DEPARTMENT OF FAMILY AND COMMUNITY SERVICES

THE UNIVERSITY OF MELBOURNE

HILDA PROJECT DISCUSSION PAPER SERIES
NO. 1/01, MARCH 2001

Structuring the HILDA Panel: Considerations and Suggestions

Joachim R. Frick and John P. Haisken-DeNew
German Socio-Economic Panel
German Institute for Economic Research (DIW-Berlin)

The opinions reported in this paper are the responsibility of the authors and should not be attributed to the Melbourne Institute of Applied Economic and Social Research, the Australian Council for Educational Research, the Australian Institute of Family Studies, or the Department of Family and Community Services.

The HILDA Project was initiated, and is funded, by the Commonwealth Department of Family and Community Services

MELBOURNE INSTITUTE of Applied Economic and Social Research

ACER

Australian Institute of Family Studies

1. Introduction

This paper deals with selected survey-related issues and the data structure of the HILDA (Household, Income and Labor Dynamics in Australia) Survey. Its purpose is to make suggestions as to the data structure to be implemented by the HILDA team, keeping in mind optimal user-friendliness and ease of data administration. We begin by briefly discussing the need to differentiate between survey items that are asked only once in a biography setting, and potential updates of those items asked yearly. We then propose, in Section 3, an ‘optimal’ data storage scheme for the HILDA data set, while discussing the nature and requirements of cross-sectional and longitudinal data. Further, a system of household and person identifiers, a variable naming scheme, and a file-naming scheme are proposed. In Section 4 we discuss the Cross-National Equivalent File project, and illustrate the substantial potential gain for the HILDA project in joining this consortium, which would immediately aid in increased scientific usage. Finally, we include an extensive appendix to illustrate various data structures.

2. Survey Related Issues

The HILDA questionnaire for individual respondents in Wave 1 includes a comprehensive section on biographical issues (marriages, work history, social background, parental information), which will become an additional instrument for first-time respondents entering the panel survey, beginning with Wave 2. This population is made up by: (a) new persons entering existing survey households; (b) persons living in households in which existing survey members move; and (c) young adults reaching respondent age of 15. While the first two groups constitute temporary sample members (TSMs) according to the currently discussed follow-up rules, the latter are most likely to be continuing sample members (CSMs). All respondents of wave 2, (i.e., those already responding in wave 1 as well as first-timers) will answer a common questionnaire. However, due to the time-dependency of some biographical data (e.g., marriages), the questionnaire of wave 2 also has to include questions targeted at changes, which may have occurred since last year’s interview. Included here are the marriage biography (to be updated using changes in marital status), children (to be updated from “process-produced” data collected in the household roster), education, and work history.

3. Data Related Issues

Data Structure

The following proposal for the future data structure of HILDA is based on the assumption of multi-wave data. Given that the starting wave of a new panel is largely just a cross-section, we want to illustrate certain features of a data structure that will pay off with each additional wave of data. However, we think it would be worthwhile to employ this structure from the very beginning, thus reducing the need for future restructuring.

Principles of a proposed data structure

In principle, according to the proposed data structure, any information as of a given year is stored in single files at the individual and household level, respectively. In both cases, there should be files consisting of the target (or "gross") population for which an interview was to be achieved, as well as for the population which was successfully interviewed ("net" population).

In order to support researchers to make use of the panel features of HILDA data, we suggest creating a set of "meta"-files as well as longitudinal files with biography data.

Cross-sectional files

The proposed set of cross-sectional files contains information on each year's data, thus making use of this data for simple cross-sectional analyses rather intuitive. These files include data collected by the interviewers (i.e., the surveyed data), as well as additionally derived information, which might also rely on information gathered in previous years. The overall structure is depicted in Figure 1, with the proposed file names starting with a $ specifying the wave (A for wave 1, B for wave 2, etc).

We suggest creating files consisting of the following.

(i) The target population (or "gross" population), with information gathered by the interviewer about basic characteristics on the household and the individuals living therein as well as on the field work process at the level of:

- households (proposed file name $HLOG); and
- individual household members (file $PLOG, including respondents, children and non-respondents).

(ii) The successfully interviewed population (or "net" population), with surveyed information at the level of:

- households (file $H); and
- individual respondents (file $P, population given by interviewed persons of 15 years of age and over);[1] and

with derived information (status variables and generated variables) on:

