THE LIS DATABASE: GUIDELINES

Table of contents

I. Introduction

This document records the rules, practices, and definitions applied during the harmonization process to ensure consistency over the LIS datasets. It gives information on general categories of LIS/LWS variables and provides references to documents containing detailed procedures for creating each variable. The information provided here refers to the 2011 Template.

The first part of this document serves to explain the general file structure and rules and practices of the *LIS Database*.

Further, the general practices used to identify and define the variables in the five general categories of LIS variables are described. These five categories are:

1. File identifiers and data information (including weights)
2. Household characteristics variables
3. Socio-demographic variables
4. Labour market variables
5. Flow variables (incomes, consumption and other flows)

Specific, detailed information about the contents and coding of the LIS variables is contained in the *LIS Variables definitions*, but references are in this document.

This document, together with the *LIS Variables definition*s one, constitutes the generic documentation valid for all LIS datasets, and thus indicates the ideal contents of the LIS data. Specific LIS datasets may have contents that differ from these ideal definitions. In all such cases the *LIS codebook* of the dataset in question contains a note warning the user about the deviation from the ideal situation. It is thus very important that when working with LIS data, users consult both this generic information and the dataset-specific documentation included in the *LIS codebook* of the dataset used.

II. Overall Structure

a) Naming of the datasets

The datasets in the LIS database are named according to the 2-character country abbreviation coded according to the ISO-3166 (see Appendix) and the "income reference year".

The income reference year is ideally the calendar year for which income data has been collected. If the income reference period crosses over two years, the year with the longest period will be chosen; if the reference period is equally split between two years, the most recent year will be used. In case the income reference period is shorter than one year, the year in which that period is contained is chosen as the income reference year (and all the incomes are transformed into annual amount if not already expressed as such). Please note that the income reference year may differ from the year following which the survey was named by the data provider, and/or the year in which the survey was conducted.

b) Structure of the datasets

Each LIS dataset is composed of two files, a household level file and an individual level file.

The household level file (henceforth referred to as the LIS H-file) contains a record for each survey unit of the sample; the survey unit is typically the household (whereby the most common definition of household used by most data providers is the single person or the group of persons living in one dwelling and sharing a budget), but may slightly differ in a few cases (e.g. tax unit, family unit, etc. – see variable SVYUNIT for more information).

The individual level file (henceforth referred to as the LIS P-file) always includes as many records as there are individuals in the survey unit (even if the data were originally not reported individually for all persons).

In all datasets, the individual level files contain the same variables, and the same is true across household files. This means that all variables are physically present in the file even if the information was not available for that country and year.

III. Generic Rules and Practices

a) Variable standardisation

The harmonization component of the LIS datasets is reflected both in the file structure (as explained above) and in the contents of each variable. Most LIS variables are standardised in two senses: in terms of conceptual content (the variables are as comparable as possible across datasets in terms of concepts/definitions) and in terms of coding structure, i.e.:

i. Continuous standardized variables report information expressed in the same unit across different datasets (e.g. flow variables report annual units of national currency, hours variables report number of hours worked per week, age variables report number of years).
ii. Categorical standardized variables report information expressed with the same value codes and labels.

There are some variables that are not standardized (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: those variables are always categorical and their exact content and coding structure is indicated in the (dataset-specific) variable and value labels.

b) Generic missing values policy

In LIS data the system missing value (dot) is used to code observations for which the information is not available.

The information may be not available for the following reasons:

- the information is not applicable (e.g. the person does not work and hence cannot have an industry).
- the information is applicable but has not been collected by the data provider for a given subset of the sample or the totality of it.
- the information is applicable and has been collected by the data provider, but the respondent did not answer (don't know/refusal).
- the information has been collected, but is not available at the level of detail necessary for the LIS variable in question (for most continuous variables, i.e. weeks, hours and flow variables).

Please note that a LIS variable may be coded with a valid value (and not the dot) even for observations for which the data provider did not explicitly collect the information (e.g. a person who is not unemployed will have a valid zero value whether the data provider asked him about amount of unemployment benefits received or not). This will ensure that the LIS data available to the users are independent of the structure of the data collection (which may vary considerably across datasets).

c) Sample selection and household membership

The sample included in the LIS files represents a cross-section of the total population during a single time period.

Household sample: the final sample included in the LIS H-file includes only those household-level observations belonging to the valid cross-section of the population. As a result, every observation in the sample will always have a valid cross-sectional weight (as the weight must have been constructed for that specific sample).

Individual sample: the LIS P-file includes observations for all individuals belonging to the households included in the LIS H-file. An individual belongs to the household if he/she is eligible for the "main" individual interview and/or is supposed to be counted for the total weighted sample to be representative of the total population.[1] In most cases, all such persons are defined as ***household members*** (i.e. individuals sharing the budget in the household - see variable HHMEM), their incomes and expenditures are added up to create household aggregates, and they are included in all the household level counters), but there are some instances where they are not (this applies to live-in domestic servants, lodgers, guests and boarders).[2] This is reflected in

[1] If the survey was well designed, this should mean that it is not possible to sample that same individual as belonging to another household.

[2] Non-household members are excluded from the equivalence scale for LIS key figures.

the difference between the variables NPERS (which records the number of individuals in the LIS P-file and for which the weight has been created) and NHHMEM (which is the number of household members).

