Schmollers Jahrbuch 130 (2010), 643 – 654
Duncker & Humblot, Berlin

PanelWhiz: Efficient Data Extraction of Complex Panel Data Sets – An Example Using the German SOEP

By John P. Haisken-DeNew and Markus H. Hahn

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 panel data can be very complex and vary dramatically as described in Haisken-DeNew (2001). Some panel data sets provide many files per year ("wide format"), differing by their population, or level of aggregation etc., creating many obstacles for researchers. If one wants to put together variables across time ("long format"), this is typically much more difficult, but ultimately the format which is required for estimation.

PanelWhiz is a collection of subroutines that allows researchers to use an intuitive "common" graphical interface for accessing many panel datasets directly within the statistical package Stata / SE 10 or better (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 panel 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 single wave, but rather in retrieving the variable for several waves, i.e. an item. 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 (vector of variables) 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 2006, now has several hundred registered users, using the common interface to access many different datasets, such as the German SOEP, British BHPS, the Australian HILDA, the German IAB Establishment Panel, the American CPS, etc. Recently, support has been extended to the American PSID. This paper describes using PanelWhiz for the German SOEP, providing specific examples for this panel data set. However, due to the generalized nature of PanelWhiz, the interface is almost completely identical for all other supported datasets and thus this paper can also be used as a general reference for other supported data sets.

2. Overview of PanelWhiz

2.1 Getting Started with PanelWhiz

PanelWhiz is described in detail on the PanelWhiz Website at http://www.panelwhiz.eu and has been presented several times at international data conferences, the UK Stata Users Group and the German Stata Users Group. Due to the extensive size of PanelWhiz, the user downloads only a very small startup Stata Add-On program from the PanelWhiz Website. This small startup program then downloads automatically the component parts required for the full installation over the internet. All component parts are stored in compressed format, and as such, typically require only one-tenth of the usual download time and are automatically expanded locally on the user's hard disk.

PanelWhiz is installed as a collection of Stata Add-On programs and loads every time Stata is started. For example in the following Screen 4 Shot 1, one can select by mouse click the desired data set to be supported.

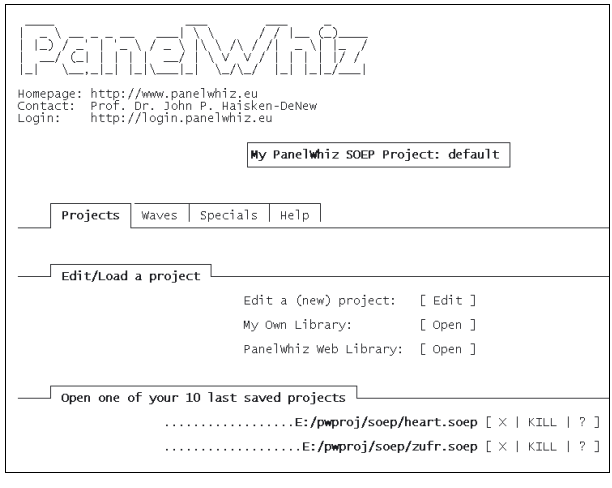
2.2 Some Details Using the German SOEP

In the example (Screen Shot 2), the German 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 keeps a note of the last 10 saved projects and allows easily loading by simply clicking on the link indicating the project name.

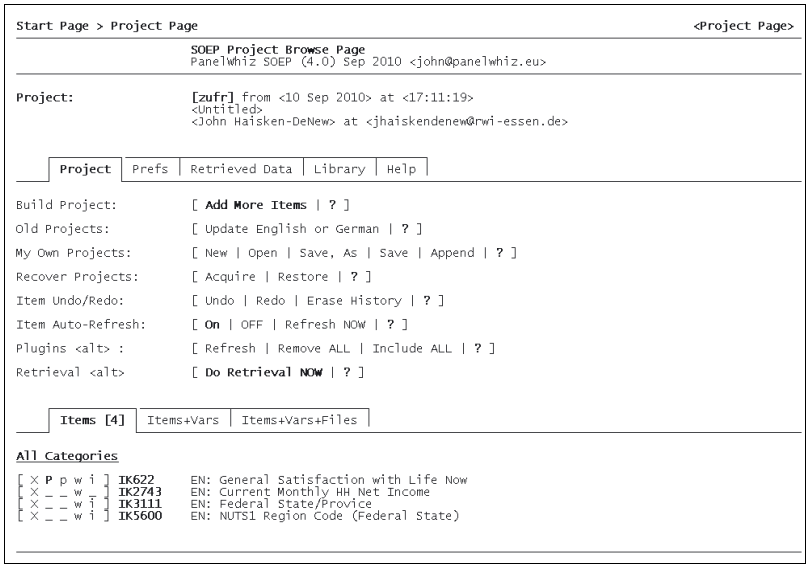
Here we have indeed opened the PanelWhiz project `zufr.soep`, and have a heads-up display indicating the contents of the project and the possible pro-