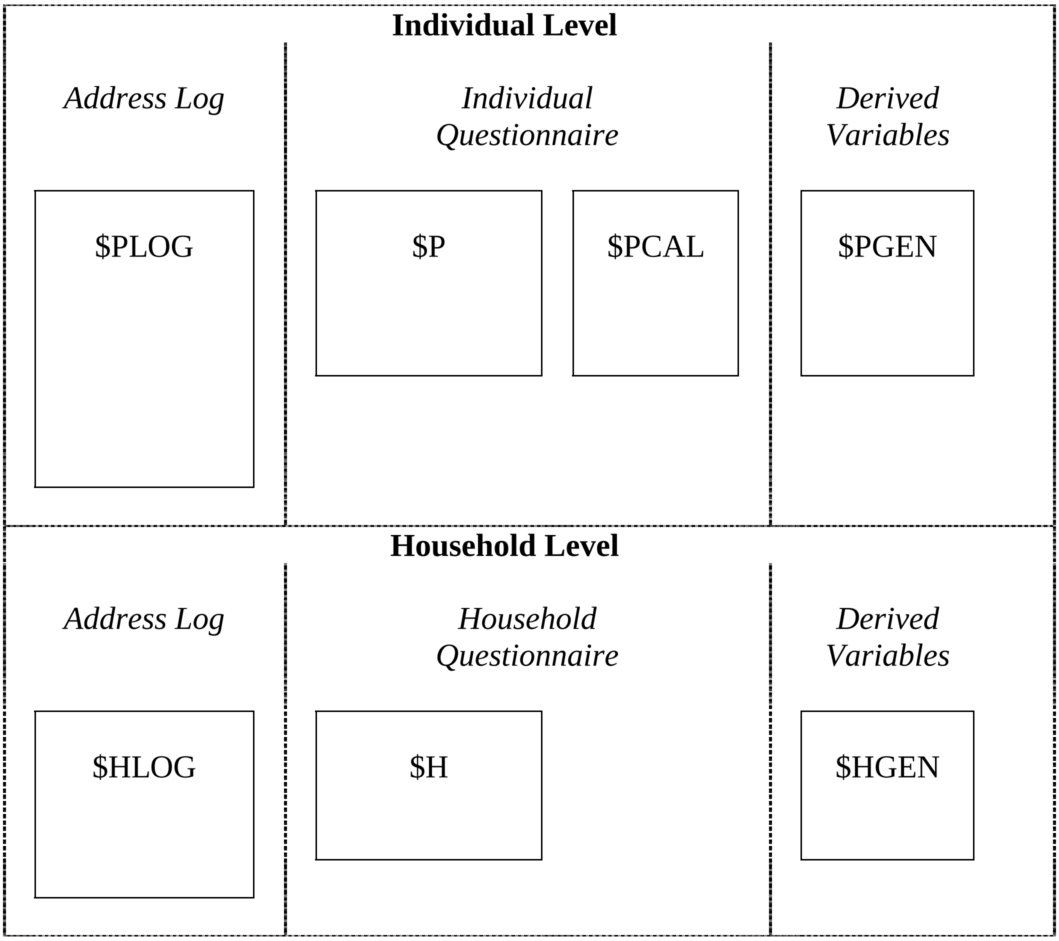
- households (file $HGEN); and
- individual respondents (file $PGEN).

Derived variables

The following considerations are thought to provide an idea as to why derived variables are very helpful for database management purposes as well as from a user's point of view.

[1] Given the presumably high number of variables necessary to cover the information collected in the activity calendar (3 time periods per month times 17 months from July 2000 through November 2001), we suggest storing these data in separate files ($PCAL).

Figure 1: The HILDA Data Structure - Yearly Cross-Sectional Files



$: Wave specification: A, B, C = wave 1, 2, 3

List of Cross-Sectional Files

$PLOG	All persons in target sample (respondents, children, non-respondents)
$P	All persons actually surveyed
$PCAL	Calendar information on previous year's monthly activities from those surveyed
$PGEN	Generated / derived variables from $P
$HLOG	All households in target sample
$H	All households actually surveyed
$HGEN	Generated / derived variables from $H

Status variables

Most panel surveys face the problem that:

(a) some information is gathered in the course of the first interview only;
(b) some information is asked for in separate questions for different subpopulations, respectively; and
(c) beginning with wave 2, in some cases old respondents are asked for changes since last year's interview only, while new respondents have to fill in the current status.

An example at the individual level is ‘Years with current employer’ and, at the household level, ‘Number of rooms in dwelling’.

A solution to these problems could be the following. In all these cases, the surveyed information is stored in different variables. In order to minimise computing efforts for the user, HILDA might provide yearly status variables on individual and household level, which integrate all of these information in a common variable showing the current status for all respondents. Thus, this involves nothing other than a re-organisation of already existing data. There is almost no assumption or normative setting involved in the generating process.

Generated variables

In addition to the above mentioned status variables, HILDA might provide generated variables for households and individuals, which require some assumptions as well. Again, the provision of these variables is targetted at enhancing the use of HILDA data by the scientific community. Examples are, at the individual level, ‘Institutional years necessary to receive current degree of education’ and, at the household level, ‘Household typology’.

Longitudinal files

“Meta-files”

These files contain information about the surveyed data as described above, rather than the survey data itself. For example, at the level of individuals, this file PMETA (see Figure 2) gives one variable for each year, showing a given person’s survey status and one variable indicating the household identifier of the household this person lived in. Additionally, it will prove to be very helpful if, at the individual level, this file contained basic demographic information as well (see Table 1). The population is made up of all observations ever contacted in the course of the HILDA-survey, thus the number of observations in this file will be cumulatively increasing year by year. Its principal purpose is to support longitudinal analyses by allowing one to restrict the data set to the sample of interest prior to matching to any cross-sectional file.

In detail, this file supports the definition of:

- the population of interest (basic demographics: year of birth, sex, immigrant status, year of death);
- the observation period; and
- the data structure: balanced vs. unbalanced panel design.

This file might also include “weighting factors”. Where to store these weights also depends on the definition of weights for longitudinal populations (i.e., it could be in a cross-sectional file as well, both at the household and individual level).