Residency in the dwelling and household membership - In most cases, in order to be defined as belonging to the household, a person must be resident in the household at the time of interview, or during a specified minimum length of period during the income reference period. However, many data providers accept that some non-resident persons may be considered as part of the household; this is typically the case for students temporarily away for study reasons, adults temporarily away for work or other reasons (e.g. hospitalization, imprisonment). Such persons, including absent heads or spouses, if available in the original data, are kept in the LIS P-file even if not interviewed only in case they are part of the individual sample for which weights have been constructed. The indication about their temporary non residence is not kept in the LIS file.

d) Weighting

LIS does not calculate weights but uses the weights provided by the data provider, both at the household and individual level. Both the household and the individual level weights are always present: if either of the two is not provided LIS creates it from the other one; more specifically, in case only an individual level weight was provided, the household level weight is created from the individual level one by averaging the weights of the individuals in the household, while in the opposite case, when only the household level one was provided, the individual level weight is set equal to the household level weight for all household members.

There are a few general requirements that the main weights used in the LIS datasets must comply with:

- The weight must make the sample representative of the total national population (or the total covered population, which often excludes a small percentage of the total national population, such as collective or institutionalized households, as well as some other specific groups). See Section c) above on the consistency between the sample and the weight; please note that in a few instances the only weight available simply corrects for sampling bias, but not for unit or item non-response, in which case perfect representativeness cannot be guaranteed.
- The weight must be *cross-sectional*; in case of panel surveys, the longitudinal weight is of no interest for LIS data.
- If there is more than one cross-sectional weight, LIS chooses the weight - and the corresponding sample - that *focuses on income* (e.g. the weight to be used in connection with the sample of households that answered the income section of the questionnaire).

e) "Gross", "net" and "mixed" income datasets

LIS datasets are classified into either gross, net or mixed income datasets depending on the extent to which income taxes and social security contributions are captured in the original data.

Gross income datasets - In the ideal case, all LIS income variables (with the exception of the disposable household income variables, DHI and DPI) report amounts gross of income taxes and social security employee contributions (but not employer contributions), and the overall amount of taxes and contributions is also available separately, so that it can be deducted from the total gross amount in order to obtain total disposable income. All the datasets satisfying these criteria are referred to as "gross" datasets (see variable GROSSNET – category 100 "taxes and contributions fully captured"). Please note that often taxes and contributions are only available at the household level even for datasets that have individual level incomes.

Net income datasets - Very often, however, respondents are asked to report only net amounts (as those are the ones they know best), and there is no information about the income taxes and social security contributions paid on those amounts. In these cases, all LIS income variables report net amounts, including the overall total income variable, which will thus be the same as disposable income. These datasets are referred to as net datasets (see variable GROSSNET – category 200 "taxes and contributions not captured").

Mixed income datasets - There are some datasets that are neither purely net nor gross. This can happen in cases when information was available only for taxes but not for contributions (or vice versa), or in cases when the information on taxes and contributions was only available for certain subcomponents of total income, or was available for total income but not subdivided by income subcomponents. Depending on the specific situation, the LIS income variables may either be all net, or all gross, or gross of only taxes or contributions, or partly gross and partly net (a note in the *LIS codebook* will inform the user about the specific situation). All those cases are flagged as being "mixed" income datasets (see variable GROSSNET – category 300 "taxes and contributions insufficiently captured").

Please note that the term gross and net as explained above should not be confused with the situations where the same terms refer to incomes that are gross or net of costs (e.g. rental income and self-employment income can be recorded either before or after deduction of the costs incurred); with respect to this definition of gross/net, LIS variables should ideally be net (and a note would warn users when this is not the case).

f) Monetary versus non-monetary flows

A flow is classified as ***monetary*** if it involves a cash or cash equivalent transaction between the household and a second party. In this instance, the terms monetary and cash can be used interchangeably.

In the case of incomes, monetary refers to an income received directly in cash or cash equivalent. Please note that if a cash or cash equivalent income is tied to a certain good or service (by conditioning its occurrence to the consumption of that good or service) it is still considered monetary (e.g. food stamps, receipts conditional on the payment of certain costs, or on the acquisition of certain goods or services).

In the case of consumption, this refers to a consumption stemming from a monetary transaction (a good or service that has been paid for by the household).

By default, all assets and liabilities transactions are monetary.

A flow is classified as ***non-monetary*** if it concerns the movement of goods or services themselves, without an associated cash or cash equivalent transaction. In this instance, the terms non-monetary, and non-cash can be used interchangeably.

In the case of incomes, this refers to an income received in goods and services (often referred to as in-kind incomes).

In the case of consumption, it refers to a good or service that has been consumed without having being paid for by the household, but either given to it by someone else, or self-produced.

In the case of non-consumption expenditures, a non-monetary expenditure may occur if a third party pays employee contributions (whether mandatory or voluntary) on behalf of the household (the monetary transaction has in fact occurred between the insurance fund and the party who paid the contribution on behalf of the household).

All non-monetary inflows into a household have a counterpart among the non-monetary outflows (any good or service received is considered both as a non-monetary income and a non-monetary consumption; similarly employee contributions paid by a third party on behalf of the household can be seen as both a non-monetary income and a non-monetary non-consumption expenditure).

All LIS non-monetary variables are monetized, i.e. they report the money value of the goods and services being transferred.

g) Annualisation

All LIS 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), LIS annualizes the amounts. In the latter case (where the reference period is one year, but the data are not recorded as annual amounts, LIS simply multiplies the regular amount by the number of periodicities in the year (e.g. if wage income is recorded on a monthly basis, LIS multiplies the amount by the number of months the income was received during the year). In the case of current income surveys (where the incomes refer to the last payment received), or surveys with shorter reference period than a year for the incomes/expenditures (this is often the case for household budget surveys, which may collect for example all the inflows and outflows during a given month), the annualisation carried out by LIS involves the assumption that those flows were occurring during the whole year with the same pattern as during the reference period (and reported amounts are simply multiplied by 52 (if weekly), 12 (if monthly), 4 (if quarterly) or any other number of reference periods in one year).

