# THE LIS USER GUIDE
# 2019 Template

## Table of contents



**THE LIS USER GUIDE**

**2019 Template**

# Introduction

The *LIS Database* is the largest available income database of harmonised micro-datasets collected from about 50 countries in Europe, North America, Latin America, Africa, Asia, and Australasia spanning over five decades.

Harmonized into a common framework, LIS datasets contain household- and person-level data on labour income, capital income, pensions, public social security benefits (excl. pensions) and private transfers, as well as taxes and contributions and other non-consumption expenditures, consumption, together with socio-demographic and labour market information. The framework used for the harmonisation of the LIS data has evolved over time as a response to changing survey methods, revisions of theoretical foundations of key concepts, and integration of data from new countries. The set of LIS variables, their definitions and their harmonisation rules underlying the current version of the data is referred to as the *2019 Template*.

This document provides generic guidelines for the usage of LIS data, their overall structure and harmonisation practices, as well as an overview of the main blocks of variables included in the *LIS Database*.

More detailed information about the contents and coding of the LIS variables is available in the METadata Information System (METIS) under the *Variables definitions*. In addition, METIS provides dataset-specific information included as *Contents and Notes*; this field includes both warnings about specific situations as well as the contents of income and consumption variables in each LIS dataset. Information about the original data source, such as the data provider, data collection information, and additional technical information is also available in METIS in the *Dataset information* section.

# I. Overall Structure and generic rules and practices

## 1. 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. 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.

## 2. 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 household of the sample (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).

Note that in a handful of cases the data are based on concepts different than the household (e.g. tax unit, family unit, etc. – see *Dataset information* in METIS for more information).

- The individual level file (henceforth referred to as the LIS P-file) includes as many records as there are individuals in the household.

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.

## 3. Variable standardisation

Most LIS variables are standardised along two dimensions: 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 standardised variables report information expressed in the same unit across different datasets (e.g. monetary variables report annual units of national currency, hours variables report number of hours worked per week, age variables report number of years).
ii. Categorical standardised variables report information expressed with the same value codes and labels.

*Country-specific variables* - There are some variables that are not standardised (variables denoted by "_c" suffix). While the variable name is the same across all datasets, the variable label may differ to indicate the actual (dataset-specific) content for the dataset in question. Both the exact content 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 (as can be seen in the *Statistics* or *Codebooks* sections of METIS), while additional information may be provided in the *Notes*.

## 4. Generic missing values policy

In the original data underlying the LIS Database, treatment and documentation of missing values (refused to answer or does not know) and not applicable cases, vary substantially. In order to simplify the information and usability across the datasets, LIS has standardised its approach to missing values in the following way.

*Missing values policy for income and consumption variables* - This section applies to the blocks Current income, Income deductions, transfers paid and loan repayments, Extraordinary income, Imputed rent, Consumption expenditure, and Major economic aggregates. LIS codes to value '0' in the following three scenarios:

i) the information has not been collected.

ii) the information has been collected, but is not available at the level of detail necessary for the LIS variable in question.

Vice versa this means that only observations where respondents refused to answer or indicated that they do not know the amount, are reported with missing value 'dot'.

*Missing values policy for other LIS variables* - 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).

## 5. 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 *household members*, i.e. individuals belonging to the households included in the LIS H-file and sharing the budget in the household. Note that in some surveys, information is collected also for individuals who are linked to the household but do not share the budget (this applies to live-in domestic servants, lodgers, guests and boarders, and in some rare case also to persons not living in the same dwelling, such as relatives); in this cases, LIS does not retain the information, as the incomes of those non-household members are not to be considered when aggregating individuals into household to compute household level aggregates; similarly, they should not be counted in the equivalence scale as they do not consume resources from the same budget as other household members.

## 6. Weighting

All LIS files include weight variables (at the household level in the H-file 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 LIS files are those calculated by each data producer; generally, those correct at least for sampling bias, but not necessarily for unit or item non-response bias, in which case perfect representativeness of the total national population cannot be guaranteed. For more information about the data provider methodology for the construction of the weight, see the *Dataset information* in METIS.

On the basis of the original weight(s), LIS carries out the following adjustments:

- If the data provider provides weights only at one level (either household or individual), LIS creates the weight at the other level from the one provided; 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.

