TWINNING CONTRACT JO/13/ENP/ST/23

Strengthening the capabilities of the Department of Statistics in Jordan

Microdata integration and schema reconciliation

Leonardo Tininini
ISTAT

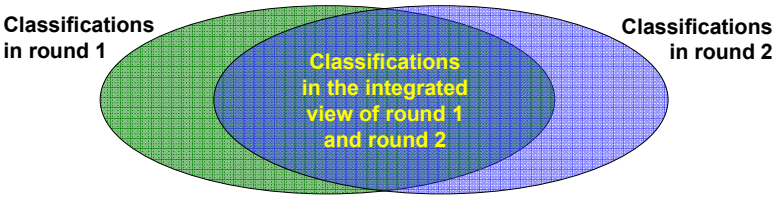
Amman, March 2014

Outline

- Integrating microdata from different rounds of the same survey
  - Naming issues
  - The role of the classification repository
  - Schema reconciliation
  - Managing distinct classifications for the same variable
- Integrating microdata from different surveys
  - sampling vs non-sampling surveys
  - managing missing/incoherent data

Integrating different rounds of the same survey

- Requires (at least)
  - **Same classifications** (dimensions) for the aggregate data or at least **same variable and corresponding classifications can be "reconciled"**



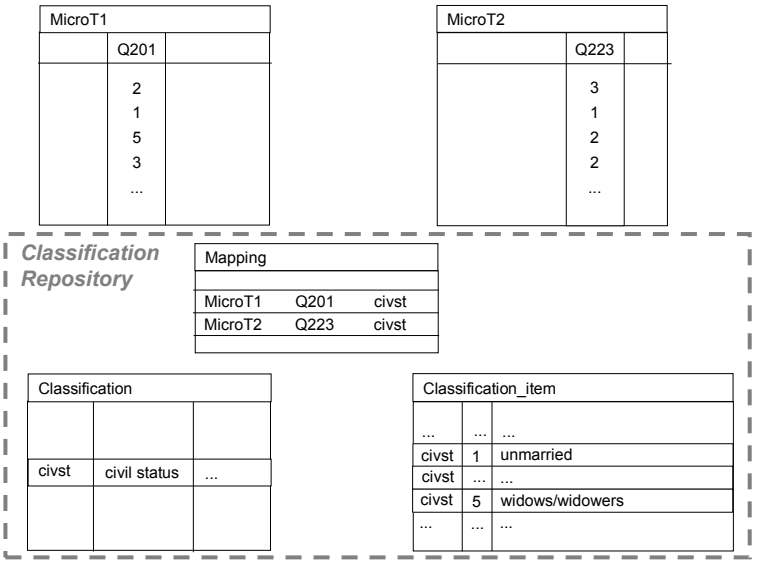
Naming issues

- In different surveys (or different rounds of the same surveys):
  - the **same statistical variable** may be stored in columns with **different names** (e.g. the "civil status" variable may be q221 in the 2010 round and q201 in the 2011 round)
  - the **same column name** may be used to store **different variables** (e.g. q201 may refer to the "civil status" variable in the 2010 round and to the "year of marriage" in the 2011 round)
- However, the fact that the same column name corresponds to the same statistical variable is not sufficient
  - the **same column name** may refer to **different classifications** (e.g. a column named "age" may refer to a 5-years classification in one case and to a 10-years classification in the other)
  - even if the classification is the same, the **codes used in the classification may differ** (e.g. 1 for "male" and 2 for "female" in one case, while M for "male" and F for "female" in the other)

The classification repository

- One of the components of the metadata repository
- Stores (at least) information regarding:
  - **classifications**:
    - code (necessarily unique)
    - name (preferably unique and possibly in different languages)
    - descriptions (possibly in different languages)
    - etc.
  - single **classification items**:
    - code (unique inside the corresponding classification)
    - name
    - descriptions
    - etc.
- In order to enable a semi-automated reconciliation of data the repository should also contain:
  - **mappings** from each **microdata table column** to the corresponding **classification** in the repository