Lump-sum incomes are taken into account as such, not multiplied by 12 or any other multiplier, unless there is a clear reason to do so (e.g. the original survey indicates that the lump-sum income is received twice per year).

h) Correction for inflation

In datasets where a country experienced serious inflation during the period of the data collection (defined as a 10 percent annual rate or higher), it is impossible to compare nominal currency values across households at different data collection times. LIS therefore corrects the amounts for inflation: all flow variables are deflated (or inflated) to mid-year equivalent values using the Consumer Price Index (CPI) indices available from official sources The correction factor used for each observation (which depends on both the CPI index and the time of collection) is then reported in the DEFLATOR variable, so that if users wish to recalculate the nominal values as reported by the data provider, they may do so by simply multiplying the monetary variables by the DEFLATOR value.

i) Aggregation rules

Aggregation of sub-categories into overall categories in categorical LIS variables – Most categorical variables have a multi-digit coding structure whereby the codes starting with the same digit belong to the same overall category (e.g. the codes in the 100s for variable CLFS – current labour force status - stand for ILO employed persons, while the codes 110 and 120 are subcategories of that overall category, namely ILO employed at work, and ILO employed not at work). Please note that in one and the same dataset observations can be coded either at the higher level or at the lower level (i.e. some persons may have been coded 110 (employed at work), others 120 (employed not at work), yet others directly with 100 (ILO employed) in case it was not possible to determine whether they were at work or not). As a result, by selecting all the subcategories, one does not necessarily get the total of one overall category (i.e. by selecting the employed at work (110) and the employed not at work (120), one does not necessarily select all the ILO employed; the selection of the higher code 100 is needed as well).

Aggregation of individual level incomes into household level incomes - In general, household income amounts are aggregated from individual-level values of all household members, so that the sum of any LIS individual income variable in the P-file over all household members is identical to the amount in the corresponding LIS household level variable. In some cases though, individual level incomes do not sum up to household income; this may be because only part of those incomes were collected at the individual level, while the rest was only available at the household level (e.g. children’s wages are often recorded only at the household level). If individual-level information is not available, all the individual level income variables remain at missing and the recorded household values are used to fill in the household level LIS variables.

In some cases, the individual level income variables and the household level ones are filled independently from each other (this happens if the data provider provides us with two sets of variables, either because they were collected independently, or because only one of the two sets is imputed/grossed up).

Aggregation of LIS sub-variables into upper level variables – Whereas conceptually LIS flow variables are constructed over several aggregation stages (whereby for example total current income is the sum of income from labour, capital and transfers, and, on their turn, each of the three sub-components is aggregated from further sub-components), the situation in which all sub-components sum up exactly to their higher level aggregate only arises when the original data were provided with exactly the same structure as the LIS variables.

The guiding principle is that each original variable is recorded in the LIS variable where it fits best, which implies that, depending on the aggregation level of the original variables, some will be reported at the lowest detailed LIS variables, and others directly in higher level aggregates, so that the higher level aggregate may contain some amounts which appear only there, and others which come from lower level variables. As a result, when summing up the sub-components of a LIS aggregate, one may not necessarily find the same amount reported in the aggregate.

For flow variables entering the construction of disposable household income (i.e. all the current income variables and the income taxes and social security contributions variables), the direct filling of an upper level variable can lead to situations where a detailed variable is filled, but not its immediate higher level one. Take as an example the case where occupational pensions are reported separately (and hence put in HITSILO), but employment-related public pensions are lumped together with universal and/or assistance pensions, so that this latter lumped amount ends up directly in HITS (as it is not possible to separate out insurance from universal and assistance transfers); in such a case, the variable HITSIL is also not filled, as it could only be filled with a part of its total (only occupational pensions). At the highest level (total current income), the aggregated variables will always be filled (as by definition all LIS datasets include a measure of total household income). Note that individual level aggregated variables though can only be completely filled in case all the corresponding subcomponents were available at the individual level (i.e. the fact that variable PITSI is filled indicates that there was no mix of insurance transfers with either universal or assistance transfers at the individual level, so that that variable could be easily created, but it does not ensure that that variable include all insurance transfers, as those that were only available at the household level would obviously not be included).

For other blocks of flow variables (i.e. those not entering the construction of total disposable household income), the upper level variables are always constructed, independently of whether all subcomponents are available, so that the existence of an upper level variable does not ensure that it is complete (but it is rather the sum of all the subcomponents – or the amounts that were directly put at that level - that were available). The only exception concerns total consumption (HC): that variable is blanked out in case the original data only provided for a subset of the total consumption.

j) Data editing and imputation

When constructing the LIS variables, LIS does some data cleaning and editing of the original data. This may happen at two levels:

Consistency checks – Some very broad consistency checks are applied to some sets of variables. This is mostly the case for the living arrangements and major demographic variables: it is always ensured that each household has exactly one (and only one) head and that there is not more than one spouse (unless it can be explained by polygamy); age and marital status are checked against relationship to the head and to each other. Some other variables are also checked against impossible or highly improbable values (e.g. a group of oddly looking numbers at the high end of a continuous variable, such as a series of 8s or 9s).