Table 1: Suggested list of variables in Meta-File for Individuals "PMETA"

Variable name	***Meaning***
HIDFIX	Original household identifier (case) from wave 1
PID	Unique individual identifier (time invariant)
SEX	Gender (longitudinally verified)
YBIRTH	Year of birth (4 digit) longitudinally verified
HID01	Household identifier 2001
HID02	Household identifier 2002
HID03	Household identifier 2003
...	
HID$$	Household identifier year $$
SSTAT01	Survey status 2001
SSTAT02	Survey status 2002
SSTAT03	Survey status 2003
...	
SSTAT$$	Survey status year $$
ESTAT01	Employment status 2001
ESTAT02	Employment status 2002
ESTAT03	Employment status 2003
...	
ESTAT$$	Employment status year $$
YENTRY	Year in which individual entered the survey (4 digit)
YFIRST	Year in which first individual interview was conducted (4 digit)
YEXIT	Year in which individual left the survey (4 digit)
YLAST	Year in which last individual interview was conducted (4 digit)
YDEATH	Year of death (4-digit)
YIMMIG	Year of first immigration to Australia (4 digit)
OZBORN	Born in Australia
CORIGIN	Country of origin

Biography Files

“Bio-files” contain biographical information (work history, parental and marriage/partnership biography, social background, etc.) gathered in the biography section of the wave 1 questionnaire. The population covered by these files will at first be identical with the cross-section of wave 1 and will cumulatively increase due to the inclusion of first-time respondents beginning with wave 2.

Attention should be given to the need for potential updating of biographical information, given it can be time-independent (e.g. the year of first immigration to Australia, or occupation of father/mother when respondent was 14 years of age), in which case no update is necessary, or time-dependent with a potential need for updates (like in case of first or repeated marriages).

We suggest breaking down biography data in three topic-related files.

- BIOBIRTH for information on own and adopted children.
- BIOPAREN for information on parents (education, labor market experience, etc.).
- BIOWORK for information on individual labor market entry and work history.

Concerning biographic information on marriages we suggest alternatively storing this information from the biography module in spell form. In principle, a spell-system

with h = household 1, ... , m
i = individual 1, ... , n
s = spell 1, ... , o

can be stored in the following way:

HIDFIX	PID	SPELLID	SPELLTYP	BEGIN	END	CENSOR
h_1	i_{11}	S_{111}				
h_1	i_{11}	...				
h_1	i_{11}	S_{11o}				
h_1	...	...				
h_1	i_{1n}	S_{1n1}				
h_1	i_{1n}	...				
h_1	i_{1n}	S_{1no}				
...	...	...				
...	...	...				
...	...	...				
h_m	i_{m1}	s_{m11}				
h_m	i_{m1}	...				
h_m	i_{m1}	s_{m1o}				
h_m	...	...				
h_m	i_{mn}	s_{mn1}				
h_m	i_{mn}	...				
h_m	i_{mn}	s_{mno}				

Below is an example for the successful conversion of biography-data in “user-friendly” spell-data taken from the SOEP-file BIOMARSY (i.e. the marital status information for each respondent based on data collected in the biography section of wave 1 and updated using consecutive data from yearly interviews since then). This file gives information on a yearly basis, measuring begin and end of each marital status spell in years of age (BEGIN of very first spell: age = 0).

Variable SPELLTYP

value	Marital status
(1)	Single
(2)	Married
(3)	Divorced
(4)	Widowed
(5)	separated (no differentiation possible between divorced and widowed)
(8)	missing because of item-non-response
(9)	missing because of unit-non-response

HIDFIX	PID	SPELLNR	SPELLTYP	BEGIN	END
19	101	1	1	0	24
19	101	2	2	24	28
19	101	3	3	28	32
19	101	4	2	32	59
19	102	1	1	0	22
19	102	2	2	22	49
19	103	1	1	0	24
27	201	1	1	0	27
27	201	2	2	27	39
27	201	3	3	39	71
27	202	1	1	0	31
27	203	1	1	0	37
35	301	1	1	0	24
35	301	2	2	24	33
35	302	1	1	0	23
35	302	2	2	23	32

Occupational calendar

Calendar data on occupational activities per month is often stored in time-series format (see suggested file $PCAL). An alternative way of storing the same information is again in spell-format, which has several advantages.

- It is very efficient in terms of number of variables as well as disk space.
- It is straightforward to use for duration analysis.
- Mismatching information in the case of overlapping periods can easily be handled. In fact, this will most likely happen in wave 2 of HILDA, when the recall period of the calendar will start in July 2001 (i.e. overlapping the recall period of wave 1 for up to five months.
- It is easy to generate a variable indicating the censor status for each spell (i.e., telling whether this observation is left- and/or right-censored).

