The *LWS database:* user guide

The *LWS Database* focuses on wealth and debt of households. The goal of *LWS* is to enhance studies on understanding of households' financial stability through both the analyses of wealth distribution and other related dimensions of economic well-being. It draws upon the *LWS* project that started already in 2007; that pilot version of the wealth database facilitated the exchange of expertise on wealth among scholars in that field, and as a consequence, the new framework of *LWS* has been introduced.

Generic information

The *LWS Database* consists of several *ex post* harmonised datasets from different countries and years. Each dataset refers to one specific country at one point in time, and is named according to the 2-character country abbreviation coded according to the International Standard for country codes (*ISO-3166*) and the reference year. The reference year is the calendar year, which includes the data point to which wealth data refers to. Please note that the reference year may differ from the year used by the data provider in the official documents. For the exact reference period of the values of wealth and income variables, please consult the *Survey Information* document that is available for each specific dataset on the LIS website.

Structure of the LWS datasets - Each *LWS* dataset contains at least two files, namely the household level file and the individual level file; additionally for some datasets, there might also be a replicate weights file.

- *Household level file (LWS H-file)* – This file includes all household level variables (see technical, household characteristics, household balance sheet and other wealth, behavioural, and flow variables reported under the heading H-file of the *LWS Variables List*). The unit of this file corresponds to the survey unit, which, for wealth surveys typically is the household or the primary economic unit, but may differ in some cases (e.g. tax unit, family unit, etc. – see variable *SVYUNIT* in *LWS Variables Definitions* for more information); irrespective of the exact definition of the survey unit, the unit of the file is always referred to as “household”. The *LWS* H-file contains at least one observation for each household of the sample. If the data provider conducted a multiple imputation procedure to impute missing values, all imputation replicates (implicates) are included in this file. Since the imputations are usually stored as five successive implicates of each record, the number of observations in the file is five times the actual number of households. This file is uniquely identified by the household identifier as well as the implicate number (variables *HID* and *INUM*). If there was no imputation or a single imputation procedure applied by the data provider instead of multiple imputation, then the variance of variable *INUM* is equal to 0.
- *Individual level file (LWS P-file)* – This file includes all individual level variables (see technical, socio-demographic, labour market, pension asset, behavioural, and flow variables reported under the heading P-file of the *LWS Variables List*). The unit of the file is the individual, and the P-file includes at least one record for each individual

belonging to the household (even if the data were originally not reported for all household members). If the data provider conducted a multiple imputation procedure to impute missing values, all implicates are included in this file. This file is uniquely identified by the household identifier, the person identifier, and the implicate number (variables *HID*, *PID*, and *INUM*). If there was no imputation or a single imputation procedure applied by the data provider instead of multiple imputation, then the variance of variable *INUM* is equal to 0.

- *Replicate weights file (LWS R-file)* – This file contains as many replicate weights as the data producer has created (up to 1,000) and it includes exactly one observation for each household of the sample. By using the replicate weights, users obtain more precise confidence intervals and significance tests; in other words, theoretically the replicate weights generate more informed empirically derived standard error estimates by simulating multiple samples form a single sample. This file is uniquely identified by the household identifier.

Note that all *LWS* H-files include the same variables in all datasets (i.e. all variables are actually present even if they contain only the missing values). The same holds for all *LWS* P-files.

Variable standardisation - The *LWS* variables are standardised in terms of conceptual content (the variables are as comparable as possible across datasets in terms of concepts/definitions, see *LWS Variables Definitions*) as well as in terms of coding structure:

- All continuous variables are standardised and report information expressed in the same unit across different datasets. *LWS* wealth and flow variables are reported in annual amounts and in units of the national currency in force at the time of the data collection (see variable *CURRENCY*).
- Most categorical variables are standardised and report information expressed with the same value codes and labels.

There are also some categorical variables that are not standardised (variables denoted by a "*_C*" suffix). While the variable name is the same across all datasets, the variable label may differ to indicate the actual (dataset-specific) contents for the dataset in question. Both the exact contents and the coding structure will differ across datasets (see the *LWS Codebooks*).

The standardisation of some categorical variables follows a two- or multi-digit coding structure, whereby the codes starting with the same digit belong to the same overall category (e.g. the codes in the 10s for variable *BASB* – saving behaviour – implies that a person does not save, while the codes 11 and 12 are subcategories of that overall category, namely "does not save: expenses exceed income", and 12 "does not save: expenses about the same as income"). Please note that in one and the same dataset observations can be coded either at the higher level or at the lower level. For instance, some persons may have been coded 11, others 12, yet others directly with 10 ("does not save") in case it was not possible to determine whether a person does not save because expenses exceed income or a person does not save because his/her expenses are about the same as income. As a result, by selecting all the

subcategories, one does not necessarily get the total of one overall category (i.e. by selecting codes 11 and 12, one does not necessarily select all persons who do not save; the selection of the higher code 10 is needed as well).

Generic missing values policy **-** In the *LWS* database the system missing value ('dot') represents all cases of observations for which the information is not available. This includes both, the cases where the information is not applicable (e.g. the person does not work and hence cannot have an industry), and the cases where the information is applicable, but not available. The latter case includes the following situations:

- the information is applicable, but has not been collected by the data provider for a given subset of the sample or at all (e.g. only heads and their spouses/partners were routed through the individual level questionnaire);
- the information is applicable and has been collected by the data provider, but the respondent did not answer (don't know/refusal);
- the information is applicable and has been collected, but is not available at the level of detail necessary for the *LWS* variable in question (e.g. the vehicle loans - variable HLNCV - has all missing values if all consumer goods loans are collected together).

The number of "applicable missings" (as described in the three situations above) can vary substantially across datasets. This is due to the different collection practices as well as imputation practices adopted by the data producers. The *LWS* data include imputations only if they were available in the original data provided by the data producer.

