UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Workshop on Developing Data Dissemination Systems
(12 May 2008, Geneva, Switzerland)

ESTONIAN EXPERIENCE IN IMPLEMENTATION AND MAINTENANCE OF STATISTICAL DATABASE USING PX-WEB

Eda Fros
Statistics Estonia
2008

Contents

Introduction

The main task of a statistical organisation is to make figures accessible to users. Traditionally statistics are presented in the form of statistical tables. The number of accessible tables has increased enormously due to the Internet. Tables can be organised in different ways and formats. Any statistical table can be presented as a static or as a dynamic table depending on the purpose. Anyway, statistics should be available on screen and also downloadable to the user's computer for further use.

The Internet has become the primary channel for dissemination of statistical information and the dissemination database has become an integral part of statistical data web. Web services allow to apply different kinds of design and functionality in order to present statistics from either a technical or user perspective. The aim of this review is to provide a clear-cut example about how the statistical database is presented on the web site of Statistics Estonia. In the context of this review, statistical database is regarded not as a "real" relational database, but as a set of statistical tables intended for dissemination of statistics.

1 Estonian public databases

Two public databases — the Statistical Database and the Regional Development Database — are intended for publishing statistics. The Statistical Database is available on the Internet since April 2001. The Regional Development Database is available since April 2002. The aim of these public databases is to grant the users a fast and easy access to all published statistics and to allow users to download selected information in suitable formats.

The Statistical Database presents official statistics, including official regional statistics collected in the framework of the official statistical surveys approved by the Government of the Republic.

According to the strategy of regional development of Estonia, Statistics Estonia maintains also the Estonian Regional Development Database which presents official regional statistics and the data not collected in the conduct of official statistical surveys but received from administrative data sources.

The Statistical Database and the Regional Development Database are available on the web site under the heading **Statistics – Products** http://pub.stat.ee/px-web.2001/dialog/statfileri.asp.

Common principles

- The Statistical Database is meant for all official statistics published by Statistics Estonia
- All official statistics should be first disseminated in the Statistical Database
- In case a news release is issued, the data are published simultaneously in the news release and Statistical Database
- Through the Release Calendar, Statistics Estonia informs all users about new statistics being released in the databases
- Updates in databases are made public at 9.00 a.m. on the day announced in advance in the Release Calendar
- The Statistical Database and the Regional Development Database are both published in Estonian and English
- Both databases can be accessed free of charge

2 Software

For dissemination purposes, Statistics Estonia employs PX-Web originating from the PC-Axis software family developed in Sweden. PX-Web is a server-based database tool which enables dissemination of statistics in the form of dynamic tables on the Internet by using PC-Axis files. To be more exact, it is an asp- and dll-application that requires a Microsoft Internet Information Server. PX-Web can very easily be translated into any language and also customised to local layout rules.

As for the structure, the database consists of PC-Axis files, which are organised like any other files in Windows folder. The database can be divided into sub-areas/sub-folders. The PC-Axis file is an ASCII file consisting of keywords in English. It contains both data and metadata.

The selected table can be displayed on screen in two different layouts. The table can be downloaded into PC-Axis, Excel, CSV, XML, DBF, etc. file. For the PC-Axis format, the PC-Axis main module has to be installed and used. PC-AXIS is a program for Windows for working with PC-Axis files. PC-AXIS can be downloaded free of charge.

Why PC-Axis software?
In 1999–2000, various tools for browsing statistical tables on the Internet were simultaneously offered. Statistics Estonia tested all of them and decided to rely on PC-Axis software. The main advantages of PX-Web were as follows:

- We had used the PC-Axis main module (for Windows) since 1998 for presentation of statistical tables on CD-ROM
- Simple structure of PC-Axis file
- Easy to adapt and maintain the software
- Convenient to create, update and edit tables
- Good solution for the presentation of multidimensional tables
- Use of software in neighbouring countries (all Nordic countries)
- Existence of the PC-AXIS reference group to influence further development of the product

3 The structure of statistical table

To create a proper database table, the structure of statistical indicator and the structure of database table have to be understood as a prerequisite. There is no common generic model of a statistical indicator. Every statistical organisation has its own one, and the components have different labels. But it is always possible to create correlation between the components of different schemes. Persons responsible for database tables, need to understand the structure of statistical indicator and statistical table.