ject commands as seen in Screen Shot 3. 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 clickable, linked to a keyword thesaurus.

Screen Shot 1: Select Data Set



Screen Shot 2: Open a Project



Screen Shot 3: Project Page

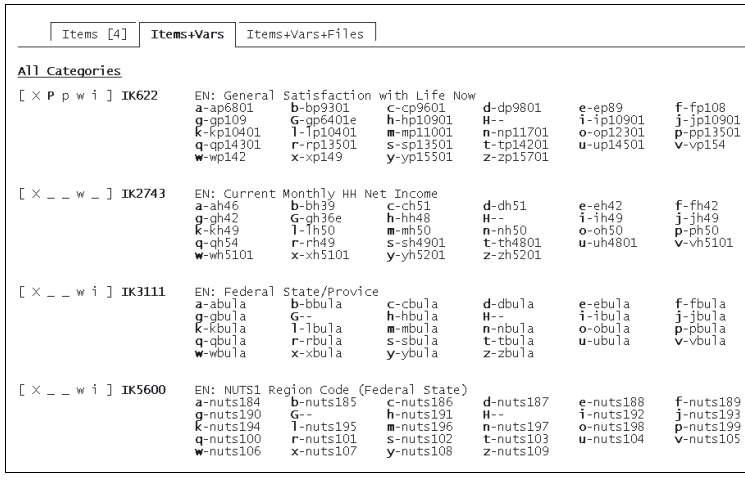
The variables underlying the 4 items in the project are listed below (Screen Shot 4). Each variable listed under the item displays the associated wave / year. In the SOEP, the "a" wave is 1984, the "b" wave is 1985 and so on. The "z" wave is 2009. Thus for the item `IK622`, in 1984 the underlying variable is `ap6801` and in 2009 it is `zp15701`.

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.

2.2.1 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.dta`, `apgen.dta`, `ah.dta`, `ahgen.dta` 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.[1] Here we will first examine the page associated with items.

[1] To learn more about the long or wide data format, see http://www.stata.com/help.cgi?reshape.



Screen Shot 4: Start Page

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).

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".

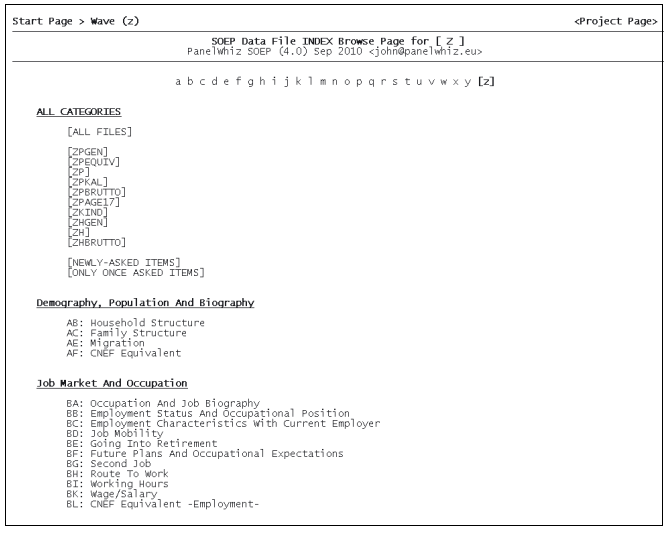
Technically speaking, PanelWhiz works because the item correspondence information (for each item, that vector of variables over all 26 years which are available currently for SOEP) is injected into each of the relevant variables as a Stata variable characteristic. PanelWhiz reads this information from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

2.2.2 Item Browse Page

To find an item available say for the year 2009, we click on the year [z – 2009]; and receive the following browse page (Screen Shot 6). The browse page contains all variables / items from all files in the year 2009. Alternatively,

Screen Shot 5: Items and Specials

one can jump only to specific variables / items in a particular file. Further, the variables for the German SOEP are sorted using the same hierarchical categorisation scheme as in SOEPinfo. See Haisken-DeNew and Frick (2005) for more information on SOEPinfo.

