

CONFERENCE OF EUROPEAN STATISTICIANS

**Joint UNECE/EUROSTAT Work Session on Methodological Issues Involving the Integration of
Statistics and Geography**
(Tallinn, Estonia, 25-28 September 2001)

Topic (ii): New technological solutions, including those based on online data access

FROM CENSUS MAPS TO THE CENTRAL SPATIAL DATABASE II

Submitted by the Statistical Office of Estonia ¹

Contributed paper

ABSTRACT

The Statistical Office of Estonia (SOE) has completed the mapping programme for the 2000 Population and Housing Census. The programme was begun in 1995 and the database was completed in 1999-2000. The delineation of enumerator areas and the printing of enumerator maps were based on this database. However, for the post-Census data processing and analysis the database structure had to be adjusted. The software for scanning and processing the Census questionnaires and an anonymous Census database were developed in Oracle relational database management system (RDBMS). Therefore, the decision to convert the existing GIS database (in Mapinfo and ArcView formats) into Oracle Spatial or into ArcSDE is under discussion. At the time of writing, the provisional database structure has been designed and data conversion will start after a new national tender has been carried out, hopefully in 2002.

I. BACKGROUND OF THE EXISTING SPATIAL DATA AND CENSUS INFORMATION SYSTEM

I.1 Description of Census mapping

1. SOE launched the mapping programme for the 2000 Population and Housing Census in 1995. After completing the test areas, the specification of digital Census maps was compiled. According to the specification, 1:50 000 maps in rural areas and 1:5 000 maps in urban areas were drawn. The specification was optimized to create a cartographic basis for the Census planning (Census area (CA) delineation) and the Census itself (maps for enumerators, maps for supervisors, etc.).

2. The Census mapping process was outsourced from SOE to two companies – one in an urban area, the other in a rural area. The production methodology was different in each case. In rural areas, paper maps of the 1989 Census digitized by the mapping company and updated by local governments were used as source material. In urban areas, the existing maps and orthophotos were used as a source and the maps were updated by the mapping company. For rural and urban areas the municipalities compiled household lists, including the number of inhabitants in each building or apartment. The household lists were to provide information about the number of inhabitants to delineate enumerator areas (EA).

¹ Prepared by Inge Nael, Külli Tiits and Andrus Tamboom.

3. SOE stores digital maps of urban areas in Mapinfo, of rural areas in ArcView software and household lists in Foxpro software. The Census maps were completed in December 1999 and were printed out in the 1st quarter of 2000.

4. The delineation of CAs was completed on 4 February 2000. In total, there were 5,323 enumerator areas, 995 Census districts (set of 4 to 6 EAs), and 165 supervisor areas (set of Census districts).

5. The printing of enumerator maps was begun in January 2000 and was completed in March 2000, one week prior to the critical moment of the Census. Four types of maps were printed:

Map type	Scale in urban areas	Scale in rural areas
Enumerator maps	1:3 000 to 1:5 000	1:5 000 to 1:50 000
Census district maps	1:3 000 to 1:22 000	1:10 000 to 1:100 000
Supervisor area maps	1:5 000 to 1:22 000	1:20 000 to 1:120 000
Wall maps for local Census offices	various scales	

For this task, two A3-size color laser printers were used by subcontractors, the third was kept in reserve at SOE as a backup for possible equipment breakdown.

6. In parallel with the enumerator area map printing, the household lists were cross-matched with the data of the Population Register and the Building Register. The selected columns from the resulting database were printed for each enumerator as auxiliary information to speed up the filling-in of Census questionnaires. The cross-matching was relatively labour consuming as, at the time, there was no reliable identifier system to build up the relationship between different registers. Automatically, only 1 to 70% of database rows were matched using addresses, the percentage being higher in urban areas. The relationship between registers was created by local governments using their expertise and information about local inhabitants.