Event data

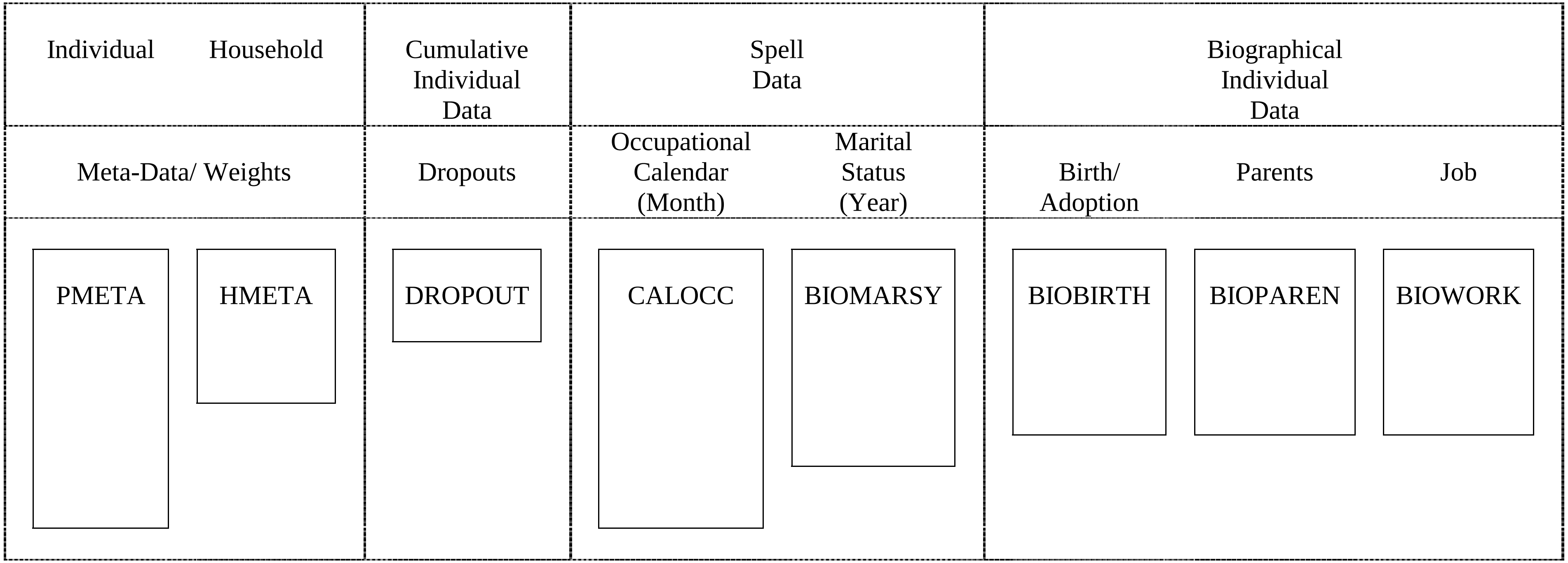
Panel data is explicitly suited to cover demographic events such as residential mobility. Although in the course of time the absolute number of observations might be decreasing from a cross-sectional perspective, the cumulative number of events covered with this data is increasing wave by wave.

In order to support event analysis, we suggest creating the file DROPOUT, which cumulatively covers drop-outs from survey households due to:

- death;
- emigration (out-migration); and
- split-off from an existing survey household and move into another household within survey territory (residential mobility).

Since a given individual might show up in such a file more than once (e.g., because of moving into a new household from wave 1 to wave 2, and emigrating from wave 2 to wave 3), a unique identification is required for each given observation per individual. We suggest the inclusion of an additional identifier YEVENT, giving the year in which this event was covered by the data.

Figure 2: The HILDA Data Structure - Longitudinal Files



PMETA All persons ever surveyed for all years, could include weighting information as well
HMETA All households ever surveyed for all years, could include weighting information as well
BIOMARSY Biography: Yearly Marital Status in spell form
BIOBIRTH Biography: Fathers/Mothers info on own/adopted kids
BIOPAREN Biography: Info on persons Father/Mother
BIOWORK Biography: Info on labour market entry, and work history (# of years in employment, in unemployment, not in employment)
CALOCC Occupational Calendar in spell form
DROPOUT List of all persons dropping out of survey due to death and emigration, and household splitt-offs (residential moves)

Identifiers

A set of variables is necessary in order to ensure unique matching and merging across files and waves. In principle, each observation within a panel survey can be traced back to the original household of wave 1. This hierarchical order should be maintained within the set of identifiers as this not only helps to check database integrity, but also allows users to match easily information from persons with a common background. Indeed, we suggest adding this original wave 1 household identifier to each single data file. Additionally, the following identifiers are necessary at individual, household, and event level:

Person level (cross-sectional files)

- Individual Identifier (suggested variable name PID, which needs to be fixed over time).
- Original Household Identifier as of Wave 1 (HIDFIX, fixed over time)
- Wave specific Household Identifier (HID$$, with $$=wave specifying suffix ‘01’ for wave 1, ‘02’ for wave 2, etc.). This variable is necessary to identify all members of a given household in a given wave.

Person level (longitudinal file PMETA)

- Individual Identifier (PID, fixed over time).
- Original Household Identifier as of Wave 1 (HIDFIX, fixed over time).

Household level

- Individual Household Identifier (HID, fixed over the “life” of the household, crucially depending on the definition of a “longitudinal” household, see above).
- Original Household Identifier as of Wave 1 (HIDFIX, fixed over time).

