Income Database of Harmonized Microdata: LIS- Cross National Data Center in Luxembourg

Measuring Income and Wealth through Household Surveys for Welfare Monitoring
Perugia, 10-14 December 2018

Heba Omar
Data Expert, LIS
Outline of the presentation

I. LIS, the institution
II. The LIS/LWS Databases
III. Research possibilities
IV. Data dissemination & Documentation
V. Challenges of harmonizing microdata
   Emphasis on Middle income countries
VI. Major “dilemmas”
LIS golden rules for harmonization
Part I:  
*LIS, the institution*
LIS History

• LIS was founded in 1983 by two US academics (Professors Tim Smeeding and Lee Rainwater) and a team of multi-disciplinary researchers in Europe. It began as a “study”, which later grew and was institutionalized as “LIS”. (Note the “oral history” video on LIS website.)

• For nearly 20 years, LIS was part of a Luxembourg-based research institute known as CEPS (recently renamed LISER). In 2002, LIS became an independent non-profit institution (an ASBL).

• LIS is supported by the Luxembourg government (25%), by the national science foundations and other funders in many of the participating countries (50%), and by several supranational organizations – supplemented by project-specific grants and some private philanthropy (25%).
LIS Mission

To enable, facilitate, promote, and conduct cross-national comparative research on socio-economic outcomes and on the institutional factors that shape those outcomes.
LIS: new leadership structure
launched 1 September, 2016

Prof. Daniele Checchi
Director of Luxembourg Office of LIS

* He is Professor of Economics currently on leave from the University of Milan.

* He is currently working at the Italian National agency for the Evaluation of the University system.

Prof. Janet Gornick
Director of US Office of LIS

* She is Professor of Political Science and Sociology, at the Graduate Center of the City University of New York.

* The “LIS Center” has been renamed “The Stone Center on Socio-Economic Inequality”.

* The “US Office of LIS” is now an entity within the enlarged Center.
# LIS structure

**LIS - Cross-National Data Center**  
François Bourguignon, President  
LIS Executive Committee  
Daniele Checchi, Secretary General

<table>
<thead>
<tr>
<th>MAIN OFFICE: Luxembourg Office of LIS</th>
<th>SATELLITE OFFICE: US Office of LIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniele Checchi, Director</td>
<td>Janet Gornick, Director</td>
</tr>
<tr>
<td>• parent organization</td>
<td>• satellite office</td>
</tr>
<tr>
<td>• located in Luxembourg, together with the University and LISER</td>
<td>• located at the Graduate Center of the City University of New York</td>
</tr>
<tr>
<td>• independent, chartered non-profit organization</td>
<td>• one pillar of <em>Stone Center on Socio-Economic Inequality</em></td>
</tr>
<tr>
<td>• cross-national, participatory governance</td>
<td>• administrative, managerial, development support to parent office</td>
</tr>
<tr>
<td>• acquires, harmonizes, and disseminates data for research</td>
<td>• collaborative public programs (lectures, conversations)</td>
</tr>
<tr>
<td>• venue for research, visiting scholars, conferences, and user training</td>
<td>• venue for research, teaching, and graduate student supervision</td>
</tr>
</tbody>
</table>

**LIS Senior Scholars**
# LIS: who’s who?

## LIS - Cross-National Data Center

**François Bourguignon, President**  
LIS Executive Committee  
Daniele Checchi, Secretary General

### MAIN OFFICE: Luxembourg Office of LIS

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniele Checchi</td>
<td>Director</td>
</tr>
<tr>
<td>Lucie Scapoli</td>
<td>Administrator Officer</td>
</tr>
<tr>
<td>Benjamin Gérard</td>
<td>System and Network Administrator</td>
</tr>
<tr>
<td><strong>Data team</strong></td>
<td></td>
</tr>
<tr>
<td>Teresa Munzi</td>
<td>Data Team Manager</td>
</tr>
<tr>
<td>Paul Alkemade</td>
<td></td>
</tr>
<tr>
<td>Andrej Cupak</td>
<td></td>
</tr>
<tr>
<td>Jörg Neugschwender</td>
<td></td>
</tr>
</tbody>
</table>

