The 2011 Template
New *LIS* Data Template / New *LIS* Documentation Framework

Motivation and Achievements

Outline of Document

I. Motivation and Philosophy

Two major factors motivated the timing and content of this restructuring:

- The inclusion of an increasing number of datasets from middle-income countries, which necessitated some conceptual adjustments and changes to our list of harmonised variables.

- The recognition that, while the previous template revision (*i.e.*, Wave V.2) increased the quality of the harmonised *LIS* data, it did not necessarily increase its user-friendliness.

Thus, the main objectives of this restructuring have been to adapt the *LIS* Data template:

- To maximize its applicability to datasets from both high- and middle-income countries.

- For a more user-friendly structure for *LIS*' data and documentation.

We aimed to meet these objectives, while maintaining the high-quality standards that we have maintained in the past.

As a result, our design of this revision was guided by these principles and goals:

- To restructure all variables, especially within the block of income variables, to achieve a more logical, comparable, and comprehensive list.

- To standardise most of the variables, which led us to use fewer country-specific codes and to eliminate the universe concept.

- To introduce easy-to-use dummy or categorical variables, to complement the more detailed ones that we will still provide.

Overall, our goal has been to produce a revised template that both increases over-time and cross-national comparability, and that requires our data users to make fewer assumptions and to do less recoding as they carry out their research.

- **We Will Continue to Grow, Improve, and Innovate**
  As summarized in this document, we have designed and implemented several important changes to the *LIS* Data template and to the *LIS* documentation framework. We are confident that our data users — both experienced users and new users — will benefit substantially from these extensive revisions. We look forward to your feedback, as you work with the new data template and the revised documentation.

II. Achievements

II.A. Changes to the *LIS* Data

The main changes to the data template can be summarized as follows:

- Renaming all of *LIS* variables, to make the variable names more logical and user-friendly. The old list, after a long succession of incremental changes, had become outdated, inconsistent, and illogical.

- More standardisation of the variables, at several levels:

  - Standardisation of the definitions/concepts, such that variables with a single name contains one only concept. In the past, the *LIS Database* contained several variables that had the same name across datasets and contained information related to the same topic, but where the concepts and definitions varied. For example, the former variable focused on immigration contained, across datasets: immigration status, country of origin, or number of years since arrival in the host country.

  - Standardisation of the value labels, such that most variables have fully standardised value labels as well. In the past, few *LIS* variables had exactly the same value categories across datasets. Now only a subset of country-specific variables — suffixed by *a_c* — contains dataset-specific codes and value labels.

  - Standardisation of missing and non-applicable values, such that non-valid values are more consistently coded across variables and datasets. Now, non-valid observations are coded in the same way for all variables, in all datasets, regardless of the values recorded in the original data. As a result, the concept of a variable- and/or dataset-specific universe is no longer present in the *LIS Database.*

- Adjustment of the disposable household income concept, such that disposable household income now includes non-monetary income from labour and from public and private third parties. Because of data limitations and comparability concerns, imputed rent is not yet included in our disposable household income measure, as is not the value of the goods and services provided by universal health and education systems. This change goes in the same direction of the revised Canberra Group Handbook on Household Income Statistics.

- Restructuring of the income and expenditure variables into five major blocks: i) current incomes, ii) windfall incomes, iii) non-consumption expenditures, iv) consumption expenditures, and v) assets and liabilities transactions. We now refer to these five blocks of variables as the "*LIS* flow variables".

- Revision of the current income variables, with a view to construct revised income aggregates that are:

  - More comparable across datasets (subject to data availability).
  - More meaningful for research purposes.

- Introduction of a set of technical variables that will make it easier to carry out multi-country analyses, including country-, dataset-, wave-, and year-level identifiers, as well as normalized weights.

- Inclusion of variables containing information on an array of new topics, including type of dwelling, internal migration, parents' education, involvement in marginal work, duration of unemployment, characteristics of second job, non-monetary incomes from private sources, and inflows/outflows of assets and liabilities.

- Creation of some easy-to-use new dummy/categorical variables. These include dummy variables that capture the following: rural residence, married/in union status, being an immigrant, being disabled, and being currently enrolled as a student. New dummies also capture labour market characteristics, including being employed, unemployed or retired, being in the military, working full-year full-time, working part-time in all jobs, and being a multiple jobholder. We also provide new categorical variables that capture educational levels (low, medium, high), as well as standardised major industry and occupation groups. These new variables were constructed, making the best assumptions possible, given data availability. (Notes, especially about assumptions made, and warnings have been added to the documentation.)

In addition to benefitting users, many of these changes will also improve our production process, especially by allowing us to harmonise datasets more quickly. In association with this template revision, we have also introduced some new processes that will prevent the possibility of errors during the harmonisation process and that will help to guarantee the consistency of construction rules across datasets.

II.B. Changes to the *LIS* Documentation

We have also revised and simplified the *LIS* documentation. The main change entailed shifting the focus of the documentation from describing the harmonisation process to documenting deviations from ideal contents.

For each *LIS* dataset, the technical documentation will now include:

- A standard document describing characteristics of the original survey (as before).

- A standard codebook, providing for each *LIS* variable, value labels, a set of descriptive values, and notes that warn users about variables whose contents depart from the ideal (as outlined in the generic *LIS* variables description).

