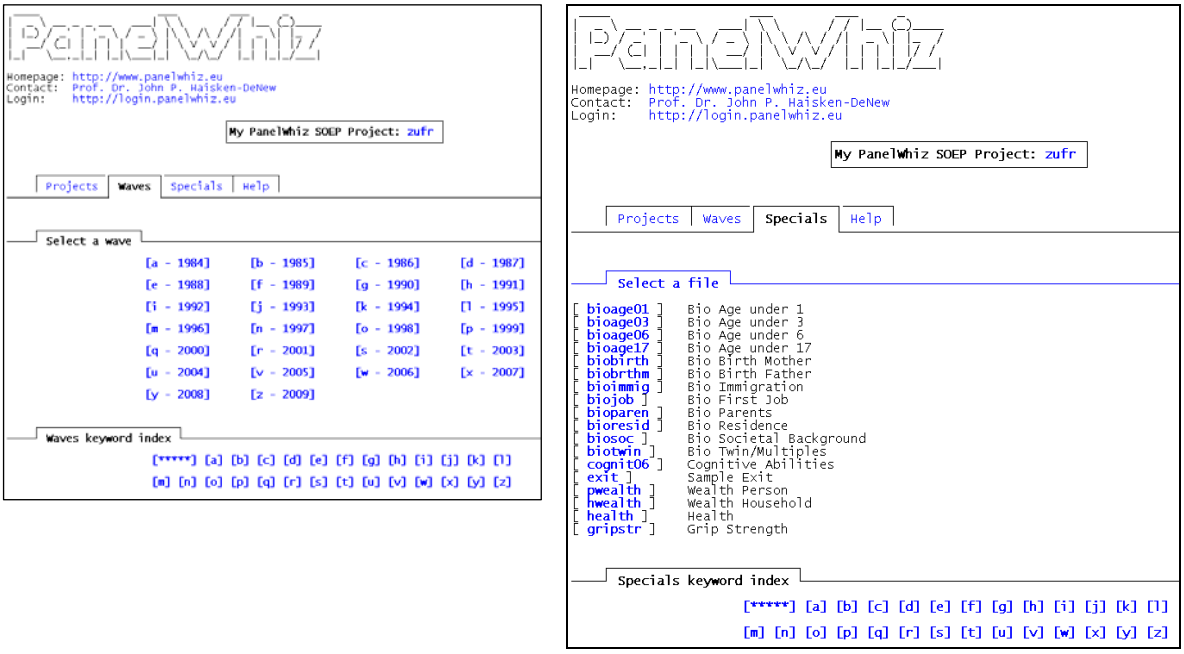
One can update the project using the automatic "Update" function on the project page. When a new data distribution becomes available, for each item, the newest variable is added automatically to the relevant item. The retrieval can be run again, having now the most recent information.

2.2 Items and Specials

Assuming that one would like to add any items to the project, one can choose between two types of concepts: "items" or "specials". Items are vectors of variables that have a standard time dimension associated with them, i.e. one variable for each year. SOEP examples of these files would be **AP**, **APGEN**, **AH**, **AHGEN** etc. Specials have a non-standard time dimension, i.e. they may have one observation per person and be time invariant, or may already be in long format, with person-year observations as the unit of analysis. Here we will first examine the page associated with items.

Clicking on a year like **[a - 1984]**, will load a browse page allowing one to click on all variables/items associated with the year 1984. Alternatively, one can click on a special file, such as **[bioimmig]** where the data is already in long format. In contrast, the information from the special file **[bioparen]** is time invariant and contains only one entry per person. PanelWhiz knows how to extract and merge the information from all of these kinds of files.

Screen Shot 5: Items and Specials



For both items and specials, all associated items have been scanned and the contents of the item labels have been catalogued into a thesaurus of keywords. Thus, if one were interested in all items or specials regarding the topic of "occupation", one would click on the **[o]** of the keyword index, to examine all keywords starting with the letter "o".

Technically speaking, PanelWhiz works because the item correspondence information (for each item, that vector of variables over all 26 years) is injected into each relevant variables as a Stata variable characteristics. PanelWhiz reads this information from a variable in one particular file/wave and automatically knows where to find corresponding information in all other files/waves.

