# Technical introduction to the LIS databases: the harmonization process, database structure and access to data

Markus Jäntti[1,2]

[1]Luxembourg Income Study [2]Swedish Institute for Social Research, Stockholm University

LIS Japan Workshop
13 January

# General Framework for LIS

- Deliverable: cross-national database that allows international comparative research using micro-data
- Scope: developed economies, to be extended with middle income countries

# Harmonization and LIS

- The harmonization process
- The LIS databases
- The LIS files + contents

# The origins of the LIS surveys

- LIS does not organize surveys
- LIS collects household surveys that have:
  - national coverage (all population groups, all geographic areas)
  - good quality income information (allows disposable income to be measured)
  - labour market variables

# The harmonization process

1. getting the data + documentation
2. file re-structuring (uniform units)
3. initialize standard set of output variables
4. mapping information into output variables
5. recode / standardize the values
6. apply standard missing values policy
7. annualize all income / expenditure
8. store final data in 3 formats (sas,spss,stata)

# The harmonization process

recommendations:

- use best practices (Canberra report)
- use international standard classifications (ISCO, ISCED, ISIC, ILO-definitions)
- provide extensive documentation
- use contacts inside NSO + country experts

# The harmonization process

comparability challenges:

- surveys change over time
- cross-section versus panel (sample)
- level of detail differs (HBS versus LFS)
- net versus gross income
- missing or imputed values

# All databases

- LIS = income
- LES = employment (discontinued)
- LWS = wealth

in total over 200 datapoints (majority LIS)

# All databases

unit of analyses per database

- LIS = individuals + household
- LES = individuals (adults)
- LWS = households

# LIS database

- the two dimensions are : country + time
  - nearly 40 countries (OECD + Latin America)
  - nearly 4 decades span, grouped by wave
- Number of countries increases with more recent waves over-time analyses (be aware: not longitudinal!)

# LIS database

LIS-specific anonimized identifiers :

- Datapoint identifier for country plus year (ccyy)
- Household identifier to merge individuals and households
- Person identifier (head, partner, etc.)

# LIS database

Final output per datapoint

- individuals : person file (ccyyp)
- household units : household file (ccyyh)

eventually plus a shadow file for each

# LIS database

LIS Person file

- demographic variables
- labour market characteristics
- income variables
- identifiers

NOTE : every var ALWAYS exists, but not always filled ! Use Variable Matrix to find out

# LIS person file

Demographic Information

- age, sex, relationship to the head
- marital status / partnership
- citizenship, immigration status
- general / vocational training
- enrolled in education
- disability status

see Quick Reference Manual

# LIS person file

Labour market variables

- labour force status
- job characteristics, such as occupation, industry, sector
- caregiving status , work experience

# LIS person file

**Income Variables** that are meaningful at individual level
therefore, no such income as:

- housing allowance
- child allowance

# LIS household file

- household composition and characteristics
- socio-demographic characteristics of head and spouse
- income variables
- expenditure variables

# LIS household file

- household characteristics
  - tenure of residence, region, urbanization
- composition (derived from hhld-members)
  - number of persons, earners, elderly
  - number of children under 18*
  - age of youngest child*
  - number of persons under 14 (needed for “modified OECD” equivalence scale)

* “child” = not head, nor spouse, never married

# LIS household file

- socio-demographic characteristics of head and spouse
- simply derived (copied) from person file
- includes age, marital status, lfs, education etc.

# Income Variables

- large set of variables, covering all sources such as income from labour, property income, transfers, taxes
- available over 3 levels of detail :
  - main level, names starting with v such as unemployment benefits (v21)
  - highly detailed sub variables
  - aggregated summary income variables

# Income: subvariables

V8: Cash Property Income

- interest and dividends v8s1
- rental income v8s2
- private savings plans v8s3
- royalties v8s4
- cash property income n.e.c. v8sr

available from wave IV onwards only

# Summary Income Variables

MI: Market Income

- Wages and Salaries v1
- Self-employment Income v4 + v5
- Cash property Income v8
- Occupational and other pensions v32 + v33

# Summary Income Variables

TRANSI: Total transfers

- government transfers: social insurance benefits v16-v24
- government transfers: social assistance (means-tested) v25-v26
- private transfers v34-v35

DPI = net disposable household income

# Income variables

## WARNING

net versus gross datasets!
see overview on-line www.lisproject.org/techdoc/netdatasets.htm

## DPI is gold standard

since this guarantees comparability
What is not included in DPI?