7. As a result of mapping, SOE has created a data set of about 400,000 buildings of approximately 300 urban settlements and about 200 rural municipalities. Digital maps are associated with alphanumeric data – household lists, which in turn are associated with data from the Building Register and the Population Register. The data set is unique in Estonia in terms of accuracy, completeness, up-to-dateness and scope, and is worthy of being maintained in a better IT-environment than has been feasible so far.

8. The way in which the Census map database was processed until the completion of map production was not perhaps the most “high-tech” method, but was completely appropriate for the purpose. However, the following disadvantages may hamper further development:

- different software environments for storing the Census data and spatial data would be difficult to handle;
- in the case of paper maps the overlapping areas around urban areas were unavoidable, but in the spatial database it creates unnecessary duplication;
- in GIS data files it is difficult to ensure data consistency and security;
- the data split between a number of files and file formats is difficult to analyze and use for the generation of small-scale maps.

9. The full Census database (data on persons and households) in Oracle will be archived and stored in a highly secured archive. During the next stage, primary and secondary person identification information will be removed. As a result, an anonymous Census database will be created. The anonymous database will be processed by software dedicated to the generation of tabulations and providing open access to the database. The anonymous database can be used with Oracle Discoverer and a number of other tools, including GIS software.

10. The link between GIS and the anonymous database processing software has been designed. For example, the list of buildings can be given as a selection criterion for the software. If the list contains buildings around power lines, tabulations for the population most likely affected by an electromagnetic field will be calculated. However, the anonymous database does not include geographic information at present. The map data must be stored and processed by GIS software (see Figure 1).

11. Enumerators have been instructed to mark corrections on the EA maps during the field work. The majority of corrections concerned addresses of houses. Some new buildings were also added and some demolished ones removed from the databases. These corrections were entered into the ArcView and MapInfo files. The result will be the most detailed and up-to-date database of Estonian buildings and roads (as of 31 March 2000).

12. There now exists a single database of all dwellings with unique addresses, XY_IDs and x and y coordinates. This data will be linked to the Census database through addresses and each dwelling house will be given a XY_ID.

13. To connect the anonymous Census database and the GIS database at the building level, the address matching was achieved using the PL/SQL procedure. The cyclic databases linking process consists of the automatic linking of entries and an operator's work. The working cycles will be repeated until all entries of the Population Census database have been linked to the GIS database.

14. The houses were then given a unique identifier, XY_ID, which was calculated from the x and y coordinates of the building. From the XY_ID, the location of a house can be recovered with an accuracy of less than 6 meters if the respective function is known. This XY_ID is necessary in order to maintain information on the location of the dwelling houses in the Census 2000 database after the exact addresses (street name, house number and flat number) have been deleted to ensure the anonymity of the database.

15. The software working process will be described as follows:

- Entry data check – before searching for the link the accuracy of entry data will be checked. Inaccurate entries will not be subjected to further processing.
- Search for the corresponding link of an entry from the GIS table – a search for the link from the GIS table in accordance with linking rules. If no link was found or several links were found, then an error message will be included in the error message table.
- Inclusion of an error message in the error table – if an error appears an error message will be included in the error message table. Error messages are t for operators.
- Linking of an entry to the GIS table – the GIS table field XY-ID and the city district code will be copied into the housing questionnaire table in the Census Database.

16. For security reasons, only the Population Census GIS Section of the Statistical Office will be able to undertake this procedure, since this Section knows how to calculate the coordinates from the XY_ID and has the spatial database with addresses, XY_IDs and x and y coordinates.

17. The Census data processing software is powered by a Sun Ultra Enterprise 450 server and Oracle ver. 8.0.5.

18. To improve possibilities to analyse Census data and to overcome drawbacks in the current method of storing spatial data, SOE launched a project to design a central GIS database.

