TWINNING CONTRACT JO/13/ENP/ST/23

Strengthening the capabilities of the Department of Statistics in Jordan

Microdata integration and schema reconciliation: some practical examples

Leonardo Tininini
ISTAT

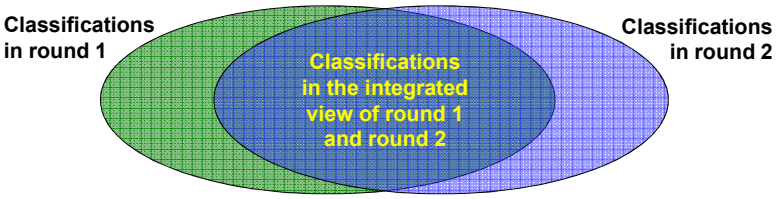
Amman, August 2014

Outline

- Integrating microdata from different rounds of the same survey (typically to produce time-series)
  - Naming issues
  - The role of the classification repository
  - Schema reconciliation
  - Managing distinct classifications for the same variable

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"**



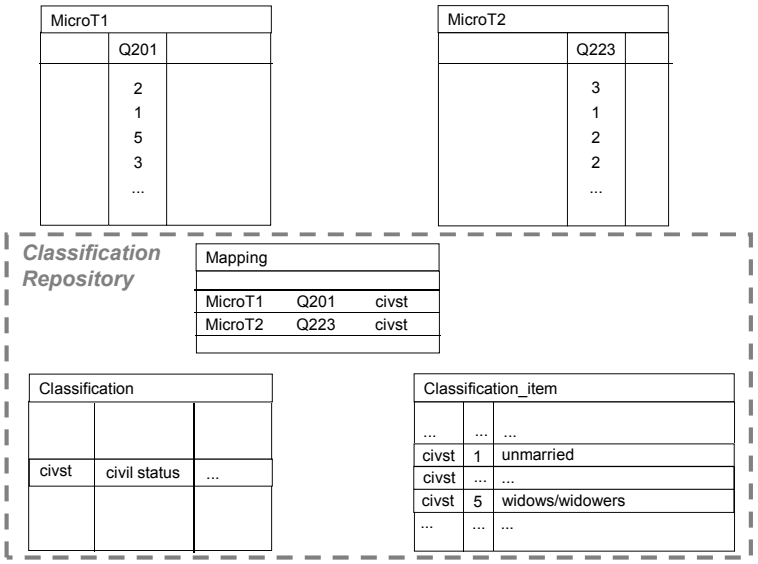
Naming issues

- 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



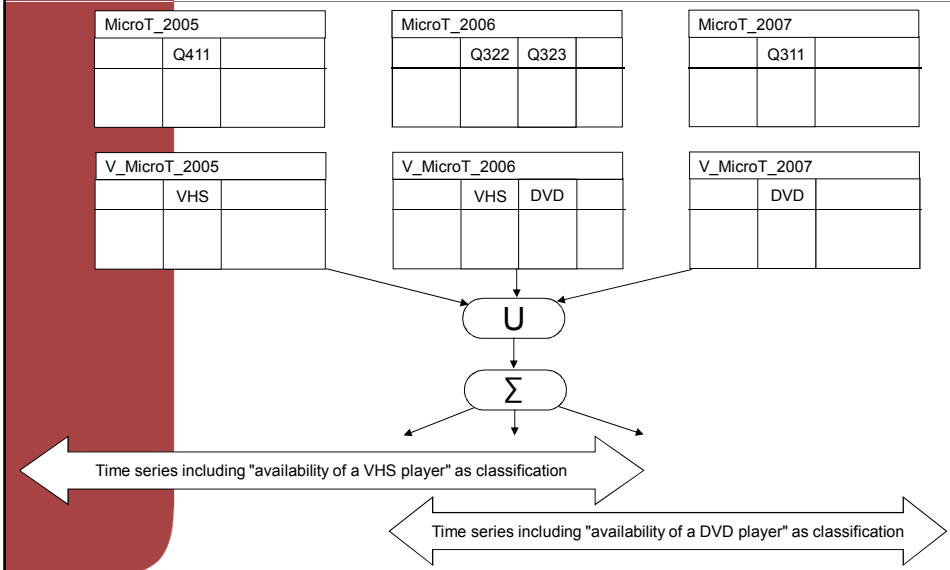
Creating reconciled views from mappings

- 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



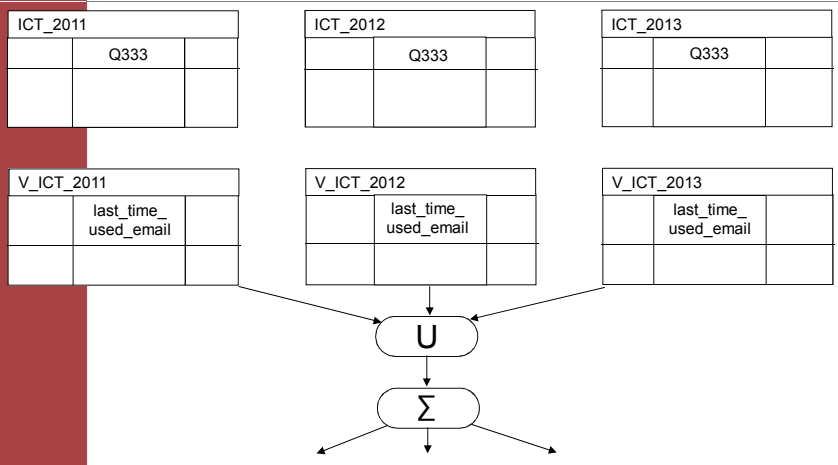
Q333 example - the repository data