- irregular lumpsums like lottery winnings, inheritance
- imputed rent
- non-cash income

REMINDER: all amounts in national currency

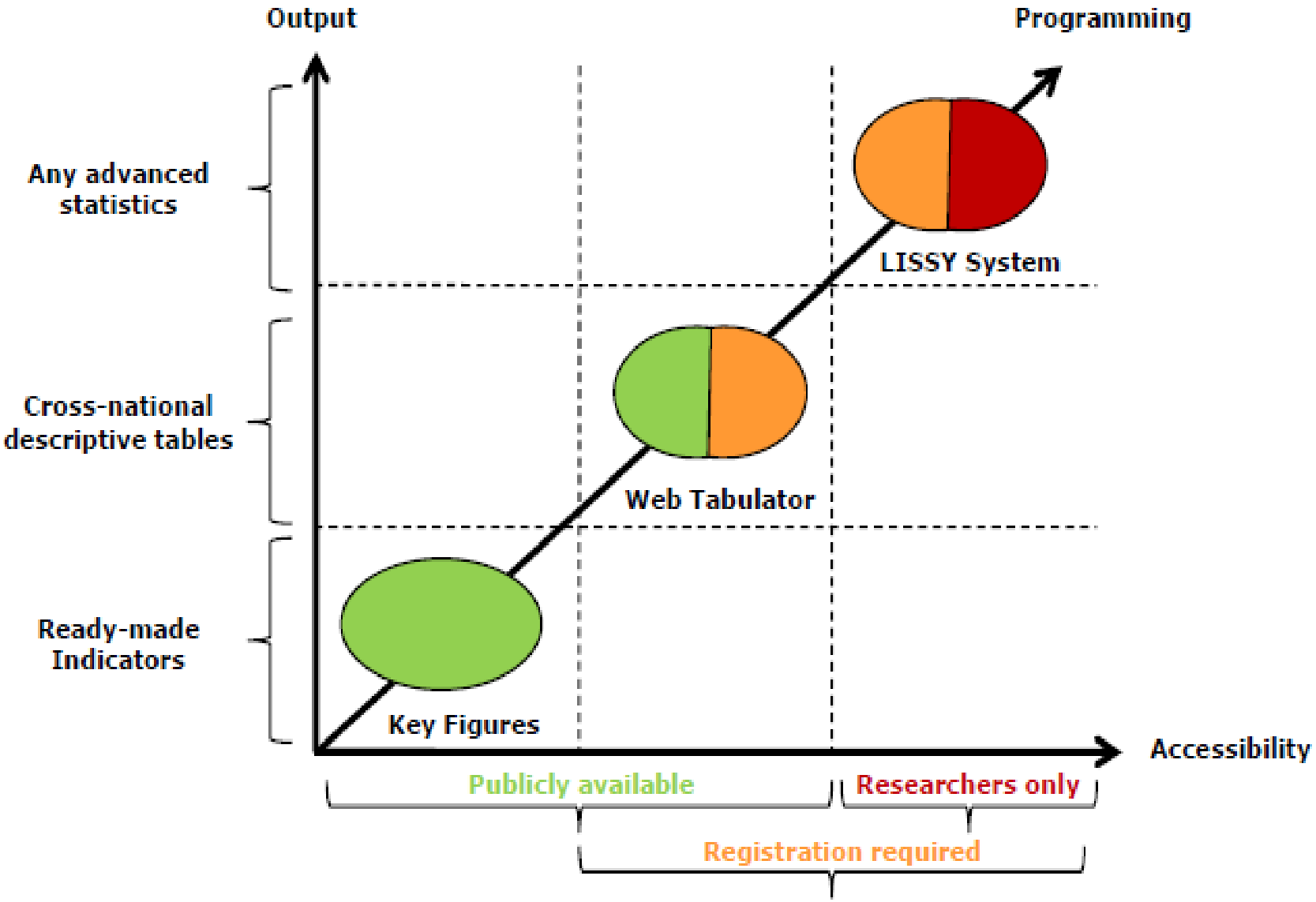
# Expenditure variables

- Coverage: Total consumption expenditure as defined in the COICOP classification
- 12 major groups, ranging from food, housing, medical expenditures, etc, to miscellaneous
- Total expenditure only filled when surveyed

# Important notes

- Weighting
- Missing values
- HWEIGHT: Sample Weight
- HWEIGHT: Sample Weight
- Missing values