### SATELLITE OFFICE: US Office of LIS

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Janet Gornick</td>
<td>Director</td>
</tr>
<tr>
<td>Caroline Batzdorf</td>
<td>Assistant Director</td>
</tr>
<tr>
<td>Mei-Ling Israel</td>
<td>Financial Manager</td>
</tr>
<tr>
<td>Laurie Maldonado</td>
<td>Senior Administrator</td>
</tr>
</tbody>
</table>

### LIS Senior Scholars

<table>
<thead>
<tr>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prof. Louis Chauvel</td>
</tr>
<tr>
<td>Prof. Daniele Checchi</td>
</tr>
<tr>
<td>Prof. Conchita D'Ambrosio</td>
</tr>
<tr>
<td>Prof. Markus Jäntti</td>
</tr>
<tr>
<td>Prof. Frank Cowell</td>
</tr>
<tr>
<td>Prof. Janet Gornick</td>
</tr>
<tr>
<td>Prof. Paul Krugman</td>
</tr>
<tr>
<td>Prof. Leslie McCall</td>
</tr>
<tr>
<td>Prof. Branko Milanovic</td>
</tr>
</tbody>
</table>
LIS’ partners

Our partners include data providers, data users, and funders, in more than 50 countries …
and in major supranational organizations, including:

Financial contributors:
World Bank (WB)
Organization for Economic Cooperation and Development (OECD)
International Monetary Fund (IMF)
International Labor Organization (ILO)
European Union (InGRID)

Dataset exchange; joint research projects; joint fundraising:
Economic Research Forum (ERF)
European Central Bank (ECB)
French Development Agency (AFD)
LISER
University of Luxembourg
Part II: The *LIS/LWS Databases*
Overview of the LIS data

**Deliverables**
Two cross-national harmonised databases that allow international comparative research using micro-data:

- LIS (focus on income): 339 datasets – this presentation will focus on LIS
- LWS (focus on wealth): 39 datasets

**Scope**
Initial focus on high-income countries, successively extended to middle-income countries

**Time span**
From the late 1960s to 2016

**Geographical coverage**
World-wide, but some regions are less covered (Africa, EECCA…)

**Main contents**
- household composition and characteristics
- socio-demographic characteristics of household members
- extensive set of labour market data
- detailed breakdown of household and individual income data
- household consumption data
- a detailed set of wealth and behavioural variables (LWS only)
Luxembourg Income Study Database (LIS)

- First and largest available database of harmonized income data, available at the household and person levels
- In existence since 1983
- Data mostly start in 1980, some go back to the 1960s (recollected every 3-5 years)
- Started with six countries; now 50 countries
- 300+ datasets (repeated cross sections)
- Used to study: poverty; income inequality; labor market outcomes; policy effects
Luxembourg Wealth Study Database (LWS)

- First available database of harmonized wealth data, available at the household level
- In existence since 2007
- 52 datasets from 16 countries – up or in process
- Revised and updated in 2016 (in coordination with Euro system's Household Finance and Consumption Survey - HFCS)
- Used to study: *household assets, debt, and expenditures; wealth portfolios; policy effects.*
The LIS Databases: LIS and LWS

LIS Database
- 339 datasets
- Time coverage:
  - Historical database
  - Wave I around 1980
  - Wave II around 1985
  - Wave III around 1990
  - Wave IV around 1995
  - Wave V around 2000
  - Wave VI around 2004
  - Wave VII around 2007
  - Wave VIII around 2010
  - Wave IX around 2013
  - Wave X around 2016
- Geographical coverage:
  - Europe: 23 EU and 6 non EU
  - America: 2 North, 3 Central, 1 Caribbean, 7 South
  - Asia: 4 East, 1 South, 3 West
  - Africa: 2 countries
  - Oceania: 1 country

LWS Database
- 39 datasets
- Time coverage:
  - Wave V around 2000
  - Wave IV around 1995
  - Wave VI around 2004
  - Wave VII around 2007
  - Wave VIII around 2010
  - Wave IX around 2013
  - Wave X around 2016
- Geographical coverage:
  - Europe: 10 EU and 1 non EU
  - America: 2 North
  - Asia: 1 East
  - Africa: 1 country
  - Oceania: 1 country
Countries in LIS and LWS Databases (N=50)