Weights **-** All *LWS* files include weight variables (at the household level in the H- and R-files and at the individual level in the P-file) that are needed to make the sample representative of the overall population. The weights included in the *LWS* files are those calculated by each data producer. All of them correct at least for sampling bias, but not necessarily correct for unit or item non-response bias, or inflate to covered population (please refer to the *Survey Information* for each *LWS* dataset for more information on weights). If the weight originally provided by the data producer (as reported in variables *HWGT* and *PGWT*) does not inflate to the total population, then an inflation to total national population is carried out by LIS. Thus, the weights reported in variables *HPOPWGT* and *PPOPWGT* are all normalised to population size. Further, if the data producer only provided an individual level weight, then the household level weight reports the average weight of the individuals in the household. If only the household level weight was provided, the individual level weight is equal to the household level weight for all household members. In addition, if the data producer provides replicate weights, those are accessible in the additional *LWS* R-file (see *Structure of LWS datasets* above). Users are strongly encouraged to use weights in their analysis to correct for bias due to the sampling and data collection processes; this is particularly important for *LWS* datasets, as in many instances the wealth surveys oversample the relatively wealthy.

Useful information on *LWS* household balance sheet

Aggregation of wealth variables – The guiding principle for the *LWS* household balance sheet variables is that each original variable is recorded in the asset/liability variable where it

fits best. *LWS* wealth variables are constructed over several aggregation stages: total assets is the sum of non-financial assets, financial assets, and pension assets and other long-term savings. These three sub-components, on their turn, are aggregates from further sub-components. For instance, non-financial assets are the sum of real estate and non-housing assets, whereas the real estate is a sum of principal residence and other real estate. The upper level variables do not always constitute the sum of the lower level variables due to the lack of sufficient detail for a specific asset/liability. For instance, the amount of financial investments might not be always equal to the sum of its subcomponents such as bonds, stocks and investment funds/alternative investments, because some of the financial investments collected by the data provider could be a mixture of these subcomponents; therefore, those amounts would be coded directly in financial investments.

Pension assets availability and the concept of net worth **-** Some surveys do not yet collect all pension assets. In general, this is mostly the case for social security and/or occupational pension entitlements, where sometimes additional computation techniques are needed. In order to separate out differences in measurement of pension assets and to distinguish these practices, LIS has adopted a breakdown of defined-benefit vs. defined-contribution schemes in addition to the breakdown in individual, occupational, and social security pensions. Some data providers just collect the current value of the pension accounts (defined-contribution schemes), whereas they do not calculate the current value of future cash flows, as required in defined-benefit schemes. As soon as some pension assets are not collected or computed, variables pension assets and long-term savings (*HAS*) as well as total assets (*HA*) contain only the missing values. Thus, for comparability reasons, three measures of net worth are offered in the *LWS* datasets:

1) *disposable net worth (variable DNW)* – this variable has always defined values; it excludes the totality of pension assets and other long-term savings (variable *HAS*);
2) *adjusted net worth (variable ANW)* – this variable has only defined values if life insurance and individual pension assets (variable *HASI*) are available;
3) *total net worth (variable TNW)* – this variable has only defined values when all pension assets and other long-term savings (variable *HAS*) are available.

Breakdown of liabilities variables **-** Liabilities are mainly disaggregated following the purpose for which the obligation was made (e.g. real estate, investments, consumer goods, education). An additional breakdown of liabilities by security is also available. Three notes of caution should be highlighted for users of this additional set of liabilities:

- these are additional variables, which overlap with the other liabilities variables, i.e. most amounts included in those variables were also included at some level of disaggregation in the main set of liabilities variables;
- in some cases, the secured loans include business debts while they are not present in the liabilities by purpose due to the fact that all business debts are deducted from business assets and reported on the asset side as business equity;
- the liabilities included in this additional set are NOT always exhaustive; this might happen when all liabilities cannot be categorised as secured or non-secured loans.

Useful information on flow variables

Annualisation - All *LWS* flow variables report annual amounts. If the original survey does not provide annual amounts, whether because of a different reference period, or because the amounts are collected as usual amount together with periodicity and number of periodicities (e.g. usual monthly wage, and number of months during which it was received during the year), the amounts are annualised.

Gross versus net incomes - The income variables in the *LWS* datasets are classified into either gross, net or mixed depending on the extent to which income taxes and social security contributions are captured in the original data (please see variable *GROSSNET* for more information).

Aggregation of flow variables - In general, household income amounts are constructed as the sum of individual-level amounts of all household members. In other words, the sum of any *LWS* individual income variable in the P-file over all household members is identical to the amount in the corresponding *LWS* household level variable. In some cases, individual level incomes do not add up to household level incomes. This is due to the fact that some incomes were collected at the household level. Also please note that in some cases, the individual and household level income variables are constructed independently from each other (e.g. individual level information was collected, whereas household information was retrieved using tax records).

Similarly to assets and liabilities variables (see above), also for flow variables it is often the case that the sub-components do not sum up exactly to their higher level aggregate. This is because some amounts are not allocated into the lower-level variables, but entered directly into the higher level aggregate.

Correction for inflation - In datasets where a country experienced 10% or higher annual inflation during the period of data collection, it is impossible to compare nominal currency values across households at different data collection times. Thus, the amounts are corrected for inflation; all flow variables are deflated (or inflated) to mid-year equivalent values using the Consumer Price Index (CPI) available from official sources. The correction factor used for each observation (which depends on both the CPI index and the time of collection) is reported in the variable *DEFLATOR*. In order to recalculate the nominal values as reported by the data provider, the multiplication of flow variables by the value stored in variable *DEFLATOR* should be carried out.