Data editing – For standardized variables, original data may be substantially re-edited. This is especially true for summary dichotomous variables such as IMMIGRANT, DISABLED, EDUC, etc. For those variables, users who do not wish to use the LIS codings (based on assumptions made at the discretion of the LIS staff, and reported in the dataset specific documentation) are welcome to use the corresponding country-specific variables (e.g. instead of using the DISABLED dummy, users can choose to work with the corresponding country-specific variables DISABLED_C, ILLNESS_C and/or HEALTH_C and create their own definition of disabled person).

Imputation – While LIS always uses the results of imputations carried out by the data providers, it never carries out its own imputations. As a result, depending on the data providers' imputation practices, some LIS datasets include item or partial unit non response while others do not. In case of major income imputations, a dummy flags the records imputed (see FPIMPU for the individual level and FHIMPU for the household level).

IV. Variable Groups

1. Technical variables

a) Identifier variables

Each file contains unique identifiers of its observations (households or individuals within households), as well as unique identifiers at the level of the dataset (dataset, country, wave, year); each variable in this last set of identifiers has a constant value within the same dataset.

b) File information variables

i. Weights: The H/PWGT and H/PPOPWGT weights are always filled and report the main household/individual level weights (see Section II.d on *Weighting* above); in addition, variables HWGTA and PWGTA report additional household/individual level weight in case only part of the household sample has been selected for some variables and the data provider calculated a special weight for that subsample.

ii. Information about survey unit (ideally household), number of persons in the unit and household membership.

2. Household characteristics variables

These variables are all household level variables and report the major characteristics of the households. They are split into four blocks:

a) Geographical characteristics

These variables contain country-specific information about the locality in which the household resides, the geographical/administrative region, as well as indications about the population density or type of area. On the basis of these country-specific variables, LIS has created a dichotomous indicator for rural areas (RURAL).

b) Dwelling characteristics

These variables contain information about the dwelling (or principal residence) of the household: the tenure (owned versus rented), the type (house versus apartment or other), the value, and an indicator for farm households.

c) Household farm

Variables containing information about households carrying out farming activity (whether owned or not, whether with crops or livestock).

d) Household composition

These variables contain information about the composition of the household. In most cases, household composition variables are derived from individual level demographic and living arrangement variables (this is not true for those few surveys where individual level information was not available for all household members). The counters are all based on the household member definition: they all report numbers of household members and not necessarily the number of persons in the survey unit (see Section on *Sample selection and household membership* for the definition of survey unit versus household).

3. Socio-demographic variables

These variables are all individual level variables and report the major socio-demographic characteristics of the household members. They are split into 6 major blocks.

a) Living arrangements

These variables contain information about the living arrangements of the household members: the relationship of each household member to the head of the household, as well as indications on whether the household members have a partner, or whether they live in the same household as their partner, parents or children. Please note that the ***household head*** is usually chosen by the data provider, and its definition may differ across datasets (main income earner, person most knowledgeable about the budgetary situation of the household, eldest person, person responsible for the dwelling contract, or simply self-defined by the respondents, etc.). Whereas the relationship to the household head is available in all LIS datasets for all household members, the presence of a partner, parents or children is often only available for household head (and possibly his/her partner).

b) Demographics

These refer to the age, gender and marital status (both *de facto* and *de jure*) of the household members. *De facto* marital status (reported both in a detailed categorical variable and a summary dichotomous one) refers to the married or cohabiting situation, and is by definition independent from (even though most often coinciding with) the living arrangements (as reported in the variable PARTNER). Age and gender are always available for the totality of the sample (with the exception of some earlier datasets for which individual characteristics were only collected for head and spouse, or up to a certain number of adults). Marital status is often available for adults only.

c) Immigration

These variables are intended to identify the immigrant minorities resident in a given country. They include information on citizenship, country of birth, duration of stay in the country (since arrival), ethnicity/race, previous place of residence and other relevant country-specific immigration characteristics (such as mother tongue, religion, etc.), and are summarized in a summary dichotomous variable (IMMIGR), created by LIS on the basis of the available information (see the definition of the IMMIGR variable for a precise indication of the rules used for its construction).

d) Health

Any available information about disability status, chronic illness and subjective health status has been put in three country-specific variables, on the basis of which LIS created a dichotomous summary variable (DISABLED) to identify persons with disabilities/chronic diseases/poor health.

e) Education

Education variables include information about highest education level, age at which left education, and an indication of whether the person is currently in education, and if so, the level thereof.

Highest education level may refer to the highest completed level or the highest attended one and is provided in country-specific classification. On the basis of that information, LIS has created a 3-category classification of highest completed education into low (less than secondary), medium (secondary) and high (tertiary) levels (EDUC), which is based on the ISCED classification (see Appendix for details about ISCED).

Country-specific information about the education level (possibly highest completed) of the mother and the father of the household members is also provided.