*approximately 65% of world population and 85% of world GDP*

<table>
<thead>
<tr>
<th>High-income countries (N=34):</th>
<th>Upper-middle-income countries (N=12):</th>
<th>Lower-middle-income countries (N=4):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>Brazil</td>
<td>Cote D’Ivoire (q1 ’19)</td>
</tr>
<tr>
<td>Hungary</td>
<td>Paraguay</td>
<td></td>
</tr>
<tr>
<td>Slovak Republic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austria</td>
<td>Iceland</td>
<td></td>
</tr>
<tr>
<td>Iceland</td>
<td>Slovenia</td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>China</td>
<td>Egypt</td>
</tr>
<tr>
<td>Ireland</td>
<td>South Korea</td>
<td></td>
</tr>
<tr>
<td>Canada</td>
<td>Spain</td>
<td></td>
</tr>
<tr>
<td>Israel</td>
<td>Dom. Republic</td>
<td>India</td>
</tr>
<tr>
<td>Chile</td>
<td>Italy</td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td>Guatemala</td>
<td></td>
</tr>
<tr>
<td>Czech Republic</td>
<td>Japan</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>Switzerland</td>
<td></td>
</tr>
<tr>
<td>Denmark</td>
<td>Taiwan</td>
<td></td>
</tr>
<tr>
<td>Estonia</td>
<td>Luxembourg</td>
<td></td>
</tr>
<tr>
<td>Luxembourg</td>
<td>United Kingdom</td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td>Netherlands</td>
<td></td>
</tr>
<tr>
<td>Norway</td>
<td>United States</td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>Norway</td>
<td></td>
</tr>
<tr>
<td>Norway</td>
<td>Uruguay</td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>Panama</td>
<td></td>
</tr>
<tr>
<td>Greece</td>
<td>Poland</td>
<td></td>
</tr>
</tbody>
</table>
LIS country coverage by end of 2018

LIS/LWS country coverage
Country not covered
Country covered
What we do at LIS

Core activity: **Data work**

**Step 1. Data acquisition**
- We identify appropriate datasets *(reliable, and high-quality data)*
- We negotiate with each data provider

**Step 2. Data harmonisation**
- Common cross-national template
- Comprehensive documentation

**Step 3. Data dissemination**
- We create national-level indicators *(LIS Key Figures)*
- We provide harmonized microdata to researchers via *remote execution*

**Other activities**
*Research-promotion activities* (conferences, work on methodological issues, collaborations with networks/users/journalists, newsletter, individual research)
Support (user support, research visits)
Ex-post harmonisation at LIS

The origins of the LIS data

- LIS does not organise surveys but collects data from existing data sources:
  - Survey data: income, household budget, living conditions, multipurpose, human development
  - Administrative records: tax records, employers records, social security records
  - Any mix of the above

- Common denominator:
  - microdata (household and individual level)
  - representative of the whole population
  - good quality income/wealth data
  - main demographic and (possibly) labour market information
Ex-post harmonisation at LIS

Final output: the LIS/LWS datasets (CCYY)
- Technical harmonisation: same file structure, same variables
- Conceptual harmonisation
  - Based on the same definitions
  - Comparable concepts

Harmonisation allows LIS users to eliminate many of the potential sources of technical and conceptual non-comparability
**LIS Database content**

- Household composition and characteristics
- Socio-demographic characteristics of hld members
- Extensive set of labour market data
- Detailed breakdown of hld / ind income data
- Household consumption data
- Some info. on assets and liabilities transactions

**Non-flow variables**

**Flow variables**

**Household disposable income**
## LIS Flow variables

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current income</strong></td>
<td>• 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.</td>
</tr>
<tr>
<td><strong>Windfall income</strong></td>
<td>• Windfall gains and other such irregular and typically one-time receipts</td>
</tr>
<tr>
<td><strong>Non-Consumption expenditure</strong></td>
<td>• Monetary expenditures (i.e. paid directly by the household and/or its members) and nonmonetary 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).</td>
</tr>
<tr>
<td><strong>Consumption</strong></td>
<td>• Monetary and non-monetary consumption items.</td>
</tr>
<tr>
<td><strong>Assets/Liabilities transactions</strong></td>
<td>• Monetary inflows that do not constitute income (sales of real estate, financial products, durables or inflows from loans) and outflows that do not represent consumption (purchase of real estate financial products of outflows from loans)</td>
</tr>
</tbody>
</table>
# LIS Household Disposable Income