- If the weight originally provided by the data producer does not already inflate to the total population, then an inflation to total national population is carried out by LIS to create the main set of LIS weight variables (*hpopwgt* and *ppopwgt).*
- LIS provides a second set of weights (*hwgt* and *pwgt*), where the original weight is normalised to 10,000 by country; those weights should be used in multi-country analyses if each country is intended to have the same weight (while the use of inflated weights implies that each country counts in the final results proportionately to its population size).

In addition to the main and the normalised weights, LIS also provides additional weight variables (*hwgta* and *pwgta*) in case the data provider had provided a weight calculated calculated for a selected sub-sample of households.

## 7. "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 *dhci*) 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* - 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 dataset note in METIS 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).

## 8. Annualisation

All LIS income and consumption variables report annual amounts. If the original survey did 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 annualises 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).

## 9. Aggregation rules

*Aggregation of sub-categories into overall categories in categorical LIS variables* – Many categorical variables have a multi-digit coding structure whereby the codes starting with the same digit belong to the same overall category. As a result, by selecting all the subcategories, one does not necessarily get the total of one overall category. To clarify this, let's make an example with variable *own* (owned/rented housing): the codes in the 100s stand for owned housing, while the codes 110 and 120 are subcategories of that overall category, namely owned outright and owned with mortgage. In one and the same dataset observations can be coded either at the higher level or at the lower level (i.e. some households may have been coded 110 (owned outright), others 120 (owned with mortgage), yet others directly with 100 (owned) in case it was not possible to determine whether they had a mortgage 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 owned outright (110) and the owned with mortgage (120), one does not necessarily select all the owned; 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:

i) 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 zero and the recorded household values are used to fill in the household level LIS variables.

ii) In other cases, the individual level income variables and the household level ones are filled independently from each other (this happens if the data provider provides two sets of variables,

either because they were collected independently, or because only one of the two sets was imputed/grossed up, e.g. cn13).

*Aggregation of LIS sub-variables into upper level variables* – LIS income variables are disaggregated over several levels of subcomponents: total current income is the sum of income from labour, capital, pensions, public social benefits (excl. pensions) and private transfers, and, on their turn, each of the five main income blocks is aggregated from further sub-components. However, while the disaggregation of total current income into the five main income blocks is always ensured (i.e. *hitotal = hilabour + hicapital + hipension + hipubsoc + hiprivate*), the disaggregation of each of these five blocks into further subcomponents is not always possible , as some amounts are reported directly at the higher level. For example, labour income (*hilabour*) will certainly include variables *hi11, hi12, hi13* and *hi14* (and in some datasets it may be exactly equal to the sum of those four sub-components), but it may also include an additional amount that was included directly at the level of *hilabour*. In datasets with very little level of detail available, total labour income was provided as a unique amount and hence directly put into *hilabour* (so that the four subcomponents are all zero).

A similar aggregation rationale applies to the individual level income variables: *pitotal* is always equal to the sum of all individual level variables (*pilabour*, *pipension*, *pi41*, *pi42 pi43* and *pi44*), whereas the subcomponents of each of them do not necessarily add up to the upper level aggregate. It should be noted that variable *pitotal* is always created, even when the subcomponents are incomplete; as a result, it cannot be ensured that it is perfectly comparable across datasets.

Finally, for other blocks of income and consumption variables (i.e. those not entering the construction of total disposable household income), the upper level variables are not necessarily equal to the sum of the sub-components (as some amounts may have been added directly at the upper level); similarly to *pitotal*, the existence of an upper level variable does not ensure that it is complete (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 expenditure (*hcexp*), as it can be empty even when some sub-components are filled (this happens if the original data only provided a subset of the total consumption).

## 10. 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 *immigr*, *educlev* and *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 *immigr* dummy, users can choose to work with the corresponding country-specific variables from which it was constructed - *citizen*, *ctrybrth*, *yrsresid* or *immigr_c* - and create their own definition of immigrant).

*Imputation* – While LIS always uses the results of imputations carried out by the data providers, it does not carry 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 by a data provider, a dummy flags the records imputed (see *fpimpu* for the individual level and *fhimpu* for the household level).

# II. Variable Groups

## 1. Income and Consumption variables

LIS datasets include a large amount of variables referring to incomes (current and extraordinary), income deductions and other transfers paid and consumption. 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 income deductions or payments are sometimes not included, while consumption is often not included (or not fully covered) and extraordinary incomes are rarely available.