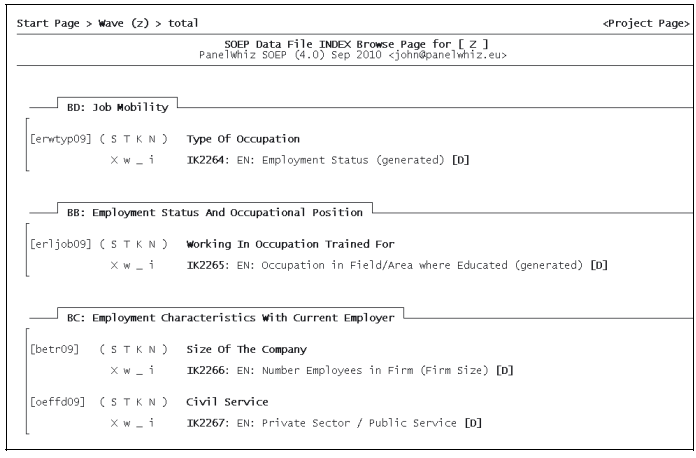


Screen Shot 6: Top Level Item Browse Page

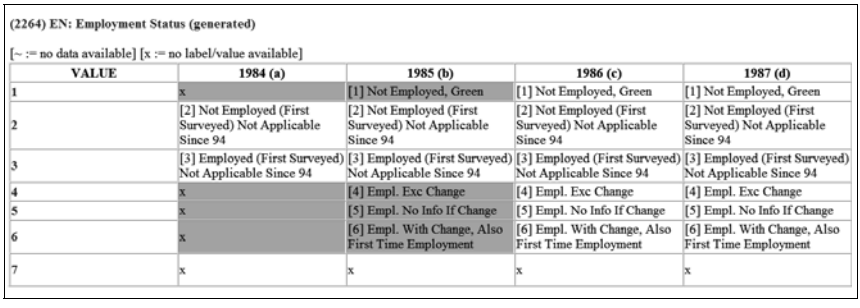
In this example, we select the clickable button [ALL FILES] from wave "z" and get the following browse page (Screen Shot 7). All variables are listed in the order they naturally exist in the respective physical data files. The example shows that for the SOEP variable erwtyp09, there is a PanelWhiz item IK2264 associated with it. By actually clicking on IK2264, one would select the entire item (potentially all underlying variables from wave "a" through "z").

One can also examine the changing nature of the item over time. The variable erwtyp09 contains value labels. Thus we can click on "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. Screen Shot 8 illustrates this example. Jumping from wave 1984 to 1985, there have been some additional outcome values added. These changes are colour coded in grey.

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).



Screen Shot 7: Look at Items

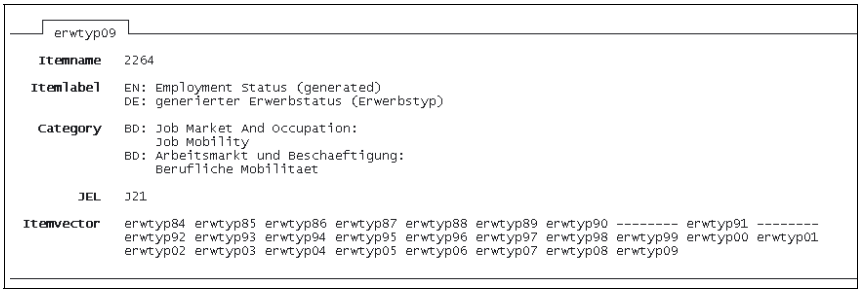


(2264) EN: Employment Status (generated)