Mapping columns to classifications in the repository



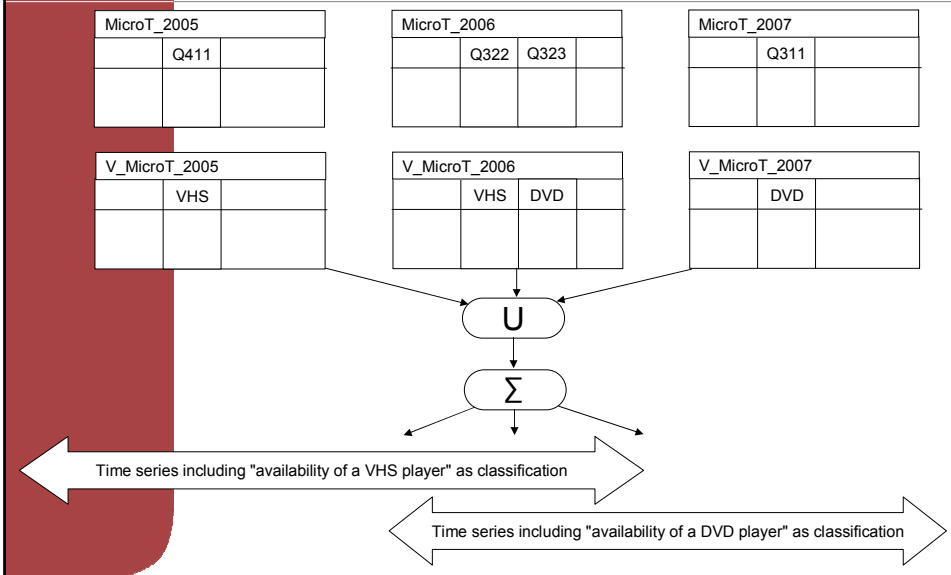
Creating reconciled views from mappings

- The **classifications** in the repository constitute a **common, shared "language"** enabling the different surveys to "talk" with each other
- The **mappings** represent the **"translations"** of the specific columns/variables of each survey/round in this common language
- Once the mappings have been determined and stored in the repository, **views can be (automatically) generated**, representing the translations of each table's contents, e.g.:

```
CREATE VIEW V_MicroT1 AS
  SELECT ..., Q201 AS civst, ...
  FROM MicroT1;
```

- The **views can be directly queried in a unified manner**, by taking the UNION of the several tables, based on the common columns and produce, for example, time-series

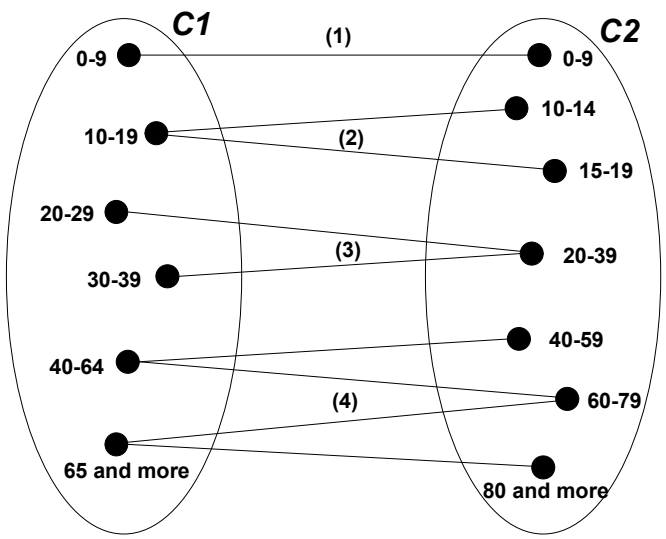
Querying reconciled schemas



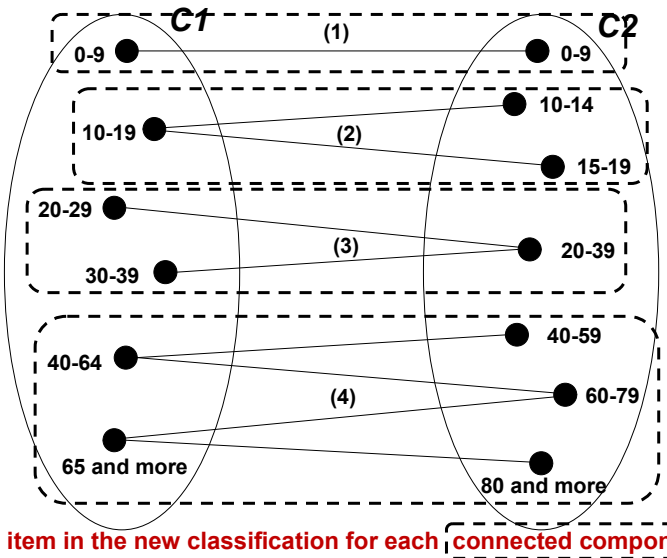
Managing distinct classifications for the same variable