All LIS income and consumption variables are reported in annual amounts and in units of the national currency in force at time of the survey (see variable *currency*). In addition to this, LIS provides a dataset that includes conversion factors for CPI adjustment and PPP adjustment to 2011 International Dollars. This dataset can be accessed online and through LISSY (see https://www.lisdatacenter.org/data-access/lissy/syntax/).

LIS income and consumption variables are split into the six major blocks:

- *Current incomes* (variables suffixed by *hi* and *pi*): these consist of cash 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;
- *Income deductions, transfers paid and loans repayments (*variables suffixed by *hx* and *px*): these consist of non-consumption expenditures such as taxes, contributions, donations, inter-household transfers and interest paid on loans;
- *Extraordinary incomes* (variables suffixed by *he* and *pe*): these consist of windfall gains and other such irregular and typically one-time receipts;
- *Imputed rent (variable hrenti)*: this section consists only of the imputed rent;
- *Consumption* expenditures (variables suffixed by *hc*): these consist of monetary and non-monetary consumption items, with the exception of imputed rent.

- *Major economic aggregates:* main aggregates mostly constructed from variables belonging to the above mentioned blocks and consisting of a series of income aggregates (including the overall concept of total disposable income and its major components) and consumption aggregates.

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 *LIS Variables List* for an exhaustive list of all the variables that exist also in the person file).

### 1.1 Current income

***Current income*** refers to all receipts whether cash or non-cash (value of goods and services, also referred to as in-kind incomes) 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 cash and in-kind income from labour, income from capital, pensions, cash payments from social security transfers (excluding pensions), and non-cash social assistance transfers, as well as cash and in-kind private transfers. There are two notable exceptions among the non-cash incomes:

- *Non-cash 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.

- *In-kind 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 in-kind 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 in-kind incomes are not available in the LIS microdata.

Although we state above that we include non-cash social assistance transfers, note that this does not mean that all non-cash 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 extraordinary (or windfall) income and inflows from sales of durables, and intake of loans.

Total current income is disaggregated into five major components:

***Current income from labour*** (*h/pilabour* and *h/pi1* variables): cash payments and value of goods and services received from dependent employment, as well as profits/losses and value of goods from self-employment, including own consumption.

***Current income from capital*** (*hicapital* and *hi2* variables): cash payments from property and capital (including financial and non-financial assets), including interest and dividends, rental income and royalties, and other capital income from investment in self-employment activity.

***Income from pensions*** (*h/pipension* and *h/pi3* variables): pension income from all pillars (private, occupational, public), all types (insurance, universal, assistance), all functions (old-age, disability, survivors).

***Income from public social benefits excl. pensions*** (*hipubsoc* and *h/pi4* variables): cash social security transfers (excluding public pensions) stemming from insurance, universal or assistance schemes, and in-kind social assistance transfers.

***Private transfers*** (*hiprivate* and *h/pi5* variables): cash transfers and value of in-kind goods and services of a private nature that do not involve any institutional arrangement between the individual and the government or the employer, including transfers provided by non-profit institutions, other private persons/households, and other bodies in the case of merit-based education transfers.

### 1.2 Income deductions, transfers paid and loans repayments

These consist of expenditures on non-consumption goods and services (such as taxes, contributions, donations, inter-household transfers and loan instalments.

### 1.3 Extraordinary incomes

Extraordinary (or 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, such as windfall labour income (severance pay, retirement packages), capital gains, inheritances, and other extraordinary income (insurance compensations, lottery winnings, etc.).

### 1.4 Imputed rent

This section consists only of the imputed rent variable, defined as the imputed value of the service of the dwelling owned by the household.

### 1.5 Consumption expenditures

LIS ideally records total consumption, including that stemming from expenditures (i.e. if the good or service consumed has been purchased by the household) and that stemming from own-production, transfers or gifts (goods and values consumed and not purchased, but either given to the household

from somebody else, or self-produced). Similarly, to the current income concept, there is a major exception in total consumption as well, i.e. imputed rent is not included.

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.

### 1.6 Major economic aggregates:

In order to facilitate the use of the LIS data, LIS has created some income and consumption aggregates that are commonly used in research. These aggregates are either derived from other LIS variables, in which case the calculation always follows the same formula to ensure prefect comparability across datasets, or are produced from original variables (in which case it is not possible to recreate them on the basis of other detailed LIS variables). Are listed below those aggregates that are not part of the current income disaggregation tree:

- *dhi*: disposable household income, which includes current total income net of income taxes and social security contributions (*dhi = hitotal - hxitsc*); *dhi* is the variable used for the LIS Inequality and Poverty Key Figures.