<table>
<thead>
<tr>
<th>LABOUR</th>
<th>MONETARY</th>
<th>NON-MONETARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent employment</td>
<td>Wages, salaries, bonuses Profit and losses</td>
<td>In-kind earnings Own consumption</td>
</tr>
<tr>
<td>Self-employment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-employement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Financial investment</td>
<td>Interest and dividends</td>
<td></td>
</tr>
<tr>
<td>Real estate investment</td>
<td>Rental income</td>
<td>Imputed rent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CAPITAL</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial investment</td>
<td>Interest and dividends</td>
<td></td>
</tr>
<tr>
<td>Real estate investment</td>
<td>Rental income</td>
<td>Imputed rent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TRANSFERS</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Social security transfers</td>
<td>Insurance pensions and wage-replacement benefits</td>
<td></td>
</tr>
<tr>
<td>- Work-related insurance</td>
<td>insurance pensions and wage-replacement benefits</td>
<td></td>
</tr>
<tr>
<td>- Universal transfers</td>
<td>Universal pensions and universal benefits</td>
<td></td>
</tr>
<tr>
<td>- Social assistance transfers</td>
<td>Minimum income guarantee</td>
<td>In-kind social assistance</td>
</tr>
<tr>
<td>Private transfers</td>
<td>Inter-household transfers, transfers from charity</td>
<td>In-kind benefits from privates</td>
</tr>
</tbody>
</table>

= TOTAL GROSS INCOME

- Income taxes
- Social security contributions

= HOUSEHOLD DISPOSABLE INCOME
Part III: Research possibilities
Measuring Inequality Across Households

Gini Coefficient for LIS countries
/latest available year/

Measuring Poverty

Household Poverty Rates

Relative poverty rates for the overall population, children and elderly
% of individuals with household income below 50% of median, latest available year

Researching Policy Impacts
Inequality reducing effect of Redistribution

Gini Index on Market Income and Disposable Income and percent reduction
latest available year

Market income is defined as the sum of labour income, capital income, occupational pensions and private transfers. The number at the top of the bar represents the percent reduction of Gini after redistribution of public transfers and taxes. Countries denoted by an asterisk are net income countries (so that redistribution includes transfers only).

Comparing Employment Outcomes
Earnings Equality between Women and Men

Men's Employment Rate (%)
Women's Employment Rate (%)
middle income countries
high income countries
Earnings gap
0.55
0.30
0.10
0.02
Measuring SDGs: goal 10.1
By 2030, progressively achieve and sustain income growth of the bottom 40 per cent of the population at a rate higher than the national average.
Measuring SDGs: goal 10.1
By 2030, progressively achieve and sustain income growth of the bottom 40 per cent of the population at a rate higher than the national average
Part IV: Data dissemination & Documentation
Data access

Level of interest

Microdata

Substantive indicators

Simple descriptives

Flexibility for the users

Ready-made indicators

Tailor-made tables

Microdata

Key Figures

DART

Web-Tabulator

LISSY System

Simple descriptives

Substantive indicators

Microdata

Data access
LIS Key Figures

- Multiple country-level inequality measures (e.g., Gini and Atkinson coefficients, percentile ratios)
- Relative poverty rates for various demographic groups
- Median and mean disposable household income

Lissy: remote-execution system

- Fully automated, running 24 hours/day and 7 days/week
- Job Submission Interface (JSI): send statistical batch programs (SAS, SPSS, Stata or R) automatically processed and reports back aggregated results
- Micro-databases cannot be downloaded and no direct access to the data is permitted. Only aggregated results from statistical requests are sent back to the users.
- Access to Lissy is granted to researchers, incl. students, working for an academic, government or non-profit organization under the condition that use of the micro-data is restricted to research purposes only
How LISSY works 1/2
How LISSY works 2/2