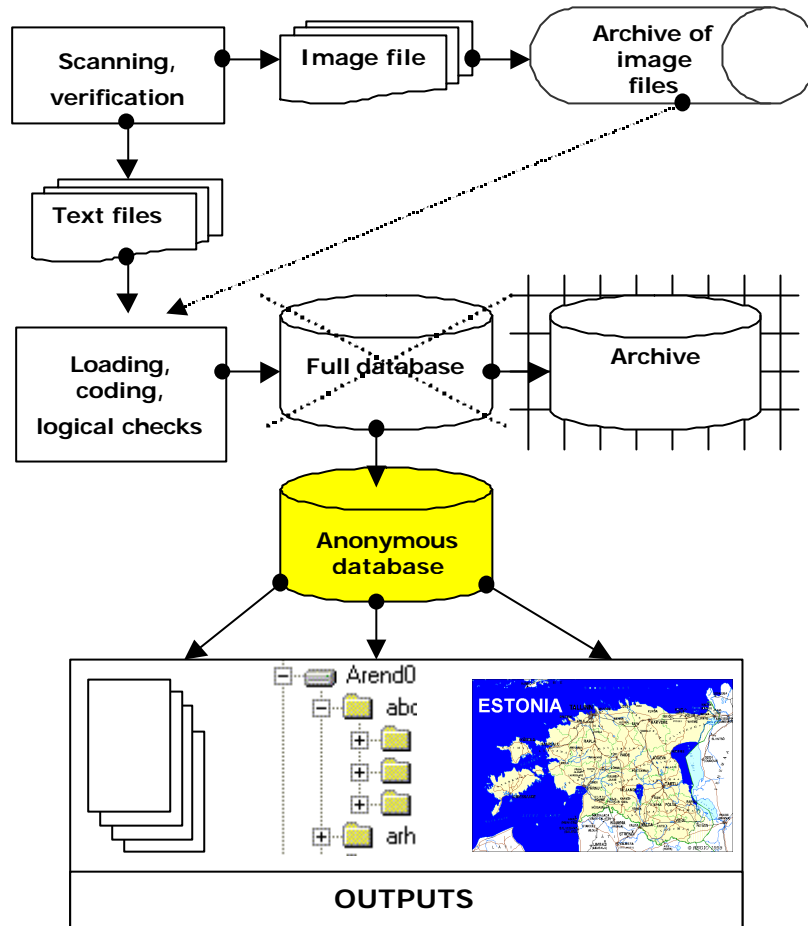


Figure 1. Simplified process diagram of Census data processing.¹ The anonymous database is “GIS-supported”, but does not include spatial data. The goal of the central GIS database is to create a tight association with an anonymous database and collected map data.

II. BUILDING OF THE CENTRAL GIS DATABASE

19. The design of a central GIS database was begun in December 1999. The objective was to create a spatial database which would satisfy the following requirements:

- enable queries over the whole territory
- guarantee user rights
- metadata is stored together with data
- data is not deleted, records are marked “not valid”
- data is suitable for small-scale thematic mapping and large-scale analysis
- the database contains information on Census maps and household lists
- the database avoids data duplication
- use open data structure, permit use of as many GIS software applications as possible
- the database structure enables export of data for Web services in the future
- enable relationship with an anonymous Census database
- enforce the technical quality and logical consistency of the data
- enable exchange of data with municipalities.

¹ Inno, V., Data Processing for Census 2000. Newsletter of the Statistical Office of Estonia, 1999. (In Estonian).

20. Deliverables of the preliminary project were:

- reality model
- data model (ER-diagram)
- metadata specification
- resource estimation for conversion to the new Central GIS database structure.

These models were described in the previous presentation.

21. The database model was created based on the reality model. The database model is presented as an ER-diagram, a familiar tool in information systems development. However, this technique has been used relatively rarely for designing GIS databases. This difference will have to be minimized in the near future, together with the erosion of the difference of the subject of GIS into “mainstream” subject - information systems. A “standard” ER-diagram was enhanced with additional columns for each entity and attribute, to accommodate GIS-specific information (such as digitizing rules, accuracy requirements, etc).