4. Labour market variables

These variables are all individual level variables and report the major labour market characteristics of the household members. They are split into 4 major blocks.

a) Activity status

Activity status can be identified following three different definitions:

i. Current labour force status (CLFS), which classifies as employed those who, during a specified brief period (the current period) carried out ANY employment (any type or any extent), even if it is just one occasional hour of paid work or irregular unpaid family work. The current period is defined by the data provider and is either the interview day/week, or any specific other day/week. The ideal definition of employment in this variable should follow the ILO concept of "currently employed".

ii. Current main activity status (CMAS), which classifies as employed those for whom work is the main activity during the current period. For all others, the main activity should attempt to distinguish between pensioners, students, and homemakers; the concept of main activity is generally defined in the original data (typically it will be self-assessed by the respondent, following some instructions such as "the activity at which you spend the most time", but it may be asked without any precise guidelines).

iii. Usual activity status in income reference period (UMAS), which classifies as employed those for whom employment was the main activity over a specified long reference period (typically the same as the income reference period, i.e. a period of 12 months), as determined by the number of weeks/days in employment. The usual main activity is sometimes provided directly in the data (this may be from a question asked to the individuals or it may be constructed by the data provider using the reported activities over the reference period), but in many cases, it is not present in the data, in which case UMAS will be constructed by LIS using a calendar of activities if available; in the latter case, LIS determines employment to be the main activity if it is carried out during at least half of the reference period. The ideal definition of employment in this variable should follow the ILO concept of "usually employed".

One, two or all three variables can be filled in one and the same dataset. In order to simplify the use of such variables, LIS has used the three variables above to derive dichotomous variables which flag employment (EMP and MAINEMP), unemployment (UNEMP), retirement (RETIRED) and military status (MILIT). Each summary variable has an ideal derivation rule: EMP, UNEMP and MILIT should be derived from the current labour force status, RETIRED from the current main activity status, and MAINEMP from the usual activity status. In case it is not possible to construct those variables according to these ideal derivation rules, different priority rules can be used. More precisely EMP, UNEMP and MILIT can be derived from CMAS or UMAS if CLFS is not available (in some rare instances, it may be constructed directly, see below the note on datasets with annual labour market information), while RETIRED can be derived from UMAS or CLFS if CMAS is not available. On the other hand, MAINEMP is only created out of UMAS. In case a variable was not created using the ideal definition, a clear warning would report the deviation.

In addition to these overall activity statuses, there are some other variables concerning the employment-related activities of individuals (information on job-search, unemployment duration, employment situation desired, caregiving responsibilities, leave from work); of particular importance for the lower and middle income countries, three country-specific variables try to capture information about the involvement in the informal sector: information on household production, on odd/marginal/irregular jobs, and on unofficial work affiliation.

Please note that all activity status variables are typically defined only for those persons who answered the labour market individual questionnaire (adult persons for most datasets).

b) Employment intensity

Several measures of overall employment intensity are reported in the second block: indicators of multiple employment, part-time employment, and the number of hours worked per week and number of weeks worked in a year. On the basis of the number of weeks and hours worked, LIS then creates an overall full-year full-time dichotomous variable (FYFT) and an overall part-time employment indicator (PTIME).

Please note that different employment intensity variables are defined for different groups of the population: variables about hours worked, part-time and multiple job indicators (the second job dummy and the number of jobs) refer to the present situation, and are hence defined only for persons who currently work (as from the EMP dummy variable); on the other hand, variables about weeks worked during the year and full-year full-time indicator, refer to the situation over a longer reference period (usually the income year), and are hence defined for all persons who were eligible to answer the labour market individual questionnaire (usually all adults – see EMP), whether they are currently working or not. As a result, for a person who is not currently employed weekly hours will not be reported (the variable will be set to missing), whereas annual weeks will be reported (at either zero, if the person did not work during the whole reference year, or any positive number).

c) Job characteristics

These variables report the main job characteristics for both the main and the second job of every individual. They include the status in employment (dependent versus self-employment), the industry, the occupation, the number of persons employed in the local unit, the sector of employment (private versus public), managerial responsibilities, duration of the employment (permanent /long-term versus short-term contract, actual tenure), the work intensity (hours worked at each job), and an indication of the income level (as indicated by the hourly wage rate).

Please note that most of these variables are fully standardized; this includes the occupation and industry codings, which even if delivered following country classifications, have been recoded to fit the ISCO/ISIC/NACE classifications as closely as possible (see Appendix for a description of those standard classifications).

Job characteristics variables for the main job are defined only for persons who have a job (as from the EMP dummy), while those for the second job are only defined for persons who have more than one job (as from the SECJOB dummy).

d) Work experience

The last block of labour market variables refers to the overall work experience, with indications for the existence of such experience, as well as its duration (years of total work experience as well as a breakdown of full-time versus part-time experience), or potential duration (given by the age when the individual started working).

These variables are typically defined for the same group of persons for which the labour status variables are defined (mostly adults).

A note on datasets with annual labour market information

In case a dataset collects information on the annual labour force status, without any indication about the current period (e.g. number of days/weeks/months in employment, number of jobs held during the year, and characteristics of the main job held during the year), the information will be adapted to the LIS variables in the following way: CLFS and MAINEMP will be empty (as there is no information about the current period), while UMAS will report the usual activity status. The variable EMP will be created directly from the original data and will be set to 1 for persons who have had any job during the year (even if not usually employed over the year). The information about multiple jobs (NJOBS and SECJOB) will refer to the jobs held at any time during the year (rather than the simultaneous jobs held at present), and the job characteristics will all refer to the main job held during the year (which is not necessarily the one currently held).

5. Flow variables

LIS flow variables refer to the monetary and non-monetary in- and out-flows of the household, including income (current and windfall), consumption, non-consumption expenditures, and other in- and outflows resulting from transactions that do not reduce or increase the household net worth.

Data collection for such variables may differ substantially between original surveys: some data are based on survey collection, others on administrative records, yet others on full simulations (e.g. taxes, imputed rents or even some flat-rate public benefits). Some datasets use one of the three sources above, while others may combine two or all three methods.