Event level

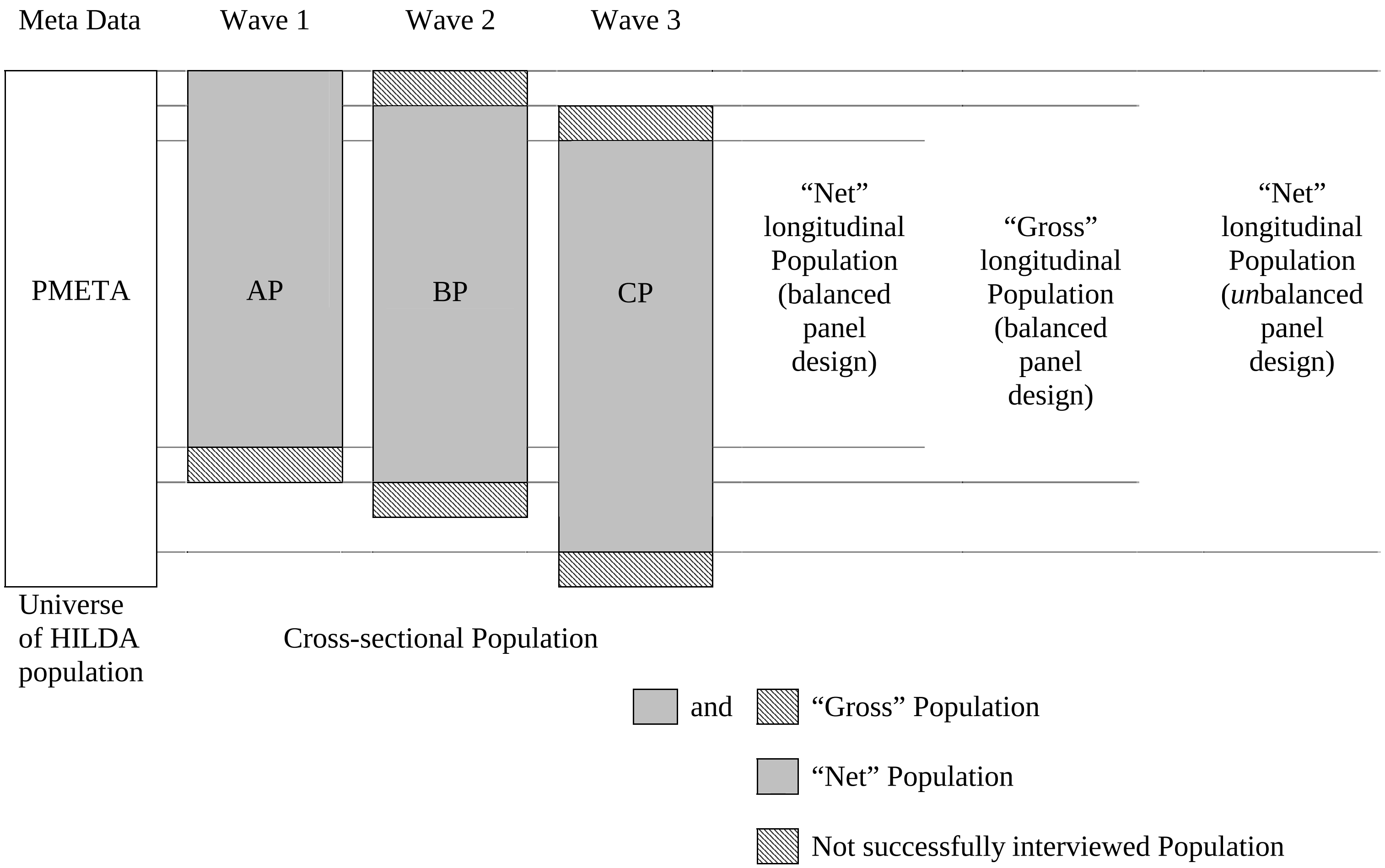
- Event Identifier (SPELLID).
- Individual Identifier (PID, fixed over time).
- Original Household Identifier as of Wave 1 (HIDFIX, fixed over time).

Biography data

- Individual Identifier (PID, fixed over time).
- Original Household Identifier as of Wave 1 (HIDFIX, fixed over time).
- Although not explicitly necessary, we suggest adding a variable indicating the year/wave in which the biography data has been collected (variable BIOYY with 2001, 2002, etc.).

Finally, Figure 3 gives an example for how the longitudinal meta-file PMETA (including all persons ever contacted over the whole panel period, which in this illustration is three years) relates to wave-specific data (files $P) and how it can be used to pre-define the population of interest for a given analysis with respect to data structure (balanced vs. unbalanced panel design).

Figure 3: The HILDA Data Structure - Cross-Sectional and Longitudinal Files


Meta Data
Wave 1
Wave 2
Wave 3
PMETA
AP
BP
CP
"Net" longitudinal Population (balanced panel design)
"Gross" longitudinal Population (balanced panel design)
"Net" longitudinal Population (*un*balanced panel design)
Universe of HILDA population
Cross-sectional Population
and
"Gross" Population
"Net" Population
Not successfully interviewed Population

Longitudinal Households

From a cross-sectional perspective, the definition of a household is rather straightforward. It is a single person or a group of persons living together and sharing resources. However, there are arguments about what constitutes a "longitudinal" household, given that the address or the composition of a household might change over time. As such, the question arises whether a household remains the same unit, given that individual household members may have left, due to death or moving out. In contrast, new household members may have entered the household, such as a new birth or persons moving in. A very important question along these lines is the definition of an appropriate identifier (i.e., the definition of the variables HID and HID$$ as mentioned earlier). It is intuitive, that a very rigid definition of a longitudinal household yields a large number of newly defined household IDs for every new wave of data. It also requires incorporating the follow-up rules into the definition of these IDs.

As an example of how these issues can be taken care of, we depict the rules employed in the case of the German SOEP. General rules include:

- by definition, household IDs of wave 1 (proposed variable name for HILDA is HIDFIX) are fixed over time;
- new households, and as such, new household Ids, evolve only in case of split-offs from existing survey households;
- a new household receives the HIDFIX of the household from which the split-off occurred; and
- as is the case for the individual PID, a HID is "permanently retired" if a household dissolves (e.g. in case of death of a one-person-household).

