UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

STATISTICAL OFFICE OF THE EUROPEAN UNION

Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses
(The Hague, The Netherlands, 10-11 May 2010)

Working paper 18
4 May 2010

Topic (vii) of the provisional agenda

REGISTERS: THE MAIN SUPPLIER OF TOMORROWS' RESEARCH WAREHOUSES?

Census data warehouse - use of census data

Note by the Statistics Finland

I. USER DEMANDS OF THE DATABASE

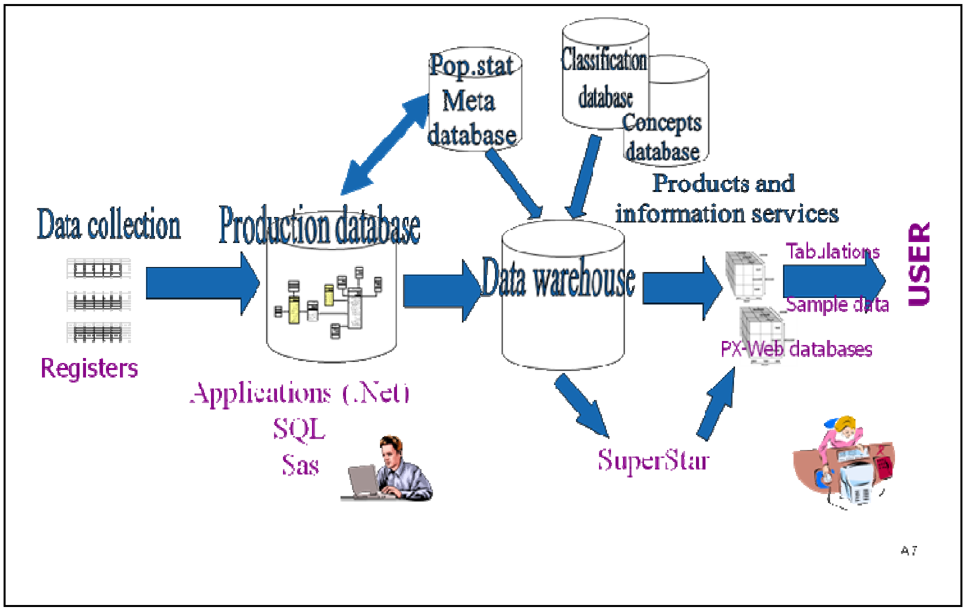
1. Statistics Finland decided to discard the mainframe environment and move all statistical production into an open environment. Annual census-related statistics on population structure, families, household-dwelling units, dwellings, buildings, and employment were the largest users of the mainframe environment and the renewal of their production was a major effort that took several years. The project started in 2002 and finished in 2007. Production started in the new environment in 2007 with the production of statistics for year-end 2006. The demands on the new system included the requirement of close connection with education statistics and geographic information system (GIS) processes.

2. The challenge in the new production environment was to create a system that could satisfy the needs of both statistical production and the use of statistical data for dissemination and research purposes.

II. DATA WAREHOUSE AS PART OF THE PRODUCTION PROCESS

3. The solution to the many different needs was to have separate databases for production and dissemination purposes. The databases are built on an MS SQL server. For production the database is organised so as to make the editing, imputing and compiling of data as effective as possible. For dissemination purposes the project decided to build a new data warehouse. The intention in the separation of production and dissemination was to make production more effective and avoid data redundancy and transfer. Instead of each set of statistic having its own data files, all data related to register-based population statistics could be updated and found in one database. For example, area codes for different years and sets of statistics could be updated in one table that could be used for all data in the data warehouse.

Figure: Data warehouse and the production process of census statistics



III. BUILDING UP THE DATA WAREHOUSE

4. When the data warehouse for census and other register-based statistics was being built at the population statistics department the main focus was on putting together all former, separate annual data files into tables containing the basic subjects and on building on the basis of content rather than the organisation of production. For example, one table was built for annual data on persons which contained all data related to persons instead of having separate files for population structure and employment /census as in the past.