Classification			
ref_code	code	description	notes
1001	last_time_used_email	The last time the person used his E-mail	

Classification_item			
ref_code	org_item_index	item_description	notes
1001	1	Today	
1001	2	Before Sveral Days	
1001	3	Before One Week	
1001	4	Before One Month	
1001	5	More than One Month	
1001	6	Don't know	

Mapping						
No_study	Type_table	Table_name	Round	Year	Real_name	Ref_code
1	1	TECPRSN	99	2011	Q333	1001
1	1	TECPRSN	99	2012	Q333	1001
1	1	TECPRSN	99	2013	Q333	1001

Q333 example - views and aggregations



We can group by last_time_used_email and aggregate the data on all 3 years

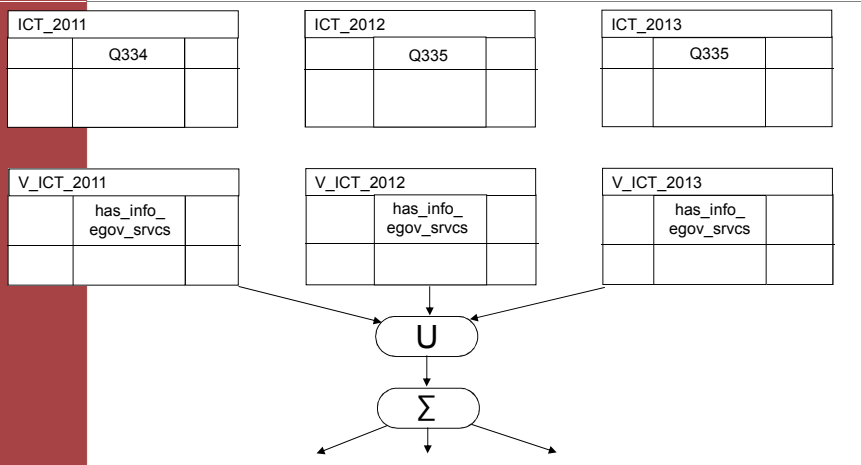
Q334/Q335 example - the repository data

Classification			
ref_code	code	description	notes
1002	has_info_egov_srvcs	The person has info on the e-gov services	

Classification item			
ref_code	org_item_index	item_description	notes
1002	1	Yes	
1002	2	No	
1002	3	Don't know	

