UNITED NATIONS **E**

Economic and Social Council

Distr.
General

ECE/CES/2006/SP/17
7 June 2006

ENGLISH ONLY

ECONOMIC COMMISSION FOR EUROPE **STATISTICAL COMMISSION**

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-fourth plenary session
Paris, 13-15 June 2006
Item 6 of the provisional agenda

SEMINAR ON POPULATION AND HOUSING CENSUSES
SESSION III

Historical data harmonization – key task for research[1]

Submitted by the National Statistical Institute of Romania

1. The National Statistical Institute of Romania (NSI) has taken a new approach to Population and Housing Censuses (PHC) microdata management as a result of increasing demand from the public at large and the internal necessity to organize the microdata and in-depth analysis of census data over the years.

2. Census data is an important element in the improvement of the data dissemination system of the NSI as an ongoing task of priority.

3. The census data, a national resource in all countries, is often underutilized. The reasons for this may be: restricted access to microdata, unavailability of some data sets, formats, limited set of pre-prepared tables, non-harmonized data, costs of answering demands (response time, staff, pricing policy) and confidentiality.

4. The NSI decided to minimize internal efforts and the cost of data dissemination, increasing the capacity and quality of census dissemination in the same time.

[1] This paper has been prepared by Victor Dinculescu, National Institute of Statistics of Romania.

5. The first step has been made to achieve the general goal of all statistical institutes – separating the production from the dissemination process, and minimizing the contact with users in the data dissemination process. The aim of this process was to “Give more at the lowest possible costs”.

6. Micro data
Inventory of micro data:
- 93% recovered for 1977 PHC
- 100% for 1992, 2002 PHC

Problems, various:
- record-types and structures
- files formats
- number of variables and definitions
- number of derived variables.

7. Classifications
a) Geography – administrative structure changed over the censuses
b) Occupations (national, ISCO)
c) Activities (national, ISIC,NACE)
d) Other minor one-level classifications

8. Tabulated data
a) paper 1977 PHC
b) electronic 1992, 2002 PHC

9. Documentation
Completion of METADATA:
a) English translation
b) Content:
- questionnaires, forms
- manuals, methodology
- definitions
- classifications, nomenclatures
- translation tables

10. Harmonization
- Translation tables
- Recoding of data
- Fields rename
- Label values rename
- Add of extra sort-codes - language specific
- Rebuilding of derived variables

11. Database
A query system was designed – Oracle DB – web access.
First stage limited access in Intranet.
Table 1 - Micro data for 1992 and 2002 PHCs.

Millions of records		1977	1992	2002
Persons	Records	**21.5**	**22.8**	**21**
	Variables	72	37	61
Households	Records	**6.8**	**7.3**	**7.3**
	Variables	60	16	19
Dwellings	Records	**6.4**	**7.7**	**8.1**
	Variables	54	29	39
Buildings	Records	**4.7**	**4.5**	**4.8**
	Variables	41	26	37

12. As for 1977 PHC, some archives of microdata failed to be transferred from the old nine-track reels, ICL format – the process of recovery of 1977 PHC micro data was a challenge for NSI.

13. The main figures of 1977 PHC micro data recovery:
- Romania districts in 1977: 40
- District's files existing in tapes: 38

14. Recovered micro data:
- for two missing districts (921,654 persons) and missing data for incomplete districts the data files(tapes) were not available - micro data can be considered forever lost;
- recovery process was 100% on available files;
- the constructed dataset covers 97.23% for 38 recovered districts and 93.1% for Romania.

Table 2

	Persons in files	Published persons	% recovered	% of Romania's population
Romania	20,071,693	21,559,910	93.097	-
35 districts	18,608,341	18,608,528	99.999	86.311
Bacau	547,484	667,791	81.984	2.539
Galati	543,142	581,561	93.993	2.519
Ilfov	372,726	780,376	47.762	1.729

15. Database content:

- 5.7 billion original cells
- 10 billion harmonized cells

16. Datasets
    (a) 100 % micro data
        - Non harmonized 1992
        - Non harmonized 2002
        - Harmonized 1992, 2002 – 80%
    (b) 10 % micro data files sample – next step
        - Non harmonized 1977
        - Non harmonized 1992
        - 2002

17. Confidentiality
    - No personal IDs
    - No names
    - No addresses
    - Several levels of access to the data
    - Users classification
    - Reduced set of variables

18. For sample files- 10%
    - Swap across geographic units
    - Merge small variable categories
    - Recode sensitive numeric variables

19. Users
    - Policy
    - Registered users
    - Costs involved
    - Users classification

20. Dissemination
    - Web-based Query System
    - All cross-tabulation, all variables and levels of aggregation (depending on user rights)
    - Limited size request result
        - Large tables to be produced by NSI and downloaded through FTP
        - Price policy
    - Various output format
    - Free of charge or not
    - 10 % sample files for free download for registered users
    - License policy for resellers
    - Metadata

21. The main results:
    - a full functioning system to be used by dissemination staff;
    - improving of time response and quality of data;
    - unlimited possibility to aggregate the data just on-clicks;
    - time series queries.

22. Further adds and improvements
    - 1977 PHC micro data
    - 10% samples datasets for researchers- free download
    - metadata
    - E-Commerce.

* * * * *