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

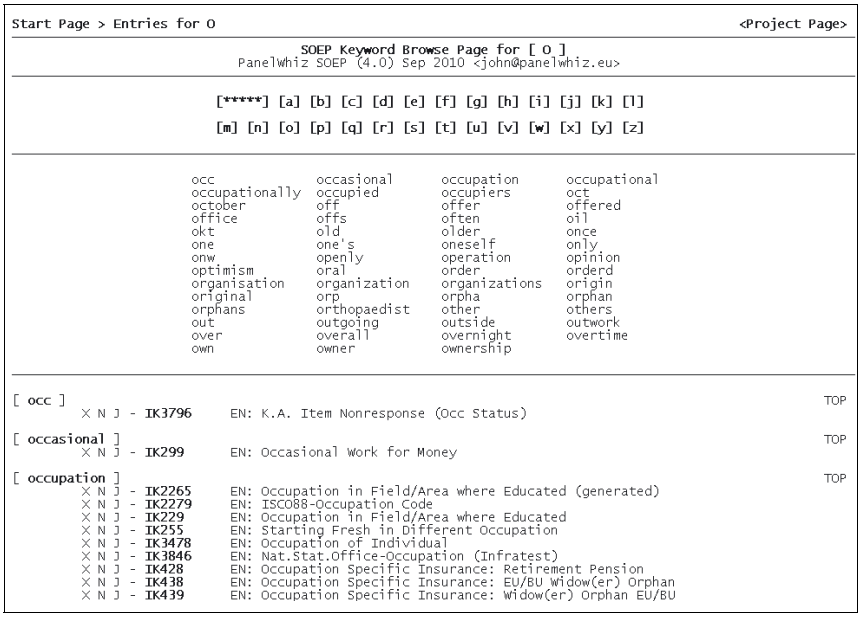
VALUE	1984 (a)	1985 (b)	1986 (c)	1987 (d)
1	x	[1] Not Employed, Green	[1] Not Employed, Green	[1] Not Employed, Green
2	[2] Not Employed (First Surveyed) Not Applicable Since 94	[2] Not Employed (First Surveyed) Not Applicable Since 94	[2] Not Employed (First Surveyed) Not Applicable Since 94	[2] Not 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

Screen Shot 8: Start Page



Screen Shot 9: Item Notes

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 (see Screen Shot 10).



Screen Shot 10: Keywords

2.2.3 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).

Depending on the dataset, various actions can be selected. For example (Screen Shot 12), using plugins, nominal money values can automatically be deflated using an integrated CPI fuction.

2.2.4 Retrievals

Once a project has been created, PanelWhiz has enough information to retrieve the actual data from the panel dataset, in this case the German SOEP. PanelWhiz dynamically creates a DO command file for Stata 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, and then merges all data chunks together in the manner prescribed by the user.

```
Start Page > Wave (z) > total_zpgen > Plugin Wizard                    <Project Page>

                              Panelwhiz SOEP - Plugin Wizard
                        PanelWhiz SOEP (4.0) Sep 2010 <john@panelwhiz.eu>


de_IK2264.ado                                                Based On: NEW Plugin
[Create] [View] [Erase] [Include Now]                [New] [Open (Original)] [Open (Personal)]

Plugin Information                                                   [Default] [Edit]
Plugin     : [IK2264] EN: Employment Status (generated)
Author     : John Haisken-DeNew
Email      : jhaiskendenew@rwi-essen.de
Web        : http://www.panelwhiz.eu
Copyright  : Copyright 2006. All rights reserved. Unauthorized copying prohibited.
Version    : 1.0
Years      : 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
Possible   : 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
Changes    : _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Plugin Initialization                                                [Default] [Edit]
Update year : 2009
Depends on  : <nothing>

Plugin Action List                                                   [Clear List]
|+|-|up|down|
```

Screen Shot 11: Plugin Wizard

Screen Shot 12: Plugin Wizard and Defining Functions

The extracted data are automatically in Stata long format, ready to be processed using Stata's rich panel "xt" commands. To document exactly the re-

trieval that was run, PanelWhiz creates an executable DO file on the fly. Screen Shot 13 shows an 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;
```

Screen Shot 13: Excerpt of a generated DO file

3. Summary

PanelWhiz is a data retrieval tool that simplifies data extractions from the many large scale data sets such as the German SOEP. PanelWhiz is directly combined into Stata / SE 10 or better, allowing a seamless interaction between the micro data and the statistics package. Entire vectors of variables, called "items" can be selected at once. Special cleaning programs written in Stata called "plugins" can clean a particular item and make it time and / or content consistent.

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 fully functional retrieval DO file, capable of recreating the identical retrieval at any time.

Groups of items 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 "wide" data in "long format". As new releases of the panel data set become available from the data provider, the user can "automatically" update his projects to include the latest wave of information.

References

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

Haisken-DeNew, J. P. / *Frick,* J. R. (2005): The Desktop Companion to the German Socio-Economic Panel Study, DIW Berlin, Germany.