Event occurring from wave 1 to wave 2	**HID wave 2**	**HIDFIX**
(A) Household Composition		
• No change	No change	Fixed
• New household members enter	No change	Fixed
• Old household members leave	No change	Fixed
(B) Address		
• No change	No change	Fixed
• All members move together to a new address	No change	Fixed
• Household split with one partition staying at the old address:		
o Partition with previous year's address	No change	Fixed
o Partition with new address	New HID	Fixed
• Household split with both partitions moving to a new address:		
o Partition with previous year's head	No change	Fixed
o Other partition	New HID	Fixed

As a result of these rules, as well as the less rigid SOEP follow-up procedures, after several years a ‘household’ might consist of totally different persons than in the first wave.

Data Storage vs Data Distribution

Although we suggest storing these files in a “compartmentalized” manner (i.e., person information only for one year in a single file), information for several years can be joined together in a very straightforward manner for distribution. Thus files can essentially be stored/administered as cross-sectional files, but distributed to the user as “ready to go” longitudinal files (in “long” format, as opposed to "wide"). That will have to be a policy decision of the HILDA group.

Variable Naming

There is a variety of variable naming conventions. The final selection of a principle for naming variables also depends on the chosen file structure. It is intuitive not to have the identical name for a variable in a person and a household level file at the same time, which might require one to add respective information to the variable name specifying the level of observation. If a cross-sectional file structure is chosen, then a year or wave specification is important. Here one could use a one-digit alphabetical code (‘A’ for wave 1, ‘B’ for wave 2, etc.) or a two-digit year code (‘01’ for 2001, ‘02’ for 2002, etc.) or, looking ahead to wave 10, a two-digit wave specification as prefix or suffix to the variable name. In the case of HILDA’s first wave being conducted in 2001, both of these approaches would yield the suffix ‘01’.

Given all these restrictions, the following principles remain to be a matter of choice, only. However, the naming schemes resemble the practice of major existing panel surveys.

(i) The variable name corresponds to the number of the question in the original instrument used in the survey, e.g. the individual questionnaire. Additionally, a specification for observation year or wave of survey is required.

- Example: AP01 would be wave A, Person level, question 01.
- Advantage: Easy for user to relate to original survey instrument in order to look up the exact wording of the question.
- Disadvantage: Due to changes in the questionnaires over time it is almost by definition that variable names change from one wave to the next.
- Employed in SOEP.

(ii) The variable has a “speaking” name, meaning that the name basically tells what information is stored in the variable. This name remains constant over time, except for a suffix of prefix defining the observation year or the wave of the survey.

- Example: PWAGE01 would be Person level, Wage in year 2001 or WAGE01 would be Wage in year 2001.

- Advantage: Easy for user to find corresponding information across waves.
- Disadvantage: for more complicated issues it might be hard to define an easy to understand “speaking” name (highly normative).
- Employed in BHPS and for derived variables in SOEP.

(iii) The variable name follows in a sequence within a larger topic (like income or employment), meaning that each major topic has variables to be counted through starting from 1 to whatever the last variable within this topic is. New variables of following waves would just add to the existing variables no matter if these are still in the survey or not. As such, this name remains constant over time, except for a suffix defining the observation year or the wave of the survey.

- Example: IP001201 would be section I (Income), Person level, the 12th variable.
- Employed in Cross-National-Equivalent File (CNEF).

(iv) All variables are counted through starting from 1 to whatever the last variable is. Again, these names could remain constant over time, except for a suffix defining the observation year or the wave of the survey.

- Example: V12301 would be the 12301th variable.
- Disadvantage: basically no structure at all – clearly inferior to any of the above.
- Employed in PSID.

Newest versions of software packages like Stata and SAS are able to handle variable names with a length of up to 16 characters, which allows rather detailed information to show in the variable names. However, it must be assumed that most potential users are working with older versions, which do not support this feature. Thus, we suggest starting out with the standard length of 8 characters.

Definition of Missing Values

Depending on the information collected in the survey and the employed data structure there are several types of missing values. Following the rules used in SOEP-data we suggest using negative values to code missing data (examples given in parenthesis) as to clearly differentiate these from valid information.

(i) User defined Missing Values

- Item-non-response: “Do not know” (code ‘-1’).
- Not applicable (code ‘-2’) [Example: Monthly Rent to be paid if respondent is owner occupier].
- A presumably valid value was deleted after extensive consistency checks, which could not be done during field work (code ‘-3’) [to be interpreted like code ‘-1’].

(ii) System defined Missing Values

- These should generally be avoided (most likely to be BLANK).

4. International Comparability

An effective use of HILDA data by the national and international scientific research community is most likely to be a very important indicator for success, and as such, for further funding as well. To improve the use by international scholars we suggest deriving information, which is cross-nationally comparable to other panel data sets. It seems most obvious to include HILDA data into the Cross-National-Equivalent-File (CNEF), which currently contains data from the US Panel Study of Income Dynamics (PSID), the British Household Panel Study (BHPS), the German Socio-Economic Panel Study (SOEP) and the Canadian Survey of Labor and Income Dynamics (SLID). Certainly with three other data sets from English speaking countries, the natural choice of countries for international comparative research is obvious.