To study the structure of statistical indicator, Statistics Estonia has used two sources

- Alfa-beta-gamma-tau-structure described in
  "Guidelines for the modelling of statistical data and metadata". Conference of European Statisticians. Methodological material. United Nations, 1995
  "Information Systems architecture for national and international statistical offices". Guidelines and recommendations. Conference of European Statisticians. Statistical standards and studies – No. 51. United Nations, 1999
  (http://www.unece.org/stats/documents/information_systems_architecture/1.e.pdf)
  and in several documents from Statistics Sweden
- "Standardization of statistical indicators and metadata" by Central Statistical Office of Poland and the Czech Statistical Office. Statistical Commission and Economic Commission for Europe. Conference of European Statisticians. Work Session on Statistical Metadata. Geneva, 22-25 November 1994

A statistical indicator is a data element that represents statistical data for a specified time, place, and other characteristics.

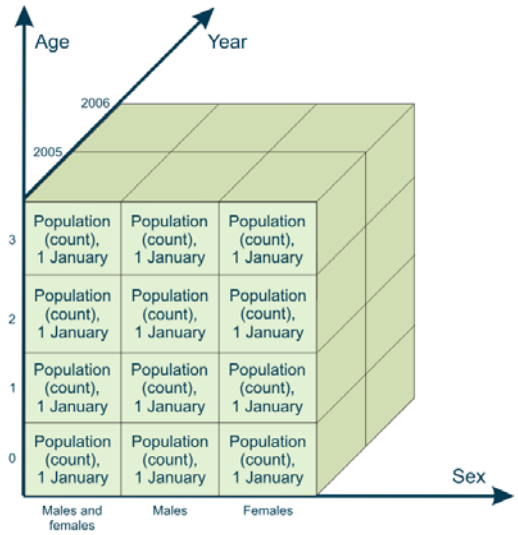
A variable is a characteristic of a unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, weight, etc. and 'occupation', 'industry', 'disease' etc).

Statistical indicator is expressed by using characteristics called variables. There are principal characteristics describing a socio-economic phenomenon or process and characteristics specifying or classifying principal characteristics. In the top of the hierarchical structure of statistical indicator there is a principal characteristic describing a socio-economic phenomenon or process, other characteristics are classifying this characteristic. Alongside with the statistical indicator, also a phenomenon or process characterised by this indicator should be described.

Traditionally, the statistical indicator is presented by using statistical table. Database table is a multidimensional matrix, the so called cube. Every dimension of a matrix determines one variable, specifying one characteristic in the table. Statistical indicator always refers to a certain time period. For this purpose time variable is required.

Example

For instance, to describe persons residing in Estonia, the statistical indicator **Population (count), 1 January by sex and age** can be used.



The values of variable **sex** are **males and females**, **males**, **females**. The values of variable **age** are **0**, **1**, **2**, **3**. The values of time variable **year** are **2005**, **2006**.
Every combination of values of variables determines one data cell in the table.
In case the table presents data for more than one time period, one dimension determines the time variable.

In case of Estonia there can be one or two time variables in the table (e.g. year and month with the respective values 2006, 2007, etc. and January, February, March, etc.). The alternative is to use only one time variable (e.g. variable time with values 200601, 200602, 200603, etc.).

One or more indicators can be presented in the table. In case there is more than one indicator presented in the table, one dimension/variable determines the principal characteristics. In our tables, this variable is called **indicator** to refer to principal characteristics. In fact, it is not correct, statistical indicator is a whole complex — the principal characteristic specified by other characteristics.

4 The structure and functionality of databases

The databases are organised in multidimensional matrixes, the so-called cubes. These cubes typically have 2 to 5 (in some cases up to 8) dimensions, which describe the observation variable called indicator (in case there is more than one indicator in the table), up to 2 time variables and classification variables. A lot of time series can be extracted from the cubes. Cubes are organised into hierarchically organized statistical areas.

Both databases are divided into four main statistical areas — the economy, environment, population and social life. The data of Population and Housing Census and Agricultural Census are presented separately. The statistical areas are divided into sub-areas according to Estonian classification of statistical domains. Every table of public databases has its own identification code. This coincides with the file name of the table. All tables of the subject area are arranged according to these file names.

The data search can be implemented in two possible ways

- Via the hierarchical structure
- Via text search based on PC-Axis files

A new function which involves the storage and usage of saved queries by registered and logged-in users is being tested and described for end-users at present.

The following steps should be taken to generate and design tabels. Users select values of the variables (the elements in the dimensions). Proceeding from these selections, the database returns a dynamically generated table. There is a feature to eliminate variables. If nothing is specified in one or more dimensions, a default selection (total) is used, which was specified by the database manager when creating the statistical table. The selected table can be presented on screen, in print and also downloaded to the user's computer in different formats. On screen, users can rearrange dimension variables in rows and columns, remove zero lines, sum values of the variable, sort values of the variable and save information in the cookie for further use. In addition, there is also a possibility to use graphical presentation.