- A standard description of relevant taxes and transfers, with links to the *LIS* variable(s) in which these institutions are captured in the microdata.

Our revised, simplified documentation will allow users to concentrate on the differences across *LIS* datasets rather than on our harmonisation process, which was the focus of the prior documentation. As with the data changes that we are now introducing, these revisions to the documentation framework will allow the LIS staff to harmonise/document datasets more efficiently and quickly.

III. Applicability

The new *LIS* Data template, and the new documentation framework, are being applied both forwards (into the future) and backwards (to earlier waves).

- Forwards. The new template/documentation will be used when we harmonised all future datasets. That includes the datasets contained in the forthcoming Wave VII (centered on 2007), as well as all datasets that will be added in the future to Waves I – VI.

- Backwards. All existing *LIS* datasets (Waves 0 – VI), and their associated documentation, have been revised to fit this new template structure — *as much as possible*. Please note that the new data template has been applied to the harmonised *LIS* datasets, and not to the original data. Thus, in the earlier waves, these revised datasets do not contain more information that was contained before. We explain this in more detail below.

IV. Implications for Experienced Users

What does this mean for users who have worked with the prior template? Experienced users ought to consider these points:

- Because all *LIS* variables have been renamed and the codes revised, current users will have to revise their existing programs/codes to make use of the new variable names and codes. We understand that this may cause some temporary inconvenience! We have taken some steps to minimize problems, which we discuss in the next section.

- At the same time, much less recoding work will be needed before datasets can be compared.

- The revised disposable household income measure, which is the basis of the *LIS Key Figures*, is now based on a slightly different concept (*i.e.*, it now includes non-monetary incomes). In practice, this change has little effect in datasets from high-income countries; we introduced this change mainly to improve the comparability between high- and middle-income countries. Note that the "old" disposable household income indicator will still be made available.

Please note that any users wishing to work with the pre-revised datasets (*i.e.*, harmonised in the pre-2011 data template) may continue to access them, as well as the associated documentation.

V. Backwards Revisions

As we noted above, the new *LIS* data template, and the new documentation framework, are being applied not only forwards, but backwards as well.

For all existing *LIS* datasets, every *LIS* variable has been remapped into the corresponding new variable(s). Depending on the variable, this process has involved a more or less automatic mapping. Here, we elabourate on this process:

- Full automatic recode. Some variables have been automatically renamed into a common new variable that is included in all datasets. For example, the old COUNTRY variable, a country/year identifier, has been renamed *did*. Other variables have been renamed, and in some cases slightly modified (*i.e.*, the variable formerly-named PPNUM is now named *pid* and is coded slightly differently).

- Dataset-specific automatic recode. Some variables have been renamed, and their contents shifted into new variables, depending on the contents of the old variable. For example, the old PDISABL variable has been disaggregated and renamed *disabl_c*, *illness_c,* or *health_c* — corresponding to information on disability status, chronic illness, or subjective health, respectively.

- Manual recode. Most variables have been completely remapped into one or several new variables. For example, the old PREL has been recoded to fit the standard codes of the new *relation* variable; the old PACTIV has been recoded to fit the standard codes of the new *status1* variable.

- Creation of completely new variables. Some variables have been created *ex novo* based on information included in the documentation; these include country-, year- and wave- identifiers, a flag for gross/net datasets, and a variable that reports the currency unit. Other variables have been constructed based on information included in the microdata; these include a series of person counters at the household level, a household composition variable, a series of living arrangement variables, as well as a set of new dummy/categorical variables, as described earlier in this document. In some cases (*e.g.*, for the rural, immigrant, and disabled dummies), these new variables were created on the basis of assumptions that we clearly state in the documentation of those variables.

- New industry and occupation variables. We carried out a large project that enabled us to standardise these variables. We recoded the original industry data into either the 17 major NACE groups, and/or into two more aggregated standardised variables, one with 9 and one with 3 categories. We also recoded the original occupation data into the 10 major ISCO groups, and/or into another more aggregated standardised variable with 3 categories. Note that the original classifications have been left alongside the new standardised variables.

In addition, other improvements were applied backwards (*i.e.*, to the existing *LIS* datasets):

- In datasets where the person-level file was either non-existent or incomplete (*e.g.*, it contained heads and spouse only, or adults only), artificial records were created, so that for all datasets the number of person-level observations will always match the number of persons in the household, as reported in the household-level file.

- In datasets that contained single and multifamily units, due to the inclusion of (overlapping) occurrences within households, the duplicate records were deleted.

- In datasets where the weights did not inflate the sample size to the size of the total population, that inflation was imposed.

- In datasets where the monetary variables were not expressed in units of national currency, corresponding to the time of the survey, the monetary variables were converted to those units.

- In some datasets, specific variables were corrected such that obvious errors were removed.

Finally, throughout the earlier *LIS Database* waves, the existing documentation has also been adapted to the new documentation framework:

- The lissification table (or lissification viewer) and the basic descriptives have been combined into a more comprehensive codebook; this codebook contains all *LIS* variables and is standardised across all datasets.

- The institutional documentation has been reviewed and revised to ensure that all referenced variables are the new ones in which the taxes and transfers are now included.