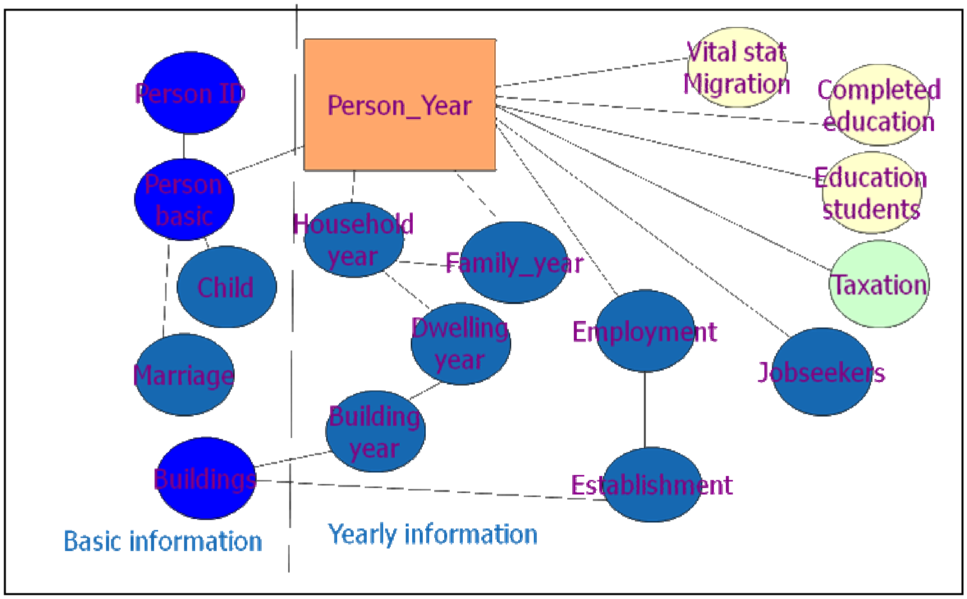
5. Data not published in any statistics but used in production have been mainly used for research purposes. For example, researchers are interested in data on all employment and unemployment periods during each year even when they are not defined like those used in employment statistics. To ensure all researchers would get the best possible quality from these kinds of data it was decided that they, too, would be added to the data warehouse.

6. Basic principles of the data warehouse at the population statistics department:

- The basic tables in the data warehouse contain data on persons and buildings
- The database contains all persons with a person ID code in any population statistics starting from the 1970 Census year
- All persons have an unchanging person number in addition to the person ID code in the Population Information System (PIS)
- Data on persons are divided into two different tables: basic information and annual data
- The table on buildings contains all buildings with co-ordinates (and geocoded co-ordinate points without buildings)
- The building code is also replaced with a surrogate key

- The table on buildings is updated with changes from the PIS
- Geographic information (GIS) will be integrated into the database (linkage to building-co-ordinates)

Figure 2. Structure of the census data warehouse database



IV. BENEFITS OF USING THE DATA WAREHOUSE

7. Since 2006 all statistical data published or disseminated from census-related population statistics are generated from the 'census data warehouse'. This has already simplified dissemination processes and the work still continues. In the 2010 Census we are trying to create a new system for updating the statistical databases. Our aim is to produce from the data warehouse just a couple of displays which would contain all the data needed for various products in web-based databases.

The users at Statistics Finland now use the data warehouse direct if they need it in their statistics, for example, for adding variables to their data, such as education or type of activity of persons, or for analysing non-response in sample surveys, etc.

Use of the data warehouse for producing products and services

8. Internet service: the free StatFin database service

Other (charged) Internet services

- Population data service
- Urban and regional indicators
- Rural indicators
- Transition from School to Work
- Statistics by postal code
- Grid Database

Special compilations
Publications
Tabulations (made at Statistics Finland)

- Time series tabulations

Microdata for research purposes

9. One of the principle arguments for the building of the data warehouse was the challenges we were facing in the use of data for research purposes. Annual production of census-related statistics started at end of the 1980s. Thus, we had in the mainframe environment about ten different statistical data files for each year, sometimes also separate files for products and services collected on various topics and years. Using these datasets we decided to compile longitudinal data files that would be updated with the latest data on persons each year. Year by year these huge data files were getting bigger and bigger but were still not comprehensive enough to satisfy all researchers' needs. They were updated after the publishing of annual data and were always out-of-date so that annual data had to be used anyway. Keeping abreast with changes in classifications also took far longer than researchers would have liked.

- Maintenance in one place
- Simplified process
- Better quality
- More possibilities
- Individual needs satisfied
- Efficient administration of large data volume.

*****