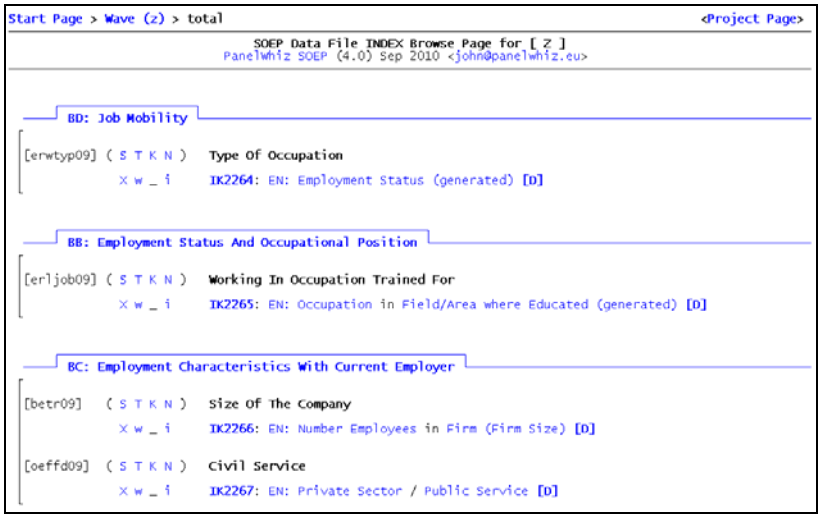
2.3 Item Browse Page

To find an item found say in the year 2009, we click on the year **[z – 2009]**, and receive the following browse page. It contains a browse page containing all variables/items from all files in the year 2009. Alternatively, one can jump to only specific variables/items in a particular file. Further, the variables are sorted using the same hierarchical scheme as in SOEPinfo. See Haisken-DeNew and Frick (2005) for more information on SOEPinfo.

Screen Shot 6: Top Level Item Browse Page

In this example, we select the blue clickable button **[ALL FILES]** from wave Z and get the following browse page. All variables are listed in the order they naturally exist in the respective physical data files. The example shows that for the SOEP variable **erwtyp09**, there is a PanelWhiz item **IK2264** associated with it. By actually clicking on **IK2264**, one would select the entire item (potentially all underlying variables from wave "a" through "z").

Screen Shot 7: Look at Items



One can also examine the changing nature of the item over time. The variable **erwtyp09** contains value labels. Thus we can click on blue "**i**" to the left of the item name and label. There is a ready-made HTML page showing all labelled values for all variables of the entire item. This will be especially useful information for data cleaning requirements. Just because a variable has been coded one way in one year/wave, it does not mean it will remain so over all time. The

following Screen Shot illustrates this example. Jumping from wave 1984 to 1985, there have been some additional outcome values added. These changes are colour coded in grey.

Screen Shot 8: Start Page

(2264) EN: Employment Status (generated)

[~ := no data available] [x := no label/value available]

VALUE	1984 (a)	1985 (b)	1986 (c)	1987 (d)
1	x	[1] Not Employed, Green	[1] Not Employed, Green	[1] Not Employed, Green
2	[2] Not Employed (First Surveyed) Not Applicable Since 94	[2] Not Employed (First Surveyed) Not Applicable Since 94	[2] Not Employed (First Surveyed) Not Applicable Since 94	[2] Not Employed (First Surveyed) Not Applicable Since 94
3	[3] Employed (First Surveyed) Not Applicable Since 94	[3] Employed (First Surveyed) Not Applicable Since 94	[3] Employed (First Surveyed) Not Applicable Since 94	[3] Employed (First Surveyed) Not Applicable Since 94
4	x	[4] Empl. Exc Change	[4] Empl. Exc Change	[4] Empl. Exc Change
5	x	[5] Empl. No Info If Change	[5] Empl. No Info If Change	[5] Empl. No Info If Change
6	x	[6] Empl. With Change, Also First Time Employment	[6] Empl. With Change, Also First Time Employment	[6] Empl. With Change, Also First Time Employment
7	x	x	x	x

By clicking on "N" to the left of the variable label, one can view the item "notes", giving an indication of the variable names of variables belonging to the item over all years.