Data availability also varies widely. A dataset included in the LIS database must by definition have full coverage of current income. However, income taxes and social security employee contributions and other non-consumption expenditures are sometimes not included, while consumption is often not included (or not fully covered) and windfall incomes and assets/liability transactions are rarely available.

All LIS 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).

LIS flow variables are split into the five major blocks:

- *Current incomes* (*I* variables): these consist of monetary payments as well as (the value of) goods and services received by the household or by individual members of the household at periodic intervals (annual or smaller), that are available for current consumption and that do not reduce the net worth of the household.
- *Windfall incomes* (*W* variables): these consist of windfall gains and other such irregular and typically one-time receipts.
- *Non-consumption expenditures* (*X* variables): these consist of monetary expenditures (i.e. paid directly by the household and/or its members) and non-monetary expenditures (paid on behalf of the household and/or its members) on non-consumption goods and services (such as taxes, contributions, donations, inter-household transfers and interest paid).
- *Consumption* (*C* variables): these consist of monetary and non-monetary consumption items.
- *Assets / liabilities transactions* (*T* variables): monetary inflows that do not constitute income (neither current nor windfall) and outflows that do not represent consumption, and do not reduce or increase the net worth of the household, but rather change the composition between cash, financial and non-financial assets and liabilities.

All these variables always exist at the household level in the household file (prefixed by "H"), and may or may not exist at the individual level in the person file (prefixed by "P"), depending on the relevance of the variable at the individual level (see *Variables List* for an exhaustive list of all the variables that exist also in the person file). Please note that, differently from the household level variables, the individual level ones are constructed as soon as some categories are reported, even if they are not complete. This is especially true for total individual level income, which may be highly underestimated in some datasets.

In addition, for each flow within each of the five blocks in either the household or person file, there can be up to three different variables:

- a *monetary variable* (denoted by an "M" in the second position of the variable name) exists for most flows;
- a *non-monetary variable* (denoted by an "N" in the second position of the variable name) exists only for items that can exist in non-monetary form (see Section II.f for a definition of monetary flows);
- an overall *monetary and non-monetary variable* (denoted by the absence of the "M" or "N" letter in the second position of the variable name) always exists, and is either equal to the sum of monetary plus non-monetary flows, or to either of the two components if only one exists.

All variables exist in total versions (monetary and non-monetary); for an exhaustive list of which variables exist in either monetary or non-monetary versions, see the *Variables List*.

On top of these blocks, there is an overall block of main household aggregates constructed from variables belonging to blocks "Current income" (I variables) and "Non-consumption expenditures" (X variables) and that report the major overall concept of total disposable income and its major components (see Section on *Major income aggregates* below). Please note that the amounts reported in these aggregate variables are always also included in the relevant variables of the main blocks (I or X variables).

Finally, there is a section containing five additional pieces of information concerning the flow variables:

i. the local currency in which all flow variables are expressed in the dataset (variable CURRENCY).
ii. the information necessary to revert to nominal values in case LIS has applied a within-the-year inflation correction to the monetary variables (see section below on *Inflation correction*).
iii. information distinguishing between net and gross datasets (see section III.e on gross, net and mixed income datasets).
iv. flags for observations where the income items were fully imputed (FPIMPU and FHIMPU).
v. variables to report the receipt or payments of certain incomes/expenditures, which are often only available as flags rather than as amounts (this is often the case for in-kind incomes, as well as for the payment of voluntary contributions).

a) Current income

Current income is defined as all receipts whether monetary or not monetary (goods and services) that are received by the household or its individual members at annual or more frequent intervals, that are available for current consumption and that do not reduce the net worth of the household. These include monetary and non-monetary income from labour, monetary income from capital, monetary social security transfers (including work-related insurance transfers, universal transfers, and assistance transfers), and non-monetary social assistance transfers, as well as monetary and non-monetary private transfers, less the amount of income taxes and social contributions paid. There are two notable exceptions among the non-monetary incomes:

- *Non-monetary incomes from capital -* These refer to the imputed value of the service of durable goods owned by the household, including the dwelling and other durables such as cars. As important as these incomes may be, they are rarely available in the income microdata and, when available, they are calculated with widely varying methodologies. For these reasons, they are excluded from DHI. Users wishing to include them can do so with the use of the LIS microdata.

- *Non-monetary universal transfers from government -* These refer to government-provided services that benefit individuals, but are provided with the primary objective of meeting the general needs of the overall population, rather than that of assisting the poor. Specifically, we do not include non-monetary transfers in the areas of housing, care (including child care), education, or health. These transfers are very hard to evaluate at the individual level and thus are typically only available at the macro-level. Thus, the value of these transfers is also excluded from DHI and, these non-monetary incomes are not available in the LIS microdata.

Although we state above that we include non-monetary social assistance transfers, note that this does not mean that all non-monetary means-tested public benefits are included in current income. We exclude means-tested public benefits in cases where they form a portion of a system in which benefits are granted to the whole population (poor and non-poor), although using different tools and programs. For example, in the case of health insurance in the U.S., we have excluded benefits received through the Medicaid program (which provides health insurance to low-income Americans) because most persons who do not receive Medicaid are subsidized either through the U.S. tax system – if employed – or through Medicare (the social insurance program for the elderly and persons with disabilities).

Note that by referring to the periodic intervals of one year or less, the definition of current income implicitly excludes all irregular and one time receipts such as windfall income and non-income inflows (see below for definitions).