5 Metadata

The display of statistical table also sets out the measurement unit, footnotes, date of update, etc.
Additional metadata to the table are presented under the heading Detailed Information. It consists of the following sections

- Terms and definitions
- Description of methodology used
- List of classifications used
  Sometimes explanations of additional groups of classification categories used in tables are provided.
- Other relevant information
- List of printed publications and links to other web sites
- Information about contact person

6 Management and maintenance of databases

Three structural units are in charge of public databases. Namely, the information technology (IT) department, methodology department, and marketing and dissemination department.

IT department is responsible for adaptation of new versions of software and for implementation of new features and functionality in co-operation with methodology department.

Marketing and dissemination department is responsible for managing databases — creating and updating tables and updating the database according to the Release Calendar, creating and updating the Release Calendar (in co-operation with subject matter units), and linguistic editing of new tables.

Methodology department is responsible for methodology and development of databases, implementation (in co-operation with IT department) of new versions and new features, compilation of instructions and other descriptions, harmonisation of new tables (in co-operation with marketing and dissemination department and subject matter units). Databases serve the purpose of harmonisation. Common databases containing data from various subject areas always require a high level of harmonisation with respect to the structure of the table and metadata (title, labels of rows and columns, etc.). To my mind this is why centralised management of databases is of extreme importance.

The majority of database tables (PC-Axis files) are converted from Excel tables by database managers. The Excel table for updating the database tabel is being prepared in subject matter departments. In most cases it contains data only for the last year, also in case of short-term statistics. The database manager converts the table into the PC-Axis format and by linking the existing table and a new table creates the updated table for database. Linking of tables enables to check the structure of the last presented table.

PC-Axis files of some subject areas are outputs of our internal macrodatabase. The internal macrodatabase contains very detailed aggregated data, also confidential data. This database is password-protected, each statistical unit has access to their own data only. The internal macrodatabase has not been launched properly for different reasons.

The PC-Axis tables of foreign trade statistics are being created historically in the foreign trade statistics unit.

7 Databases usage statistics and user feedback

The use of public databases is showing an increasing trend, but there are great differences depending on the month. The maximum number of retrievals is 100 thousand a month. Probably, new first-time users have mainly

contributed to the growth of database popularity up to date. Besides that, students comprise a large share of users.

Only some user surveys have been conducted with respect to database usage during recent years. One of the questions asked concerns the preferred format for downloading. The most popular format for downloading is Excel. The user survey conducted in 2003 dealt with the database user-friendliness. The goal was to evaluate user-friendliness and determine the need for end-user training. The results revealed that although users managed to use different features of the database, the majority still considered extra training useful. Also, some sort of user manual was deemed necessary — to provide instructions on how to use the text search engine, to provide explanations on buttons above the table on screen, etc.

We have previously introduced our databases at several events and to various target groups. No training courses have been conducted so far.

8 Developments regarding databases

Several implemented, ongoing and future developments and outputs of public databases can be pointed out.

Publications on CD-ROM
For years, the statistical database has been used as a basis for the creation and presentation of tables in electronic publications on CD-ROM. Certain procedures and special software have been worked out for the automatic creation of different formats of tables for CD-ROM.
This year we are considering to use the PX-Web CD version, allowing to significantly reduce the amount of working hours dedicated to CD-ROM compilation.

Release Calendar
To avoid duplication of any released information in different systems and to enable end-users to manage this information in a simple way, a decision has been made to store and manage release dates in a centralised manner. There is a dynamic Release Calendar on the web allowing users to select a subject area, period, type of object and type of presentation for Release Calendar. Regarding databases, the list of tables subject to updating for each object in the Calendar can be viewed by clicking on the name of object.
After the Release Calendar has already been published, all further changes will be provided with appropriate marks (added, delayed, corrected, cancelled) to inform users about changes. Release information is used under various headings of the web site.

The database table presentation standard
After a seven-year experience in the maintenance of databases, it is time to revise the current set of rules fixed for databases and to create the database table dissemination standard. It should cover not only mandatory keywords for the PC-Axis file, but also several rules regarding the title of table, variables, values of variables, etc. After the description of the database table presentation standard and once all the tables are in compliance with the standard, it will be possible to create a list of statistical indicators on the basis of statistical database to be published electronically.

Archive of database tables
It is not possible to manage different versions of database tables in the database. An archive system is needed to record all changes made to tables. Each version of table in the database should be stored in archive to allow to track all the inserted changes. This concept is in a planning phase for next year.

Predefined tables on the web
Our databases contain about 2.3 thousand tables, both in Estonian and English. For users who need only some general data, we are planning to set out some predefined tables in every subject area. Manual updating of these tables is very labour-intensive. To avoid it, a special application for describing the predefined tables based on database and automatic updating of tables are required. We have compiled and described the vision of this project, and depending on the availability of resources, we keep analysing it further.