Screen Shot 9: Item Notes

erwtyp09

```
Itemname    2264

Itemlabel   EN: Employment Status (generated)
            DE: generierter Erwerbstatus (Erwerbstyp)

Category    BD: Job Market And Occupation:
                Job Mobility
            BD: Arbeitsmarkt und Beschaeftigung:
                Berufliche Mobilitaet

JEL         J21

Itemvector  erwtyp84 erwtyp85 erwtyp86 erwtyp87 erwtyp88 erwtyp89 erwtyp90 -------- erwtyp91 --------
            erwtyp92 erwtyp93 erwtyp94 erwtyp95 erwtyp96 erwtyp97 erwtyp98 erwtyp99 erwtyp00 erwtyp01
            erwtyp02 erwtyp03 erwtyp04 erwtyp05 erwtyp06 erwtyp07 erwtyp08 erwtyp09
```

All words appearing in an item label have been added to a keyword thesaurus. Each keyword is linked to all items in the entire dataset containing the keyword.

Screen Shot 10: Start Page

Start Page > Entries for O <Project Page>

SOEP Keyword Browse Page for [O]
PanelWhiz SOEP (4.0) Sep 2010 <john@panelwhiz.eu>

[*****] [a] [b] [c] [d] [e] [f] [g] [h] [i] [j] [k] [l]
[m] [n] [o] [p] [q] [r] [s] [t] [u] [v] [w] [x] [y] [z]

occ	occasional	occupation	occupational
occupationally	occupied	occupiers	oct
october	off	offer	offered
office	offs	often	oil
okt	old	older	once
one	one's	oneself	only
onw	openly	operation	opinion
optimism	oral	order	orderd
organisation	organization	organizations	origin
original	orp	orpha	orphan
orphans	orthopaedist	other	others
out	outgoing	outside	outwork
over	overall	overnight	overtime
own	owner	ownership	

[occ] TOP
X N J - IK3796 EN: K.A. Item Nonresponse (Occ Status)

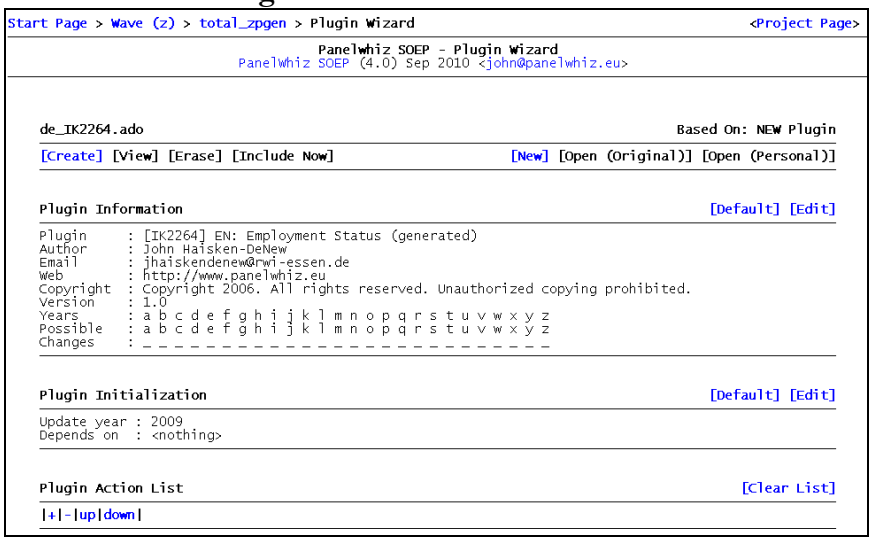
[occasional] TOP
X N J - IK299 EN: Occasional Work for Money

[occupation] TOP
X N J - IK2265 EN: Occupation in Field/Area where Educated (generated)
X N J - IK2279 EN: ISCO88-Occupation Code
X N J - IK229 EN: Occupation in Field/Area where Educated
X N J - IK255 EN: Starting Fresh in Different Occupation
X N J - IK3478 EN: Occupation of Individual
X N J - IK3846 EN: Nat.Stat.Office-Occupation (Infratest)
X N J - IK428 EN: Occupation Specific Insurance: Retirement Pension
X N J - IK438 EN: Occupation Specific Insurance: EU/BU Widow(er) Orphan
X N J - IK439 EN: Occupation Specific Insurance: Widow(er) Orphan EU/BU

