Database of demographic indicators for countries of the world and regions of Russia: new functionality

Eugeny Soroko,
Senior researcher, Institute of Demography,
National Research University Higher School of Economics

.

Abstract

The paper deals with 3 topics: 1) Overview of existing sources of information on population. It covers databases: Eurostat, Rosstat, UN World Population Prospects, World health organization, Statistics Sweden, Institut National d'Études Démographiques, Human Mortality Database, and many others. The sources are analyzed using more than twenty quality criteria – user interface, list of countries and indicators covered, age groups of population, data formats and precision, periodicity of updating, etc. 2) Description of the Database of demographic indicators for countries of the world and regions of Russia under development at the Institute of Demography of the National Research University Higher School of Economics: the principles of data collection from different sources, methods of formation of the data cubes, metadata for these arrays, the set of indicators currently covered, the reference files for coding, ... 3) New functionality of this database distinguishing it from another ones, examples of use for specific data queries and modes of work. They display several know-how options, including formation of query result from the several sources specified by the user, formation of data arrays of higher dimension for an indicator from the set of arrays of less dimension, checking the values of indicators and elimination of errors, recalculation of indicator's values when different units of measurement are used in different sources. All the options are performed by the Database "on-the-fly".

1. Existing databases and other sources of information on population

Currently the general public and scholars have access to a wide range of databases and other sources of information on population. Many of them have unique features in the breadth of indicators included, the list of countries and regions, by degree of details by age and time period covered, etc. However, they are rather different by the quality of these sources. One can easily find the indicator required, but with absent year, region, or age. If it is found successfully, the query result does not have the data format or precision the user requires. In order to aggregate the results of databases use of the type mentioned we suggest a set quality criteria for evaluation of sources. The overview covers main sources of demographic data at the international, regional, and national levels. Some of them are: Eurostat database (http://epp.eurostat.ec.europa.eu/portal/page/portal/population/introduction), UN world population prospects database (http://esa *esa.un.org/wpp/),* World Health Organization – Regional Office for Europe – European Health for All Database (http://data.euro.who.int/hfadb/), Institut National d'Études Démographiques – Population in Figures database (http://www.ined.fr/en/pop_figures/developed_countries/developed_countries_database/), The Human Mortality Database (http://www.mortality.org/), Statistics Sweden (http://www.scb.se/default____2154.aspx), and a series of many others.

The quality criteria for estimating the source of data include:

- Owner (publisher)
- Web address (URL)

- List of demographic indicators
- List of territories and regions (countries)
- Age groups (1-year, 5-year, or other)
- Formats of the data (SCV, HTML, PDF, XLS, TXT, etc.)
- Periodicity of data updating
- Precision of figures (number of decimal digits)
- Users' interface
- Opportunity of saving the web-address to the indicator
- Scientific transparency
- Remarks on the server work and access to the source
- Correspondence to another sources of similar indicators, etc.

The results of analysis of the sources in accordance to these quality criteria are organized in the form of tables at one page per each source.

2. The Database of demographic indicators for countries of the world and regions of Russia

At present the Database of demographic indicators for countries of the world and regions of Russia is being developed at the Institute of Demography of the National Research University Higher School of Economics. This work is performed due to the grant # 11-04-0039 of Academic Fund Program of NSU HSE.

Description of the Database at IDEM NSU HSE

The principles of population data collection from different sources include:

- demographic indicators aggregate the data from basic most reliable sources
- the Database is actually a collection of data arrays from these sources
- minimal portion of information in the Database is a data cube created for one indicator from one source at one moment
- each data cube is done as an Excel file with metadata specifying completely the source, dimensions, name of indicator, date, unit of measurement, distinguishing it from other arrays
- metadata are flexible and the list of descriptors may be extended if required
- all the values of categories are coded according to the extendable coding tables specific for each category
- a result query to the Database is performed "on-the-fly" according the indicator and categories selected by the user from the set of data arrays collected in it

- software for the Database work does not depend on the set of indicators, categories used, data cubes in the collection, sources of data and does not require changes in case of their updating and extending

In October 2012 the Database contains the following set of indicators: population structure by marital status, extramarital births, age-sex structure of migrants, age-specific fertility rates, age-specific death rates, marriages by age of groom and bride, citizenship of international migrants, dependency ratios, percentage of urban population, percentage of population by age groups, percentage of first marriages, percentage of working age population, natural increase, completed fertility, infant mortality rate, net migration rate, neonatal death rate, rate of natural increase, total fertility rate, median age of population, infant mortality, net reproduction rate, net migration rate, crude birth rate, crude death rate, crude marriage rate, crude divorce rate, life expectancy, population density, sex ratio, mid-year population, mean age of women at first marriage, number of abortions, number of deaths (some selected). Some indicators are for countries of the world, some – for the regions of Russia, some are available for both.

3. Some features of the Database

Is it clear that there is no reason to develop a database similar to the available well-known ones. The Database is equipped with several new options representing its know-how. Some of them are to be presented at the Conference.

One option is formation of query result in the case of diverse units of measurement used in different sources. The best typical example is population size. In some sources it is measured in persons, in other – in thousands, and in some – in millions. This situation can be solved within the Database itself by means of recalculation the indicator's values according to the unit in a specific data cube. Of course, the reference file of units should contain additional field of the recalculation rate to the "standard" unit.

New function is transition to the data array of higher dimension compared with the several source arrays of less dimension. Example is the number of migrants by country of origin, educational attainment, and cause of migration available in the source data arrays only for one specific period. Thus each of them is a 3-dimensional one. The Database gives the opportunity to create a 4-dimensional array of the indicator's time series.

Formation of the data query is performed "on-the-fly" from the collection of data cubes for a given indicator. The result of query along with the values contains the sources of the data used. If the user does not trust some of them there is an opportunity to select only the reliable ones. In this case the values will be taken only from the sources selected.

One cannot trust a database that gives the values contradicting demography and statistics. For example TFR 48 or crude marriage rate 333 per thousand per year. What is the way to guarantee the absence of such errors? One of possible solutions is to check the values at the stage of query result formation. The Database performs this task and provides elimination of erroneous values "on-the-fly".