22. Oracle Spatial was chosen as one possibility for storing spatial data. The main reasons were i) the software for storing Census data is already Oracle, and ii) general features of Oracle Spatial, described in Section IV, matched the requirements of the central GIS database. Most of the GIS software packages do not support all standard features of relational database management systems. For example, only the most important subset of SQL has been implemented. This imposes constraints on the database structure. In order to elucidate the widest range of GIS benefiting the client software from the spatial data stored in Oracle, the database structure was kept as simple as possible (for example, synonyms and composite keys were avoided). Oracle Spatial technology is used to store the geography of these objects.

23. Estimation of time and cost of converting data to the new structure was made. The most labour consuming part is the quality control of the data to be converted.

24. To implement the central GIS database, SOE plans to use 10 licenses of Oracle 8i Enterprise Edition and Oracle Spatial. GeoMedia, GeoMedia Pro and Mapinfo Professional will be used as client GIS software for handling spatial data.

III. FUTURE WORK

25. Conversion of data to the central GIS database takes place in two steps: i) common layers of administrative borders and co-ordinates of centroids of buildings were completed during the 1st quarter of 2001; ii) the conversion of other objects into a common database will start in 2002.

26. After finalising these conversion processes, SOE will be able to carry out spatial analysis in the same IT-environment as is common for alphanumeric data. In summary, data of the Estonian Census 2000 will be georeferenced at the building level. This gives SOE a very good source data for a variety of spatial analyses and thematic map applications. The official administrative units, so-called custom-sized units, can be used for thematic mapping, for example, grids.

27. As a result, SOE will be able to perform a detailed GIS analysis of Census data as well as produce various thematic maps based on the tabulation data. A powerful and flexible database system gives an opportunity to provide services for studies initiated by scientists outside of SOE, as well as various on-line services in cooperation with other organizations.

IV. STORING OF SPATIAL DATA IN ORACLE RDBMS

28. Oracle Spatial is the basic technology for the development and implementation of spatial data warehouses. It allows storage and manipulation of both spatial and alphanumeric data in a single database. Any mix of standard Oracle8i tables and spatial data tables can be used with a standard method to retrieve data — SQL. Oracle Spatial is not a GIS software. For displaying graphical data, creating

thematic maps, printing maps, etc., the GIS client software is necessary. The main functions of this software are under discussion now. All major GIS software vendors support Oracle Spatial to a greater or lesser extent.

29. Oracle Spatial supports three basic geometric forms that represent spatial data: [1]

- Points and point clusters – the points may represent locations of buildings, fire hydrants, utility poles, etc.
- Lines and line strings – the lines may represent roads, railroad lines or utility lines;
- Polygons/complex polygons with holes – the polygons may represent outlines of cities, districts, and vegetation. A polygon with a hole may geographically represent a forest surrounding a clearing.

30. In addition, rectangle, circle and compound elements with and without an area are supported. The main advantages of storing spatial data in Oracle RDBMS are:

- all data elements are in the same environment, increasing data security, ensuring data integrity, etc.;
- spatial data is stored in the format which enables its use by a number of GIS clients;
- application software developers can choose between several software tiers for optimizing security, speed, ease of use and development flexibility;
- spatial data is documented, structured and organized similarly to alphanumeric data.

31. The main disadvantages of storing spatial data in Oracle RDBMS are:

- the technology is new and technical specifications have been changed a number of times up to Oracle 8i Spatial;
- the experience with storing spatial data in RDBMS is rare in the GIS community and the learning curve is steep;
- the client software versions are sometimes unstable;
- Oracle does not support multiple coordinate systems and raster data.

32. In conclusion, Oracle Spatial does not replace traditional GIS file formats, but provides a valuable option for occasions which demand added value provided by Oracle and which do not suffer from a relatively young technology.

V. REFERENCES

[1] Oracle Spatial. Data Sheet, March 1999.

[2] Inno. V., Data Processing for Census 2000. Newsletter of the Statistical Office of Estonia, 1999. (In Estonian).