Mapping						
No_study	Type_table	Table_name	Round	Year	Real_name	Ref_code
1	1	TECPRSN	99	2011	Q334	1002
1	1	TECPRSN	99	2012	Q335	1002
1	1	TECPRSN	99	2013	Q335	1002

Q334/Q335 example - views and aggregations



(Although the original column names are different)
we can group by has_info_egov_srvcs and aggregate the data on all 3 years

Q336 example - the repository data

Classification

ref_code	code	description	notes
1003	means_hear_egov	The means that the person has heared of the E-gov	
1004	means_hear_egov2	The means that the person has heared of the E-gov	

Classification item

ref_code	org_item_index	item_description	notes
1003	1	Audio- visual media	
1003	2	Internet	
1003	3	Mobile phone	
1003	4	Jordan knowledge Stations	
1003	5	Friends and Relatives	
1003	6	Other	
1004	1	Audio- visual media	
1004	2	Internet	
1004	3	Mobile phone	
1004	4	Jordan knowledge Stations	
1004	5	Mobile phone (MobileAPP, New APP, Information APP)	
1004	6	Friends and Relatives	
1004	7	Other	

Mapping

No_study	Type_table	Table_name	Round	Year	Real_name	Ref_code
1	1	TECPRSN	99	2012	Q336	1003
1	1	TECPRSN	99	2013	Q336	1004

Q336 example - views and aggregations

ICT_2012

	Q336	

ICT_2013

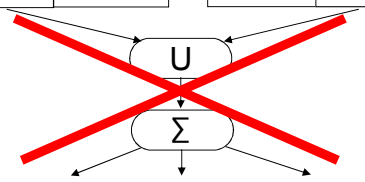
	Q336	

V_ICT_2012

	means_hear_egov	

V_ICT_2013

	means_hear_egov2	

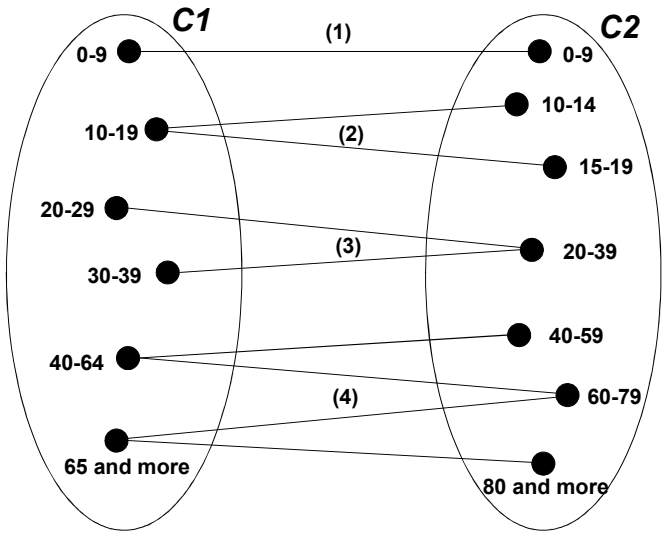


(Although the original column names are the same)
we can not combine the data in the two years (at least as they are…)

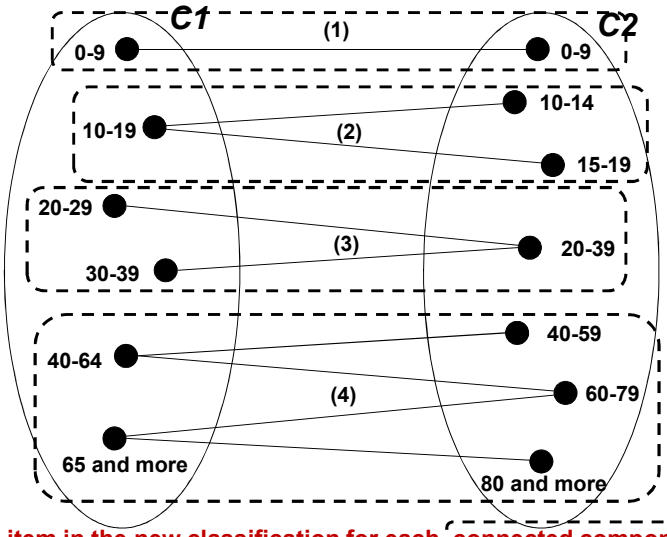
Managing distinct classifications for the same variable