2.4 Plugins

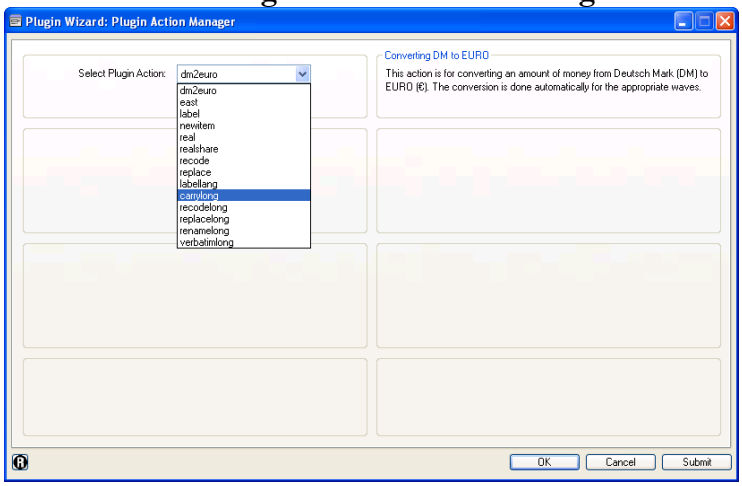
The variables underlying an item may vary of course from year to year. Using the PanelWhiz plugin system, the user can automatically create small scripts to clean time inconsistent data.

Screen Shot 11: Plugin Wizard



Depending on the dataset, various actions can be selected. For example, using plugins, nominal money values can automatically be deflated using an integrated CPI fuction.

Screen Shot 12: Plugin Wizard and Defining Functions



2.5 Retrievals

Once a project has been created, PanelWhiz has enough information to retrieve the actual data from the SOEP dataset. PanelWhiz dynamically creates a command file and executes it on-the-fly. PanelWhiz opens the files that the user specifically addressed and pulls out the variables specifically selected. It stores these variables in a temporary file and then moves on to the next data file. It does this many times until all data chucks have been extracted, then merges all data chunks together in the manner prescribed by the user.

The extracted data are automatically in Stata long format, ready to be processed using Stata's rich panel "xt" commands. To document exactly the retrieval that was run, PanelWhiz creates an executable DO file. The following Screen Shot shows an excerpt of a generated DO file.

Screen Shot 13: Excerpt of a generated DO file

```
                    /* ----------( create master )----------------------------------- */;

save                "$tmp/master", replace;
                    /* ----------( pull: hp / 1991 Person )--------------------------- */;

use                 hhnr      persnr
                    hp10901
using               "$soep/hp";

label               lang DE;

pwclone             hp10901   IK622;

drop                hp10901;

sort                persnr;
save                "$tmp/hp", replace;


                    /* ----------( pull: ip / 1992 Person )--------------------------- */;

use                 hhnr      persnr
                    ip10901
using               "$soep/ip";

label               lang DE;

pwclone             ip10901   IK622;

drop                ip10901;

sort                persnr;
save                "$tmp/ip", replace;
```

3. Summary

PanelWhiz is a data retrieval tool that eases data extractions from the many large scale data sets such as the German SOEP. PanelWhiz is directly combined into Stata SE, allowing a seamless interaction between the micro data and the statistics package. Vectors of variables, called item-correspondences can be selected all at once. Special cleaning programs written in Stata called "plugins" can clean a particular item-correspondence and make it time and/or content consistent. Groups of item-correspondences can be stored as projects. Groups of projects can be stored as libraries. This method of organizing the projects and plugins allows for a modular administration, facilitating knowledge transfer and group work. Data can be retrieved by mouse-click, providing rectangularized data in long format. As new releases of the micro data become available, the user can "automatically" update his projects to include the latest wave of information. All programs used are available in source Stata code which allows complete transparency of content. All commands used in the generated retrieval are documented in a full functional retrieval DO file, capable of recreating the identical retrieval at any time.

References

Haisken-DeNew, John P. (2001) "A Hitchhiker's Guide to the World's Household Panel Data Sets." *The Australian Economic Review*. 34(3), 356-366.

Haisken-DeNew, John P. and Joachim R. Frick (2005) "The Desktop Companion to the German Socio-Economic Panel Study", DIW Berlin, Germany