- *hvalgs*: total value of goods and services, which includes the in-kind income from labour (fringe benefits and own consumption), from the State and from private households or institutions (*hvalgs = hi13 + hi14 + hi47 + hi53*).

- *dhci*: disposable household cash income, which includes total cash current income net of income taxes and social security contributions (*dhci = dhi – hvalgs).*

- *hpublic*: public transfers, which corresponds to the social security redistribution, and includes public pensions and public social benefits excl. pensions; this variable is further disaggregated into:

  - *hpub_i*: insurance transfers: transfers stemming from social security systems where eligibility is based on the existence and/or the length of an employment status; 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;

  - *hpub_u*: universal transfers: transfers stemming from public programmes that provide flat-rate benefits to certain residents, 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;

  - *hpub_a*: assistance transfers: transfers stemming from public programmes that provide benefits especially targeted to individuals or households in need (i.e. with a strict income and/or assets 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.

## 2. Other variables

### 2.1 Technical variables

Technical variables consist of file identifiers, dataset level information, weights and imputation flags.

*File identifiers and dataset-level information* - Each file contains unique identifiers of its observations (*hid* for the H-level file and both *hid* and *pid* for the P-level file). In addition, in both files there is a series of dataset-level variables that provide information about the dataset and that are useful when working with multiple datasets (*did*, *dname*, *name*, *iso2*, *iso3*, *country*, *wave*, *year*, *currency* and *grossnet*).

*Weights* - There are two sets of weights that are always existing both at the household and at the individual level: the main weights that inflate to total population (*hpopwgt* and *ppopwgt*) and the normalised weights (*hwgt* and *pwgt*). 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.

*Imputation flags* - Variables *fhimpu* and *fpimpu* flag household or individuals in case the incomes have been substantially imputed by the data provider; this typically flags cases where a household or an individual did not report a large section or the totality of the income, which was then partially or totally imputed by the data provider.

### 2.2 Geography and Housing variables

These variables are all household level variables and report geographic as well as dwelling information.

*Geographical characteristics* 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*) trying to keep as much as possible the country-specific definition of rural areas.

Dwelling characteristics contain information about the dwelling that is the principal residence of the household: the tenure (owned versus not owned) and the type of dwelling (house versus multi-unit residential building or other type of dwelling).

### 2.3 Household composition and living arrangements

These variables contain household level information about the composition of the household and individual level information about living arrangements of household members.

*Household composition* variables are generally 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).

*Living arrangement* variables include the relationship of each household member to the head of the household, as well as indications on whether the household members have a cohabiting 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 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 typically available for all household members (with

some major exception for the data from early years), the presence of a partner, parents or children is often available only for household head (and implicitly his/her partner when there is one).

## 2.4 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 four major blocks: demographic characteristics, immigration, health and education.

a) Demographic characteristics

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, for adults, or up to a certain number of adults). Marital status is usually available for adults only, except for countries where is customary to marry at early ages.

b) 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. that could be the content of *immigr_c*), 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 in METIS for a precise indication of the rules used for its construction).

c) Health

Any available information about disability status has been used to create a dichotomous variable (*disabled*) to identify persons with disabilities/chronic diseases that limit them in their daily activities. Additional information on health is provided in the country specific variable *health_c* that includes mostly subjective health status measured on a scale from poor to excellent.

d) Education

Education variables include information about highest education level achieved by a person, total years of education, as well as an indication of whether the person is currently enrolled in education and some additional education-related information.

The *highest education level* as provided by the data provider is reported in variable *educ_c*; ideally, this refers to the highest completed level, but in some cases, it might refer to the highest attended level. Variable *educlev* contains the highest education level harmonised according to the 8 major groups of the Standard Classification of Education (ISCED) 2011. Finally, variable *educ* provides a 3-category classification of highest completed education (fully derived from the *educlev* variable),

where low stands for less than upper secondary (corresponding to ISCED levels 0, 1 or 2), medium for upper secondary and postsecondary non-tertiary (completed ISCED levels 3 or 4) and high for tertiary (completed ISCED levels 5 to 8). See Appendix for details about ISCED.

The number of education years (*edyrs*) has been derived from the highest completed level according to the average duration of each cycle in a first stage (see variable definition in METIS for more details) and when additional information will be available in the original data the variable will be filled with more details. The *illiterate* dummy flags individuals who cannot read and write in any language. Country-specific information about the education level of the mother and the father of the household members is also provided.