- 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

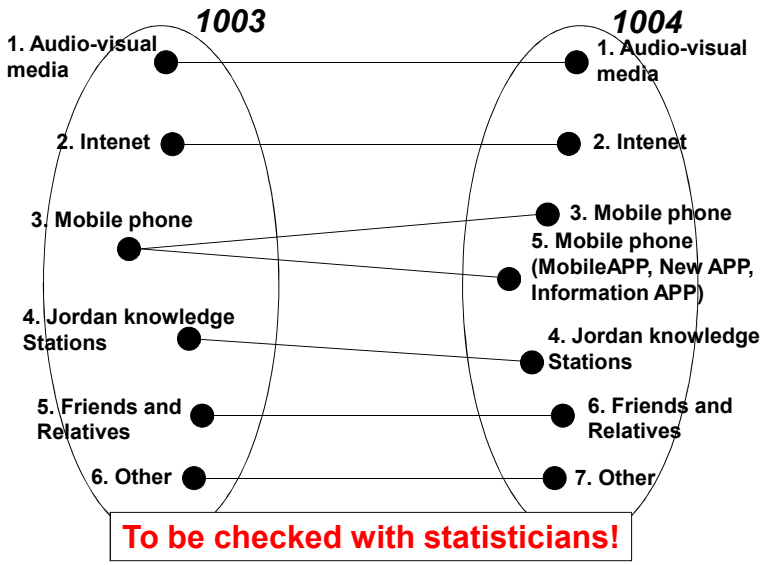
Adding the column and values to the microdata table

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

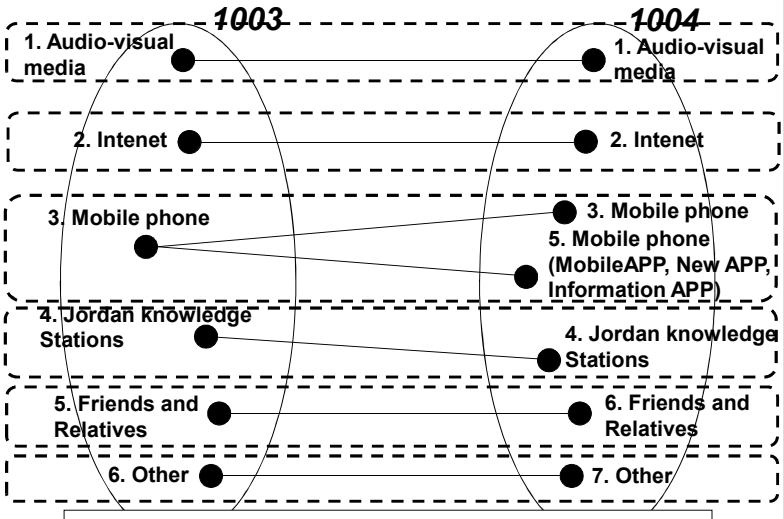
```
UPDATE
(SELECT mt1.CR, map.CR
FROM MicroT1 mt1, Map_C1_To_CRec map
WHERE mt1.C1 = map.C1)
SET mt1.CR = map.CR;
```


Q336 (2 versions) - Defining the correspondences



To be checked with statisticians!