Total gross current income is disaggregated into three major components:

Current income from labour ("IL" variables): monetary and non-monetary payments received in compensation for labour;

Current income from capital ("IC" variables): monetary payments received as a return on capital assets (including financial and non-financial assets);

Income from current transfers ("IT" variables): monetary and (some) non-monetary payments received without a counterpart; current transfers are in turn further split into the following major categories:

- ***Social security transfers*** ("ITS" variables): transfers that result from an institutional arrangement between the recipient and the government and/or the employer, with the explicit intention to relieve households and individuals of the burden of a defined set of risks or needs (please note that the role of the government may be limited to the compulsiveness of such arrangements, so that mandatory individual programmes are also considered as part of social security); these transfers may be paid either directly from public bodies, or through private bodies representing the institutional arrangement as defined above. They can be further split into:
    - ***Work-related insurance transfers*** ("ITSI" variables): transfers stemming from systems where eligibility is based on the existence and/or the length of an employment relationship; in most cases the benefits are financed by contributions paid by employers, workers or both, and their amount is usually dependent on either previous earnings or previous contributions;
    - ***Universal transfers*** ("ITSU" variables): transfers stemming from public programmes that provide flat-rate benefits to certain residents or citizens, provided that they are in a certain situation, but without consideration of income, employment or assets; note that in some cases the benefit amount may also depend on the other incomes of the individuals, which at the limit may result in some proportion of the population at the upper end of the income distribution being excluded from receipt;
    - ***Assistance transfers*** ("ITSA" variables): transfers stemming from public programmes that provide benefits especially targeted to individuals or households in need (i.e. with a strict income or asset test); the amount of the benefit is either a flat rate or is based on the difference between the recipient income and a standard amount representing the minimum subsistence needs as guaranteed by the government.

- ***Private transfers*** ("ITP" variables): these are all transfers of a purely private nature, that do not involve any institutional arrangement between the individual and the government or the employer as defined above; they include transfers from non-profit institutions, other private households or other bodies in the case of merit-based education transfers.

Additional sets of income transfers ("IAT" variables): Current transfer incomes are often disaggregated by the function they serve rather than the type of transfer itself (transfer made to assist the needs of the elderly, the disabled, the sick, etc.). In order to *i)* accommodate datasets where that is the only disaggregation provided, and *ii)* facilitate the users who may wish to use this second possible disaggregation of transfers, LIS has introduced an additional set of transfer variables that disaggregate the transfers according to the following functions: old-age, disability, survivors, sickness, family/children, education, unemployment and housing. Finally, four common transfer categories that cut across both the type dimension and the function dimension have also been added (civil servants pensions, work-injury compensation, care benefit, and war victim/veterans benefits). Two notes of caution should be highlighted for users of these additional sets of income transfers:

- these are additional variables which overlap with the other transfer variables, i.e. the amounts included in those variables were also included at some level of disaggregation in the main set of transfer income variables ("IT" variables);
- the incomes included in those additional sets are NOT meant to be exhaustive: only those transfers that can clearly be identified as being the (almost) totality of a specific additional variable are included in that variable.

As a result, these variables should be used separately from the ones from the main set (to avoid the risk of double-counting incomes) and should never be used as subcomponents aggregating up to total transfer income.

b) Windfall income

Windfall income (often referred to as capital income as opposed to current income) consists of all windfall gains and other such irregular and typically onetime receipts.

Similarly to current income, it can be disaggregated into windfall labour income (severance pay, retirement packages), windfall capital income (capital gains, insurance compensations) and windfall transfers income (inheritances, lottery winnings, lump-sum retirement compensation).

c) Non-consumption expenditure

Non-consumption expenditures consist of monetary expenditures (i.e. paid directly by the household and/or its members) and non-monetary expenditures (paid on behalf of the household and/or its members) on non-consumption goods and services (such as taxes, contributions, donations, inter-household transfers and interest paid).

d) Consumption

LIS ideally records total consumption, including that stemming from expenditures (monetary consumption) and that stemming from own-production, transfers or gifts (non-monetary). More precisely, a consumption item is considered monetary if the good or service consumed has been purchased by the household, whereas it is considered as non-monetary if it has not been purchased, but either given to the household from somebody else, or self-produced.

Consumption items (whether monetary or non-monetary) can be disaggregated by the type of good or service being consumed (LIS uses the 12 major groups of the COICOP classification of consumption goods and services – See *Appendix A*).

Non-monetary consumption items only can also be disaggregated by whether the goods and services were self-produced (goods produced for own consumption, owner-occupied imputed rent or use value of durables) or given by somebody else (from employment, the government, or others).

e) Assets / liabilities transactions

Transactions of assets and liabilities consist of monetary inflows that do not constitute income (neither current nor windfall) and outflows that do not represent consumption, and do not reduce or increase the net worth of the household, but rather change the composition between cash, financial and non-financial assets and liabilities.

By definition, transactions of assets and liabilities can only be monetary.

f) Major income aggregates

In order to facilitate the use of the LIS data, LIS has created some household level income aggregates that are commonly used in research. These aggregates are all derived from other LIS variables, and the calculation always follows the same formula to ensure prefect comparability across datasets.