# Three Pathways to access LIS Databases



# Three pathways to access LIS Databases

- The LISSY System
- The Web Tabulator
- The LIS Key Figures

# The LISSY System

- The primary means of access is a fully automated software running 24/d 7/week designed specifically for LIS
- SPSS, SAS or Stata batch programs submitted via a Job Submission Interface (JSI) or an email software
- LISSY automatically processes the jobs, generates results and returns them to users on average within two minutes

# Registration Process

http://www.lisproject.org/data-access/lissy.htm

## Initial Registration process

- Download, complete, sign and send back a pledge from the LIS website, which states the rules governing the use of the micro-data
- The `userid` and a `password` are strictly personal and must not be shared with anyone
- Researchers provide an email address when registerign. LISSY returns output to this email address

## Renewal Process

Access to the micro-data for a limited period of one year, renewable annually. Can be completed by annually updating information through the LIS website

http://www.lisproject.org/data-access/renewal.htm

# Structure of a job

Jobs submitted to LISSY require a few exceptions to the usual program syntax

- LISSY relies on a three-stage built-in alias to access datasets, consisting of
  - a statistical package specific heading (&,$ or 'blank')
  - a two-digit ISO country (`http://www.lisproject.org/techdoc/countryid.htm`)
  - an abbreviation used to identify the specific type of dataset

  $IT04h to access the LIS 2004 household Italian dataset (Stata)
- LISSY rejects programs containing commands that allow users to print or read individual records
  We filter out jobs that risk breaching the rules on data confidentiality. LISSY automatically puts the job in a security review area to be manually reviewed by the staff
- When syntax errors are detected, the language-specific error message are display within the listing

# Submission via Email

- Users submit job requests via outlook, thunderbird etc. to postbox@lisproject.org
- Three requirements must be met in order for LISSY to properly process jobs
  - All emails must be sent in ascii/plain text format. Users must ensure that this option is enabled in the chosen email package.
  - All job instructions must be written inside the body of the email and not as an attachment.
  - Each job must start with a specific four-line header at the very beginning of the email body

```
* user      = your username
* password  = your password
* package   = your statistical package
* project   = LIS or LWS
```

- Same exceptions to the user's programming style are required

# LIS Web Tabulator

http://www.lisproject.org/data-access/web-tabulator.htm

## What is this?

An online table-making service enabling the design and generation of crossnational descriptive tables based on the underlying LIS datasets without the need for programming

## Registration required

Requires a username and password, see http://www.lisproject.org/data-access/web-tabulator-registration.htm for the expedited registration process

# LIS Web Tabulator

http://www.lisproject.org/data-access/web-tabulator.htm

- Includes a restricted collection of datasets and variables
  - Datasets from LIS' Wave V and Wave VI
  - Household-level data only
  - Standardized indicators, including multiple indicators of real household income, as well as demographic and labor market variables
  - Income variables expressed into 2005 international dollars
- Two types of cross-national descriptive tables
  - Univariate Statistics and Cross-Classified Statistics tables
  - To-Do list to guide designing tables
  - Possibility to export aggregated results in EXCEL or in ASCII formats

# LIS Key Figures

http://www.lisproject.org/key-figures/key-figures.htm

## Two distinct sets of LIS Key Figures

- LIS Inequality and Poverty Key Figures comprise national-level inequality and poverty indicators such as Gini Coefficients, Atkinson coefficients, etc (all waves)
- LIS Gender Key Figures that include national-level indicators highlighting women's economic outcomes and gender inequality in poverty and employment (Wave V and soon Wave VI)

## LIS Key Figures exist in two formats directly accessible

- Each set of Key Figures is included in a downloadable Excel workbook
- The Search Engines allow the creation of subsets of the two sets of LIS Key Figures to narrow the scope of the analysis.

# LIS Key Figures

Programming routines and information are publicly available

- All methodological decisions are transparent
- Alter the programs to fit the needs of individual research projects

# Last Minutes . . .

- If you have any problems contact the LIS user support `usersupport@lisproject.org`
- For users who may be unfamiliar with batch coding US00 (United States, 2000) and IT00 (Italy 2000) (sub-sample) files are downloadable (`http://www.lisproject.org/self-teaching.htm`)
- Key figures programs (`http://www.lisproject.org/key-figures/key-figures-programs.htm`)
- If you have any problems contact the LIS user support usersupport@lisproject.org