United Nations

Working paper 5

# Economic and Social Council

Distr.: General
18 April 2012

English only

## Economic Commission for Europe

Conference of European Statisticians

**UNECE-Eurostat Expert Group Meeting on Censuses Using Registers**

Geneva, 22-23 May 2012
**Session 1: Experiences with the use of registers in the censuses**

# Model of transformation of administrative data to statistical data

**Note by the Central Statistical Office, Poland**[1]

## I. Introduction

1. Data sets transformations from administrative sources to statistics data sets aim to improve the quality of data sets from external sources and to increase their usefulness in the statistics.

2. The aim of the works on administrative data sets is to obtain a sufficiently complete dataset in terms of subjective and objective completeness, corresponding to classification standards, definitions and basic categories, and thus the effective use of administrative data.

3. Transformation involves downloading source data to the production environment and then controlling and correcting them. Using the solutions, knowledge and experience gained during the work on Agricultural Census in 2010 and the Population and Housing Census 2011 in Poland, a model of transformation collections from administrative sources to statistical register has been created.

4. Data sets transformations from administrative sources to statistics data sets within the scope of identification and address include: personal identifier (PESEL), tax identifier (NIP), business register (REGON), names, surnames, address details – country, voivodeship, powiats, municipalities, cities, streets and numbers of real estates. Cleansing PESEL, NIP, REGON identifiers requires validation of the length of identifiers and a check digit.

[1] Prepared by Pawel Murawski.

5. Obtained data can be poor or of very good quality, so it depends on evaluation whether the register can be used for statistical purpose. Analysis of the quality of source data from administrative records shall be based on measures of quality.

- Measuring the quality of administrative registers:
  - timeliness of data
  - methological compatibility
  - completeness
  - identification standards used in the register
  - usefulness
  - compatibility of data in administrative sources with data obtained in the survey.
- Measuring the quality in processing of data registers:
  - excessive coverage error rate
  - incomplete coverage error rate – subjective indicator of completeness
  - objective indicator of completeness
  - imputation rate
  - data correction index
  - integration data from various sources index.

# II. Stages of transformation registers

6. Process of transformation of administrative data to statistical data begins with importing the data that goes into the production environment, where by means of appropriate mechanisms, datasets are transformed, and then integrated. Statistical register prepared in this way is sent to the analytical environment.

7. The first stage of work on datasets is to extract them and put into the production environment based on software from SAS Institute.

8. Obtained data has various format, for example txt,. xls, .csv, xml, MS SQL databases. Import means to consolidate data from various source systems and convert data into format that is suitable for processing - SAS tables.

9. An integral part of import is to check the correctness of the data and its structures. These include in particular, the number of imported records (if agrees with the number of records submitted by the provider of information) and verify the correctness of assignment of data to individual columns (that means checking if text values contain the text, if the length of the field is suitable for data variables, etc.)

10. Extract process involves loading not all, but only the needed and selected data. For this reason, during the preparations for the Agricultural Census 2010 and the National Census 2011 methodological work on the diagnosis of contents and quality of public registers was carried out. Finally - for censuses - 29 sources were selected and obtained from public administration and outside public information systems.

11. Data transformation means a series of activities in the production environment consisting of: profiling - the creation of a report on data quality, unification data, parsing (separation) or combining variables, standardize with schemes, conversion, validation, deduplication, data integration.

12. Transform data – profiling: when dataset is successfully extracted, a process called profiling take place. We create a report of the quality of data, so we can check (at the level of numerical and percentage) the rate of errors for each variable in the set. In profiling, we can obtain information about the number of completed records, the number of unique entries, patterns and incorrect data.

13. Transform data – standardization and parsing: data standardization – unification and brought down to a defined standard – values occurring in certain columns. Parsing is a separation of variables - for example, the division of one column 'address' on columns: 'street', 'town', 'home number' or partial name and surname from one text field.

14. Transform data – conversion: is also a part of transformation process. Conversion means replacing values stored in different ways for the same information, encoding information.

15. Transform data – validation: is a process of checking the correctness of data and correcting abnormal values according to the algorithms prepared by methodologists. Sometimes it is also necessary to exclude from further processing records, which improvement is impossible. Through this process we are able to obtain better quality of data. Validation is performed on the datasets already pre 'cleaned' in the previous stages of work.

16. Transform data – deduplication is the process of removing repeating units and merge the information in the same records. It requires a detailed analysis, often including legal acts analysis. It is individual for each register. As a result of deduplication – we obtain one unique record of all the possible and unique information.

17. Transform data - data integration is a process of selection the best, most current and correct value of several or a dozen of registers. It is a process to create a statistical record, which will be available for use by analysts.

18. Statistical register is transmission from the production area to the analytical environment. In this process it is important to use mechanisms for quick loading large amounts of data. In the analytical area further work on the data goes on, such as production of summary tables or generate reports

# V. Summary

19. The process of transformation data to statistical data is a complex process consisting of several stages, which are modified depending on the type of registry. Working in this framework requires the prior preparation of substantive and methodological. The proper transformation of the data is also necessary to in-depth analysis. Identify the different cases.

20. This process is essential if we want to use the data contained in administrative registers. The ongoing digitization of the administration, expansion of existing and creation of new databases gives possibility of wider use of this type of data in statistical work.