- DHI: disposable household income, which includes total monetary and non-monetary current income net of income taxes and social security contributions (HI-HXIT); DHI is the variable used for the LIS Inequality and Poverty Key Figures.
- DPI: cash disposable household income, which includes total monetary current income net of income taxes and social security contributions (HMI-HMXIT); this variable is defined (as well as named) in exactly the same way as the old LIS DPI variable;
- FACTOR: factor income, which includes total current monetary and non-monetary income from labour and capital (HIL+HIC);

- SOCRED: social security redistribution, which includes total current monetary and non-monetary social security transfers (HITS);
- PRIVRED: private redistribution, which includes total current monetary and non-monetary private transfers (HITP);
- PENSION: total pension income, which includes private and public pensions, whether from insurance, universal or assistance systems (HITSIL+HITSUP+HITSAP+HICVIP); this is the only aggregate that also exists at the individual level in the person file (prefixed by a "P").

The amounts reported in these aggregate variables are always also included in the relevant variables of the blocks "Current income" (I variables) and "Non-consumption expenditures" (X variables). In general, they are included in the variables shown in the formulas above; however, in the cases where the original detail does not allow filling directly the variables reported in the formulas above (but some higher level variable), those formulas are not satisfied. For example, in case insurance, universal and assistance pensions are lumped together in the original data (and hence put directly in HITS), variables HITSIL, HITSUP and HITSAP would be empty, but the aggregate variable PENSION would still be filled.

Appendix A: Acronyms

- ***ISO-3166 (Codes for the representation of names of countries and their subdivisions)***

  A three-part geographic coding standard for coding the names of countries and dependent areas, and the principal subdivisions thereof, published by the International Organization for Standardization (ISO). LIS uses the two-letter country codes of ISO-3166 (ISO 3166-1-alpha-2 code).

- ***NUTS (Nomenclatue of territorial units for statistics)***

  http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

  The NUTS classification is a hierarchical system for dividing up the economic territory of the EU for the purpose of:

  - The collection, development and harmonisation of EU regional statistics.
  - Socio-economic analyses of the regions:
    - NUTS 1: major socio-economic regions
    - NUTS 2: basic regions for the application of regional policies
    - NUTS 3: small regions for specific diagnoses
  - Framing of EU regional policies.

- ***ISCED 1997 (International Standard Classification of Education, 1997)***

  http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm: ISCED 1997 was designed by UNESCO to allow comparisons of educational attainment based on levels and fields of education. Adopted in 1997, it replaces the original ISCED (1978). The following broad classifications are used in ISCED 1997:

  0 Pre-primary education

  1 Primary education; first stage of basic education

  2 Lower secondary education; second stage of basic education

  3 (Upper) secondary education

  4 Post-secondary non-tertiary education

  5 First stage of tertiary education (not leading directly to an advanced research qualification)

  6 Second stage of tertiary education (leading to an advanced research qualification)

  The groups are further broken down by (1) duration of the program; (2) the type of subsequent education or type of labour market positions for which they prepare graduates; and (3) the degree to which the program is specifically oriented towards a specific class of occupations or trades.

- ***ILO (Industrial Labour Organization)***

  http://www.ilo.org/public/english/about/index.htm: The International Labour Organization is the UN specialized agency that seeks the promotion of social justice and internationally-recognized human and labour rights. It was founded in 1919 and is the only surviving major creation of the Treaty of Versailles, which brought the

League of Nations into being. It became the first specialized agency of the UN in 1946.

The ILO formulates international labour standards in the form of Conventions and Recommendations setting minimum standards of basic labour rights: freedom of association, the right to organize, collective bargaining, abolition of forced labour, equality of opportunity and treatment, and other standards regulating conditions across the entire spectrum of work-related issues. It provides technical assistance, primarily in the fields of:

- vocational training and vocational rehabilitation;
- employment policy;
- labour administration;
- labour law and industrial relations;
- working conditions;
- management development;
- cooperatives;
- social security;
- labour statistics and occupational safety and health.

It promotes the development of independent employers' and workers' organizations and provides training and advisory services to those organizations. Within the UN system, the ILO has a unique tripartite structure with workers and employers participating as equal partners with governments in the work of its governing organs.

- ***ICSE-93 (International Classification by Status in Employment)***

  Adopted in 1993 by the 15th International Conference of Labour Statisticians, the ICSE-93 classifies employment status into the following groups:

  - employees, among whom countries may need and be able to distinguish "employees with stable contracts" (including "regular employees");
  - employers;
  - own-account workers;
  - members of producers’ cooperatives;
  - contributing family workers; and
  - workers not classifiable by status.

  ICSE-93 supersedes the original ISCE-58 (1957).

- ***ISCO-08 (International Standard Classification of Occupations)***

  ISCO-08 consists of 10 broad occupation groupings, with detail at the 4-digit level. Adopted in December 2007 through a resolution of a Tripartite Meeting of Experts on Labour Statistics, ISCO-08 supersedes the previous ISCO-88 (1987), ISCO-58 (1957), and ISCO-68 (1966) versions of classifications.

- ***ISIC Rev. 4 (International Standard Industrial Classification of all Economic Activities)***

ISIC Rev. 4 consists of 21 broad industry classifications, with detail at the 4-digit level. Released in August 2008, ISIC Rev. 4 supersedes previous classifications (Rev. 3.1, 2002; Rev. 3, 1994; Rev. 2, 1968; Rev. 1, 1958; and the original ISIC, 1948).

- ***NACE Rev. 2 (Classification of Economic Activities in the European Community)***

  NACE Rev. 2 consists of 21 broad industry classifications, with detail at the 6-digit level, the first four of which are the same in all European countries. NACE Rev. 2 is meant to be used for statistic from 2008 onwards. NACE Rev. 2 is similar in structure to ISIC Rev. 4. Adopted in 2006, NACE Rev. 2 supersedes NACE Rev. 1.1, 2002 (similar to ISIC Rev. 3).