For further information on the CNEF see:
http://www.human.cornell.edu/pam/gsoep/equivfil.cfm

Following, in Table 2, is a list of the 1999 distribution of CNEF data which, in principle, is reproducible for HILDA, given the currently proposed questionnaire.

Table 2: CNEF Variable List, 1999

Variable Name	***Label***
Identifiers (Equivalent file)	
X11101LL	Unique Person Number
X11102$$	Household Identification Number
X11103$$	Individual in Household at Survey
X11104LL	Oversample Identifier
Demographic indicators	
D11101$$	Age of Individual
D11102LL	Gender of Individual
D11103$$	Race of Household Head
D11104$$	Marital Status of Individual
D11105$$	Relationship to Household Head
D11106$$	Number of Persons in Household
D11107$$	Number of Children in Household
D11108$$	Education With Respect to High School
D11109$$	Number of Years of Education
D11110$$	Disability Status of Individual
D11111$$	Satisfaction With Health
Employment indicators	
E11101$$	Annual Work Hours of Individual
E11102$$	Employment Status of Individual
E11103$$	Employment Level of Individual
E11104$$	Primary Activity of Individual
E11105$$	Occupation of Individual
E11106$$	1 Digit Industry Code of Individual
E11107$$	2 Digit Industry Code of Individual

Table 2 (cont'd)

Variable Name	***Label***
Income indicators	
I11101$$	Household Pre-Government Income
I11102$$	Household Post-Government Income
I11103$$	Household Labor Income
I11104$$	Household Asset Income
I11105$$	Household Imputed Rental Value
I11106$$	Household Private Transfers
I11107$$	Household Public Transfers
I11108$$	Household Social Security Pensions
I11109$$	Total Household Taxes
I11110$$	Individual Labor Earnings
I11111$$	Household Federal Taxes
I11112$$	Household Social Security Taxes
I11201$$	Impute Household Pre-Government Income
I11202$$	Impute Household Post-Government Income
I11203$$	Impute Household Labor Income
I11204$$	Impute Household Asset Income
I11205$$	Impute Household Imputed Rental Value
I11206$$	Impute Household Private Transfers
I11207$$	Impute Household Public Transfers
I11208$$	Impute Household Social Security Pensions
I11209$$	Impute Total Household Taxes
I11210$$	Impute Individual Labor Earnings
Weights	
W11101$$	Individual Weight
W11102$$	Household Weight
W11103$$	Longitudinal Weight
W11104$$	Population Factor
W11105$$	Individual Weight - Immigrant Sample
W11106$$	Household Weight - Immigrant Sample
W11110$$	Detailed Official U.S. Equivalence Weight
W11111$$	General Official U.S. Equivalence Weight
W11112$$	Official German Equivalence Weight
W11113$$	ELES Equivalence Weight
W11114$$	OECD Equivalence Weight

$$ = year of observation (e.g. '88' for 1988)

UPDATE-ALERT Beginning with the data distribution as of 2001, new variables will be added describing household composition. Variables on alternative equivalence scales will be dropped.

5. Conclusions

We see tremendous potential in the HILDA data set, not only as the HILDA project is able to benefit from several decades of experience from other panel data set providers in setting up their data, but also as Australia is now able to provide a longitudinal data set, comparable to those in the US, Germany, Britain and Canada. By choosing an efficient data structure now, which will be both easy to administer for the HILDA staff and also straightforward to use by Australian and international researchers, the HILDA project will get going quickly and maximise its success.

APPENDIX

The following figures depict popular data structures used for analyses of panel data: cross-section, repeated cross-section, pooled cross-section as well as longitudinal data structures like balanced and unbalanced panel design, the latter being used in “wide” format as well as “long” format (pooled data).

Basically, the data structure suggested in this paper for the storage of HILDA data, enables researchers to produce any of these structures without major difficulties.

A. Preparing Data for Analysis: Cross-Sectional Structure

- Single cross-section data
- Repeated cross-sectional datasets
- Pooling of two cross-sectional datasets

B. Preparing Data for Analysis: Longitudinal Structure

- Complete case analysis with a balanced panel design
- Downstream model (cohort)
- Complete information analysis with an unbalanced panel design
- Pooling longitudinal data (two longitudinal datasets of two waves each)

A. Preparing data for analysis: Cross-Sectional Structure

a) Single cross-section data (*i*=individuals *1, ... , n* ; *v*=variables *1, ... , m*)

	V_1	.	.	.			V_m
i_1							
.							
.							
.							
i_n							

b) Comparison of two cross-sectional datasets (for *t* = time period *1, 2*)

	V_{11}	.	.	.			V_{m1}
i_{11}							
.							
.							
.							
i_{n1}							

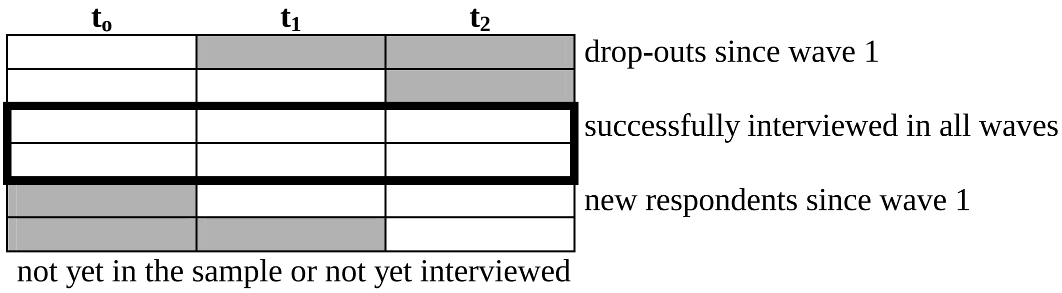
	V_{12}	.	.	.			V_{m2}
i_{12}							
.							
.							
.							
i_{n2}							