How LISSY works 2/2

How LISSY works 2/2
METadatas Information System (METIS)

METIS enables **browsing**, **aggregating**, and **exporting** LIS/LWS Databases documentation tailored to the users’ needs

1. **Overview of datasets and variables**
Overview of the contents of the LIS/LWS databases in terms of datasets and variables

2. **Compare datasets**
Select datasets and compare generic information among them (original survey, Social Security, Key Figures)

3. **Compare variables**
Select variables and compare generic information among them (variables definitions and standard labels)

4. **Cross-compare** – main functionality
View the availability of the selected variables in the selected datasets and compare dataset-specific information (statistics and notes)
<table>
<thead>
<tr>
<th>COUNTRY</th>
<th>WAVE X</th>
<th>WAVE IX</th>
<th>WAVE VIII</th>
<th>WAVE VII</th>
<th>WAVE VI</th>
<th>WAVE V</th>
<th>WAVE IV</th>
<th>WAVE III</th>
<th>WAVE II</th>
<th>WAVE I</th>
<th>HISTORICAL WAVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>2013</td>
<td>2009</td>
<td>2011</td>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>2013</td>
<td></td>
<td>2007</td>
<td></td>
<td></td>
<td>2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td>2013</td>
<td>2010</td>
<td>2007</td>
<td>2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dominican Republic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egypt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Georgia</td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guatemala</td>
<td>2014</td>
<td></td>
<td>2011</td>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LIS/LWS Working papers

Working Paper Series

- 813 WP of which over 400 published in peer reviewed journals
- Aldi Award: best paper by an under 40 author
- included on RePec (Research Papers in Economics)

- **New WP:** 69 new WP added in 2017 and 2018.

- **RePEC:** Since May 2017, LIS/ LWS working papers have been downloaded more than 9,000 times

- **H-index:** LIS = 166  
  LWS = 50
Part VI:
Challenges of harmonizing microdata
Emphasis on Middle income countries
Challenges of harmonization - overall

Our main work calls for rendering comparable original data that are:

- from various countries
  - different institutional / societal features
- over time
  - changes in institutions and original surveys
- available at the household- and individual-level data (microdata)
  - confidentiality issues
- from various existing sources (in which we generally had no input)
  - output (or ex-post) harmonization
The challenges of *ex-post* harmonisation

- **Different types/purposes of original collection instrument**
  - Survey versus administrative data (coverage and contents)
  - Cross-sections versus panels (sample selection)

- **The concepts used in the original data collection are different**
  - Different definitions (employment concept, total household disposable income concept, gross vs net incomes, assets vs net worth, etc.)
  - Different universes and reference periods
  - Country-specific categories (especially in education, social security benefits)

- **The level of detail of information collected differs**
  - Labor market (e.g.: LFS type of survey)
  - Incomes/wealth (detailed breakdown vs. overall questions)

- **Different statistical techniques**
  - Different sampling procedures (e.g. oversampling of the rich)
  - Weighting procedures (self-weighted, sampling weights, etc.)
  - Treatment of missing values, imputation methods
The challenges of harmonising labour market data

Very difficult to create comparable variables

- Many international guidelines and recommendations, but those are usually only applicable to data from LFS
  - different definitions of employment and labour force

- Rigid routing of the labour market questions
  - different universes (labour market data availability, job characteristics availability)

- Labour market information refers to a variety of time points or ranges
  - different reference periods (current versus annual labour market data)

- Country-specific codes and classifications
  - different categories (occupation and industry, but also employment status)
The challenges of harmonising income data

- Scope of total household disposable income (irregular payments, non cash incomes, imputed rents, non-taxable incomes, “informal” incomes…)

- Net versus gross (or in between...)