## 2.5 Labour market information

These variables report the major labour market characteristics of the household members (all are individual level variables with the exception of *farming* which is at the household level). They are split into two major blocks: labour market activity and characteristics of the main job.

a) Labour market activity

This section reports variables about the labour activity status, as well as information about informality, employment intensity and work experience.

*Labour activity status* is reported in the labour force status variable (*lfs*), which aims at capturing the main current activity status. If the current main activity status is not available, the main activity status in the income reference time will be used. In case the latter is not available either, the employment status according to ILO criteria in current period will be used instead (dataset-specific notes in METIS inform about the exact situation in each dataset). The labour force status distinguishes between the employed (those for whom work is the main activity during the current period), unemployed (job seekers) and not in labour force (inactives). Among those not in labour force the variable distinguishes between those retired from a job or business, disabled, those enrolled in education 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).

The dummy variable *emp* (fully derived from *lfs*) provides an employment flag that is existing in most datasets and that is mostly consistent with the job characteristics variables. On the other hand, the dummy *emp_ilo* (which flags individuals employed according to the ILO criteria) is filled only when the information was available in the original data. Are considered ILO employed persons who worked for at least one hour for pay or profit in the short reference period or had a job but did not work in the short reference period due to temporary absence from the job because of sickness, maternity leave, holidays, etc. or due the nature of their working time arrangement, such as shift work, etc.

Please note that activity status variables are typically defined only for those persons who answered the labour market individual questionnaire (adult persons for most datasets). The only household level variable of this section (*farming*) flags the households that are carrying out farming activity (crops and/or livestock).

Of particular importance for the lower and middle income countries, the *informal* dummy aims to capture information about the involvement in the informal sector. For employees, this could refer to a situation where they work without a working contract or they do not contribute to the social security system, they work in an unregistered business, they do not beneficiate of legal rights (right to pension, paid leave, etc.) or their wage is under-declared. For the self-employed, this could refer to a situation where they own an unregistered business when the legislation in the country requires them to register it, they do not pay taxes and/or contributions if they have to pay them. Those who produce goods and services only for their own consumption are not included in the definition of informality of this variable.

*Several measures of overall employment intensity* include indicators of multiple jobs holder (*secjob*), the number of hours worked per week (*hourstot*) and number of weeks worked in a year (*weeks*). On the basis of the number of weeks and hours worked, LIS then creates an overall full-year full-time dichotomous variable (*fyft*). There is also a dummy that flags parents in maternity/adoption or parental leave (*parleave*).

Please note that different employment intensity variables are defined for different groups of the population: variables about hours worked and multiple job indicator 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), 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 if the person worked at some point during the year).

Finally, variable *wexptl* refers to the overall work experience, with indication of the existence of such experience, as well as its duration (total years of total work experience). This variable is typically defined for the same group of persons for which the labour status variable is defined (mostly adults).

b) Characteristics of the main job

These variables report the most important job characteristics for the main job of every employed individual. They include the status in employment (dependent employee versus self-employment), the industry, the occupation, the sector of employment (private versus public), duration of the employment (permanent/long-term versus short-term contract), the work intensity (weekly hours worked), and an indication of the income level (as indicated by the hourly wage rate).

All the job characteristics variables are fully standardised, when possible referring to existing international classifications (see Appendix for a description of those standard classifications). In addition, occupation and industry are also available in country-specific format (*occ1_c* and *ind1_c*), which report the classifications at the most detailed level available in the original data.

Job characteristics variables for the main job are defined only for persons who currently have a job /typically those employed in *emp*).

# 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 (Nomenclature of territorial units for statistics)***
  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 2011 (International Standard Classification of Education, 2011)***
  Adopted in November 2011, the current ISCED 2011 replaces ISCED 1997 (see below) with improved definitions for types of education and an updated framework for the comparative analysis of educational levels across countries. The detailed classification contains 3 digits coding of each level, however for the construction of our EDUCLEV variable we use only first digit levels as follows:

  ISCED level 0 – Early childhood education

  ISCED level 1 – Primary education

  ISCED level 2 – Lower secondary education

  ISCED level 3 – Upper secondary education

  ISCED level 4 – Post-secondary non-tertiary education

  ISCED level 5 – Short-cycle tertiary education

  ISCED level 6 – Bachelor's or equivalent level

  ISCED level 7 – Master's or equivalent level

  ISCED level 8 – Doctoral or equivalent level

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

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).