Defining the "connected components"



In this case CRec is coincident with 1003

Adding new col and mapping 1004 to CRec (1003)

Item_mapping

ref_code	org_item_index	mapped_ref_code	mapped_org_item_index
1004	1	1003	1
1004	2	1003	2
1004	3	1003	3
1004	4	1003	4
1004	5	1003	3
1004	6	1003	5
1004	7	1003	6

ICT_2013

	Q336		N336
	3		3
	7		6
	2		2
	5		3
	6		5
	2		2
	…		…

```
UPDATE ICT_2013 mt SET N336 =
  (SELECT map.mapped_org_item_index
   FROM Item_mapping map
   WHERE map.ref_code = 1004
   AND map.mapped_ref_code = 1003
   AND mt.Q336 = map.org_item_index)
WHERE EXISTS /* optional */
  (SELECT map.mapped_org_item_index
   FROM Item_mapping map
   WHERE map.ref_code = 1004
   AND map.mapped_ref_code = 1003
   AND mt.Q336 = map.org_item_index);
```

Q336 example (revisited)

Classification

ref_code	code	description	notes
1003	means_hear_egov	The means that the person has heared of the E-gov	
1004	means_hear_egov2	The means that the person has heared of the E-gov	

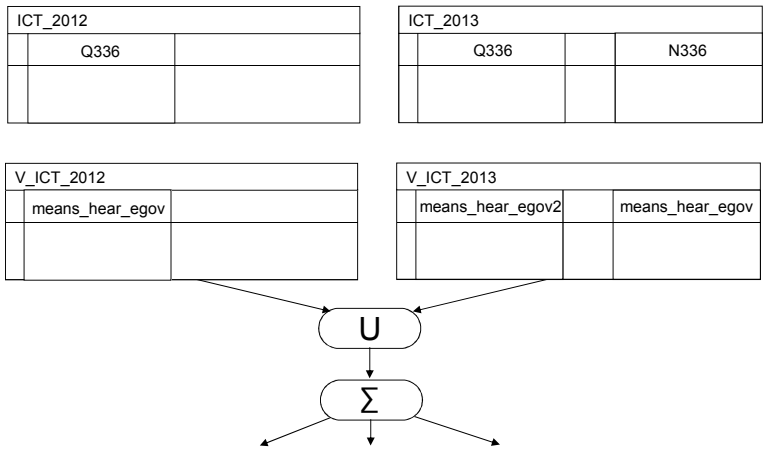
Classification item

ref_code	org_item_index	item_description	notes
1003	1	Audio- visual media	
1003	2	Internet	
1003	3	Mobile phone	
1003	4	Jordan knowledge Stations	
1003	5	Friends and Relatives	
1003	6	Other	
1004	1	Audio- visual media	
1004	2	Internet	
1004	3	Mobile phone	
1004	4	Jordan knowledge Stations	
1004	5	Mobile phone (MobileAPP, New APP, Information APP)	
1004	6	Friends and Relatives	
1004	7	Other	

Mapping

No_study	Type_table	Table_name	Round	Year	Real_name	Ref_code
1	1	TECPRSN	99	2012	Q336	1003
1	1	TECPRSN	99	2013	Q336	1004
1	1	TECPRSN	99	2013	N336	1003

Q336 example (revisited) - views and aggregations



Using the new column and the mapping defined, we can now group by means_hear_egov and aggregate the data on the two years

Q405 in 2011/2012/2013 - a tentative solution

Connected components		CRec
2011	**2012 and 2013**	**2011, 2012 and 2013**
1. Connecting Through Pre-paid Cards Dial –Up 4. Connecting Through Dail –Up Connection	1. Connecting Through Dial –Up	1. Connecting Through Dial –Up
2. Connecting Through Pre-paid Cards ADSL 5. Connecting Through ADSL Line	2A. Connecting Through ADSL Less Than 256 KBS 2B. Connecting Through ADSL More Than 256 KBS	2. Connecting Through ADSL
3. Connecting Through Pre-paid Cards Wireless Internet (Wimax) 6. Connecting Through Wireless Internet (Wimax)	3. Connecting Through (Wimax)	3. Connecting Through (Wimax)
8. Connecting Through Mobile, WAP, GPRS, etc…	5. Connecting Through Mobile, WAP, GPRS, etc	5. Connecting Through Mobile, WAP, GPRS, etc
7. Internet Service with Some one 9. Other (Specify)	4. Cable – TV 6. Connecting Through Mobile Broadband 7. Other	Other

To be checked with statisticians!