- Top- and bottom-coding

- Level of detail (e.g. total pensions) and different aggregation (e.g. pensions by type of system versus by function)

- Classification of incomes:
  - Public versus private
  - Social insurance versus universal versus social assistance systems
The challenges of harmonising data from middle income countries Relative to high-income countries, middle-income countries/surveys:

- Urban versus rural (sample composition, population coverage)
- Household membership and treatment of incomes (live-in domestic servants, family members temporarily absent)
- Complex households (multigenerational households, definition of head, polygamy)
- Employment definition and labour market characteristics (informal employment, child labour, multiple jobs, status in employment)
- Education (attended versus completed, highest level versus highest qualification)
- Enlargement of income concept to in-kind incomes (consumption from own production, in-kind individual public goods, subsidies)
- Classification of income:
  - Employer-provided pensions and benefits (labour income, social security)
  - Social insurance versus assistance versus universal benefits
- Treatment of taxes
**Issue:** analyses centered around the head/spouse may become biased

**Solution:** importance of providing detailed living arrangements for all household members (possibly pointers to partners and parents)
**Issue:** measurement of employment (artificially high), inconsistency with earnings

**Solution:** importance of providing detailed information about marginality / informality of employment
Treatment of non-cash incomes

- Relative importance of non-cash incomes (especially consumption from own production, in-kind individual public goods, subsidies) w.r.t. high income countries
- LIS includes those in total disposable income

BUT

- Not always available (incomparability)
- Risk of double-counting when in-kind incomes are collected in the consumption module and added to the incomes (own consumption of food is often counted in the incomes from self-employment farming activity as well)
Measurement of self-employment incomes

Issue is of particular relevance in middle income countries as self-employment is much more common

- Underestimation: self-employment incomes are much harder to capture in general, even more so in presence of informal/marginal employment.

- Risk of double counting: when collected both at the individual level and in the household business sections of the questionnaire.

- Individual level self-employment income not measurable: very small portion collected at the individual level (possibly only the incomes of the liberal professions), while other self-employment incomes are collected at the household level.
Measurement of individual level self-employment income

Percent of total labour income available at the individual level
(countries where self-employment income is not fully available at the individual level)
Missing (or inconsistent) income data is much more prevalent

**Issue:** bias due to non-random distribution of households with no income information

**Solutions:** calculation of the weights and imputation
Part VII: Major “dilemmas”
1. Harmonization dilemma: Determining the optimal level of detail

Microdata, by definition, incorporates fine-grained detail. In constructing the LIS Database, we expanded the level of detail over time – sometimes gradually, sometimes in “large bursts”.

- e.g., additional upper-level variables,
- more sub-level variables,
- more highly disaggregated value categories,
- more variables available in both standardized and country-specific forms, and
- multiple sets of aggregates.

On the one hand users often request high levels of detail. Potential strategy: maximize accuracy as well as comparability.

- E.g., we replaced one categorical employment variable with three variables, to allow us to not mix original data based on different definitions and reference periods.
1. Harmonization dilemma:
Determining the optimal level of detail (continued)

On the other hand, too much detail can threaten comparability.

- E.g., when income transfers are too highly disaggregated, it can be difficult to fill lower levels and still maintain cross-dataset comparability.
- Excessive detail can reduce user-friendliness; too much disaggregation prompts users to carry out their own collapsing and recoding – which, decreases comparability across research projects based on the LIS/LWS data.

Finally, filling an overly-detailed variable list is labor-intensive for the data team, creating a concrete trade-off between within-dataset detail and the volume of datasets that can be produced.
2. Data quality dilemma: Establishing our level of intervention in the microdata

LIS does not field surveys. We harmonize and make available “other people’s data”.

- we are not the “owners” of the data; we are the “keepers” and “custodians” of others’ data.
- We are indebted to our data providers, and it is crucial that they trust us and have confidence in our work.

→ That framework suggests that we make **as few changes as possible** when we harmonize.

On the other hand, over a period of 30 years, we have acquired expertise that would allow us to improve the quality of the microdata that we provide, e.g. by cleaning, editing, imputing, simulating, or otherwise altering the original data.
3. Expansion dilemma:
Weighing data “quantity” versus data “quality”

As we plan and carry out our work, we grapple with a fundamental question: How should we balance, essentially, data quantity versus data quality?

On the one hand, we feel pressure, largely from users and funders, to …
• add more countries, more years per country, or both.
• expand in other ways – by adding new blocks of data (e.g., material deprivation, subjective wellbeing, health) and new types of data (e.g., panel data, administrative data).