c) Pooling of two cross-sectional datasets

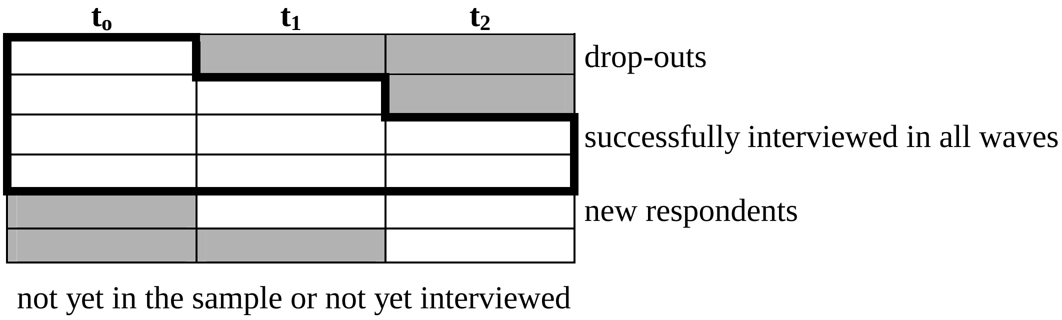
	V_{1t}	.	.	.			V_{mt}	t
i_{11}								1
.								.
.								.
.								.
i_{n1}								
i_{12}								2
.								.
.								.
.								.
i_{n2}								

B. Preparing data for analysis: Longitudinal structure

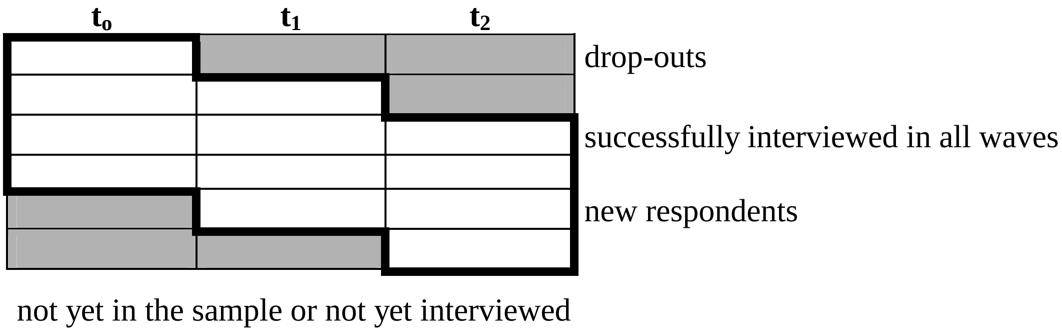
a) Complete case analysis with a balanced panel design


t_0
t_1
t_2
drop-outs since wave 1
successfully interviewed in all waves
new respondents since wave 1
not yet in the sample or not yet interviewed

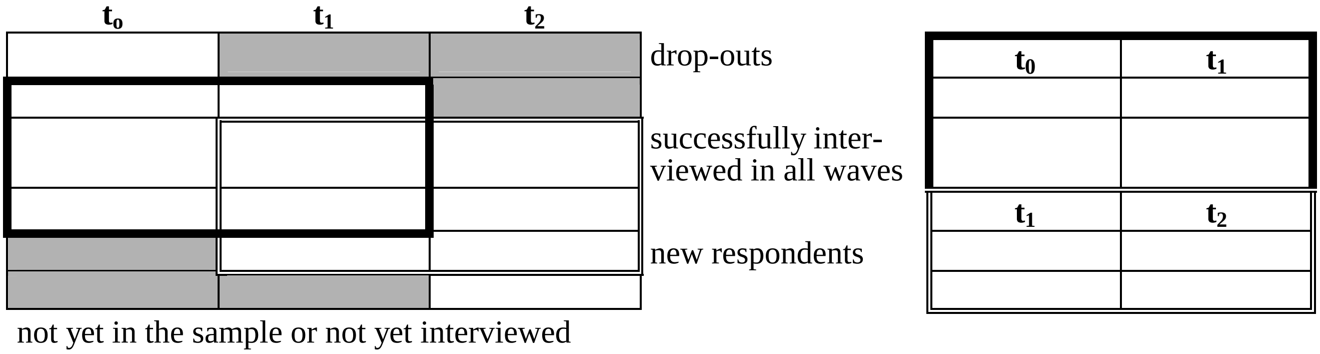
b) Downstream model keeping population as of t_0 constant (cohort)


t_0
t_1
t_2
drop-outs
successfully interviewed in all waves
new respondents
not yet in the sample or not yet interviewed

c) Complete information analysis with an unbalanced panel design


t_0
t_1
t_2
drop-outs
successfully interviewed in all waves
new respondents
not yet in the sample or not yet interviewed

d) Pooling longitudinal data (two longitudinal datasets of two waves each)


t_0
t_1
t_2
drop-outs
successfully inter-viewed in all waves
new respondents
not yet in the sample or not yet interviewed
t_0
t_1
t_1
t_2