Q405 in 2011/2012/2013 - a 2nd possible solution

Connected components		CRec
2011	**2012 and 2013**	**2011, 2012 and 2013**
1. Connecting Through Pre-paid Cards Dial –Up 4. Connecting Through Dail –Up Connection	1. Connecting Through Dial –Up	1. Connecting Through Dial –Up
2. Connecting Through Pre-paid Cards ADSL 5. Connecting Through ADSL Line	2A. Connecting Through ADSL Less Than 256 KBS 2B. Connecting Through ADSL More Than 256 KBS	2. Connecting Through ADSL
3. Connecting Through Pre-paid Cards Wireless Internet (Wimax) 6. Connecting Through Wireless Internet (Wimax)	3. Connecting Through (Wimax)	3. Connecting Through (Wimax)
8. Connecting Through Mobile, WAP, GPRS, etc…	5. Connecting Through Mobile, WAP, GPRS, etc	4. Connecting Through Mobile, WAP, GPRS, etc
	6. Connecting Through Mobile Broadband	5. Connecting Through Mobile Broadband
7. Internet Service with Some one 9. Other (Specify)	4. Cable – TV 7. Other	6. Other

To be checked with statisticians!

No values for item 5 of CRec in 2011

Q405 - if the question is "single-response"…

- Define the **mappings** from the classifications used in 2011, 2012 and 2013 **to the reconciled classification CRec**
- **Add a new column** in the microdata table (see example before)
- **Insert the proper codes** in the new column (see example before)
- Group and aggregate data **classified by CRec** in the several years

Q405 - if the question is multi-response…

- **The solution is more complex** and depends on how the data are stored in the microdata table
- Generally, there will be **one column for each classification item**, e.g. nine columns for 2011, named Q405_1, Q405_2, … , Q405_9, each of them with only two possible values therein (typically Y and N or an equivalent numeric value)
- **For each item of CRec a new distinct column** will be required, e.g. N405_1, … , N405_5
- The values in each new column are obtained by **taking the OR of the values** in the mapped columns, e.g.

```
UPDATE ICT_2011
SET N405_1 = 'Y'
WHERE Q405_1 = 'Y'
OR Q405_4 = 'Y'
```

- The UPDATE instructions could be **manually written** or **automatically generated** and executed by dynamic SQL

Q336 example (revisited for multi-response)

Classification			
ref_code	code	description	notes
1003	means_hear_egov	The means that the person has heared of the E-gov	
1004	means_hear_egov2	The means that the person has heared of the E-gov	

Classification item			
ref_code	org_item_index	item_description	notes
1003	1	Audio- visual media	
1003	2	Internet	
…	…	…	
1003	6	Other	
1004	1	Audio- visual media	
…	…	…	
1004	7	Other	

Mapping (new version with one added column)							
No_study	Type_table	Table_name	Round	Year	Real_name	Ref_code	**Org_item_index**
1	1	TECPRSN	99	2012	Q333	1001	**(null)**
1	1	TECPRSN	99	2012	Q336_1	1003	**1**
…	…	…	…	…	…	…	**…**
1	1	TECPRSN	99	2012	Q336_6	1003	**6**
1	1	TECPRSN	99	2013	Q336_1	1004	**1**
…	…	…	…	…	…	…	**…**
1	1	TECPRSN	99	2013	Q336_7	1004	**7**
1	1	TECPRSN	99	2013	N336_1	1003	**1**
…	…	…	…	…	…	…	**…**
1	1	TECPRSN	99	2013	N336_6	1003	**6**

Further (possibly useful) fields

- In general the tables on classifications and items may have additional fields for the **codes and descriptions to be used for dissemination**
- The **classification table** may contain some additional fields
  - the corresponding **value set/variable** in PX-Web
  - the **code** for the **"total" value** to be used in PX-Web
- The **classification item table** may contain an additional field for the **codes to be used for dissemination** (that may differ from microdata)
  - **Example**: the item "15-19 years" may have the code "4" in the microdata table (and also classification item table) and the code "A1519" in the dissemination database
- …