On the other hand, adding more data, in all forms, is labor-intensive, and we have to preserve time for other data-related tasks aimed at increasing the quality, usability and visibility of our harmonized data. These include investing more time in assessing datasets, conducting external checks, testing new practices, and other products next to the data work (e.g. documentation, indicators, self-teaching packages, reports and newsletter).
4. Documentation dilemma: Balancing back-end (staff) resources against front-end (user) needs

We have always constructed documentation with many elements. We provide information about …

- e.g., the original surveys/datasets,
- the content of our harmonized variables, and
- the rules of corresponding tax and transfer programs.

On the one hand, we have long felt an obligation to produce extensive documentation, to provide transparency for our data providers, to create records for our internal purposes (which is crucial over time), and to enable users to carry out the highest quality research.

On the other hand, producing documentation is labor-intensive, forcing a trade-off with other types of work.
5. Evolution dilemma: Balancing improvements with preservation of trends and practices

Over LIS’ life course, we have introduced countless changes to our microdata, to reflect internal learning on our part, changing external environments, and/or new international data practices.

These changes have included small- and large-scale template revisions, revisions/corrections made to specific datasets, and many incremental shifts in our harmonization practices and guidelines.

On the one hand, these improve data quality and modernize our practices.

On the other hand, users share the cost of this change, as it forces them to revise and update their programs and research practices – which can cause genuine hardship.

What we have learned: At best carry out revisions consistently, prepare the change well in advance and apply it systematically to the past when possible, and at best release it at once. Keep the versions of data to a minimum.
6. Institutional dilemma:
Balancing independence with long-term sustainability

LIS is a small, non-profit institution. Although we collaborate with public institutions and universities, we are an independent non-profit (ASBL) – a structure that we have purposely maintained over a many years.

On the one hand, our structure ensures that we maintain independence and autonomy. LIS’ small size and research orientation – compared with larger, public or quasi-public institutions – grant us a high level of flexibility.

On the other hand, we have to scramble to maintain our funding base. Although we have done so, successfully, over three decades, there have been periods when we were operating on a short time horizon – which is a challenge for staffing and planning.

Our independence also means that we are less able to draw on institutional supports (e.g., IT, web design, communications teams) that are available to other data projects (at, e.g., OECD, WB, UNU-WIDER, university-based entities).
7. Data acquisition dilemma:
Balancing pursuit of “priority countries” with ease of access

We have grown from six to 50 countries. Adding new countries has required complex decision-making.

At times, we prioritize bringing in countries because they are large, especially interesting, and/or in demand by data users. Sometimes we pursue countries to fill out a region (e.g., Latin America, East Asia) or other analytical grouping (e.g., BRICs). We also prioritize countries that have participated in the past (e.g., Sweden, Belgium, Romania), then drifted away or declined further participation. This targeted acquisition work can be extremely labor-intensive. Some cases have taken many years, sometimes without success (yet).

At other times, we have added datasets that came in largely by chance. E.g., a data provider wants to participate, a funding opportunity is linked to adding a specific country, acquisition is unusually easy (e.g., via unrestricted download). While we welcome country expansion in general, this can add a “randomness” to our dataset collections. (Additional randomness is added by the eclectic nature of the countries that have declined to participate).
How do we respond to those challenges?
LIS golden rules for harmonisation

- Set clear definitions for LIS variables
  - Maximise comparability by setting clear definitions for each variable (and trying to stick to them as much as possible)
  - Document very well any deviation from the general definition

- Complement easiness of use with flexibility of use
  - Enhance user-friendliness by providing fully standardised variables (standard variables, recodes, dummies, aggregate variables)
  - Allow users the flexibility to create other concepts by leaving a large amount of detailed information

- Adapt the LIS template to the changing environment (over time and space)
  - Template revisions
  - Revisions of previous datasets

Overall guiding principle: OPERATIONAL COMPARABILITY
Final remarks

• Sample coverage
• Individual level income (high non response affecting overall household income).
• Constructing upper level aggregated variables to help in the consistency checks
• Clearer documentation and definitions
Many thanks!

Data is a public good that should be non-exclusive and non-competitive

omar@Lisdatacenter.org