- Given two **distinct classifications** C1 and C2 for the **same variable** V, we can try to reconcile them into a new classification CR, that will act as a kind of "common denominator" for C1 and C2
- **Fundamental pre-condition: statisticians should confirm** if grouping two or more classification items is feasible/meaningful
- **Possible combinations:**
  1) Item i of C1 is exactly cohincident with j of C2
  2) Item i of C1 exactly corresponds to 2 or more items j1, …, jM of C2
  3) 2 or more items i1, …, iN of C1 exactly corresponds to item j of C2
  4) 2 or more items i1, …, iN of C1 exactly corresponds to 2 or more items j1, …, jM of C2
- **Worst case**: only all items of C1 correspond to all items of C2 (hence no reconciliation is possible)

Defining the correspondences for single items



Determining the "connected components"



One item in the new classification for each connected component

The new classification CRec

- **One classification item for each connected component** of the link graph
- The details need to be **added** to the **classification repository**
- A **new column** has to be added to each microdata table and the corresponding values inserted according to the **mapping** between old and new classification:

CRec	
Code	Description
1	0-9
2	10-19
3	20-39
4	40 and more

MAP_C1_TO_CRec		
C1	Optional_desc	CR
1	0-9	1
2	10-19	2
3	20-29	3
4	30-39	3
5	40-64	4
6	65 and more	4

MAP_C2_TO_CRec		
C2	Optional_desc	CR
1	0-9	1
2	10-14	2
3	15-19	2
4	20-39	3
5	40-59	4
6	60-79	4
7	80 and more	4

Adding the column and values to the microdata table

MicroT1				
	C1			CR
	1			1
	4			3
	2			2
	6			4
	2			2
	...			...

Map_C1_To_CRec		
C1		CR
1		1
2		2
3		3
4		3
5		4
6		4

```
UPDATE
(SELECT mt1.CR, map.CR
FROM MicroT1 mt1, Map_C1_To_CRec map
WHERE mt1.C1 = map.C1)
SET mt1.CR = map.CR;
```


Integrating microdata from different surveys

- **Hard problem (to be analyzed by statisticians to verify the feasibility)**
- Generally requires (at least)
  - **Same time and territory of reference** (the data in the sources to be integrated need to refer to the same time and territories)
  - **The units of analysis must be the same** in the sources to be integrated
  - **The sources must share the same ID** for the units of analysis. If this is not the case some **matching algorithm** is required, but the reliability of the matching technique and of the inferred data has to be **verified by statisticians**
- In case of **sampling**
  - **Almost impossible to combine data**, unless the samples were specifically designed to do it. **Which weight** should be chosen? How can the **characteristics of the sample be maintained**, when integrating the variables from the several sources?

Integrating microdata from different surveys (2)

- Typically done by **integrating** microdata from **different registers sharing the same ID** (e.g. a SSN or a Taxpayer Identification Number)
- There will be (almost certainly) **unmatched units**
  - Units in S1 that have no counterpart in S2 and units in S2 that have no counterpart in S1
  - Adding all unmatched units (from both sources) will almost certainly produce **overestimation**, while discarding all unmatched units will almost certainly produce **underestimation**
  - Often one of the two sources is considered the **"master"** (most authoritative one). Consequently, all its units are kept, even the unmatched ones, while the unmatched units of the other source are discarded
  - The unmatched units of the master source have missing values and require some kind of **statistical imputation**
- There may be **mismatches on common variables**
  - **Different, incoherent values** of the same variable for the same unit
  - Some kind of **statistical imputation** is required (and statisticians should choose the specific technique to be used)

Integrating microdata from different surveys (3)