Distr.
GENERAL

Working Paper
9 April 2013

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2013)
(Paris, France, and Bangkok, Thailand, 23-25 April 2013)

Topic (ii): Streamlining statistical production

Business Case for Industrialisation in Statistics Estonia: Small Example of a Large Trend

Prepared by Allan Randlepp and Tuulikki Sillajõe, Statistics Estonia

I. Introduction

1. Statistics Estonia conducted Population and Housing Census (PHC) 2011 in 2012. The census moment was 31 December 2011. The Census was carried out from 31 December 2011 until 31 March 2012. From 31 December 2011 to 31 January 2012, an e-Census took place where the residents of Estonia could fill in questionnaires online. From 16 February to 31 March 2012 enumerators visited those who had not participated in the e-Census. Revision of collected PHC data ended 31 October 2012.

2. The paper gives an overview of the changes made to statistical production system to conduct PHC 2011, related costs and gained efficiencies.

II. Business case for Population and Housing Census 2011

A. Data collection phase

3. Mixed-mode data collection (CAWI + CAPI + registers) was used for the Population and Housing Census (PHC) 2011. For that purpose, a considerable part of the information systems at Statistics Estonia were modified. Although a new data collection system was developed which can be used as a generic system for other statistical domains, but ensures cost-efficiency even if used only for data collection for PHC 2011.

4. Table 1 compares the costs of data collection of PHC 2000 and PHC 2011. The data in the first row of the table are based on real costs of PHC 2000, taking into account the consumer price index of the past ten years (meaning that if in 2011 the same data collection method had been used as in 2000, the cost would have been 13.2 euros per enumerated person). If in 2011 new generic software and a different method of data collection

had been used (compared to 2000) and the online participation rate had been 25%, the cost of data collection would have been 8.14 euros per enumerated person. The real cost per enumerated person was 6.50 euros.

Table 1. Costs of data collection of the Population and Housing Census 2011

	e-Census participation rate	Number of enumerated persons (preliminary)	Number of enumerators	Cost of data collection, euros	Cost of data collection per enumerated person, euros
PAPI (prediction)	0%	1,340,000	4,592	17,710,000	13.2
CAWI+CAPI (prediction)	25%	1,340,000	2,970	10,910,000	8.14
CAWI+CAPI	66%	1,294,236	2,131	8,410,000	6.50

5. The cost per enumerated person might seem high compared to countries with a bigger population where the economies of scale are higher. But it was at least twice as small as it would have been if PAPI had been used; and it was at least 25% lower than predicted for use of the new method of data collection. So, in the Estonian context, efficiency has been reached.

6. The main reasons for lower costs of data collection during PHC 2011, compared to PHC 2000, were as follows:

(a) There were no special data entry costs;
(b) Automatic checks diminished mistakes during interviewing in both cases (CAWI and CAPI);
(c) Less time was required for interviewing than planned, and therefore fewer enumerators were hired;
(d) Printing, communications and archiving costs were considerably smaller;
(e) Management of data collection was much cheaper because much fewer employees were needed, thanks to the management module of the new software.

7. Talking about PHC 2011 in Estonia, one of the considerable achievements was the outstanding participation rate in the e-Census. It could not been achieved without the widespread use of Internet in Estonia, but was significantly influenced by the well-planned, strictly targeted and executed public campaign. The public campaign was deeply integrated and coordinated with the data collection activities. A special awareness survey was launched about a year before actual data collection started. Based on that survey, the PR activities were corrected and targeted. So, in February 2012, after the e-Census, awareness of the census among the population aged 15–74 was 99%, compared to 98% in January 2012 and 57% in August 2011. At the same time, 95% considered the census necessary (72% answered ’necessary’ and 23% ‘quite necessary’); 90% considered their knowledge about PHC ‘good’ or ’very good’, compared to 45% in August 2011. Already in November 2011, 60% of the respondents planned to participate in the e-Census.

8. One of the communication activities during PHC 2011 was reporting about the progress of the Census on the web site of Statistics Estonia. The number of enumerated persons in total and by county was updated hourly. This information was highly appreciated and widely discussed by the society as a whole, especially in the media. The comparison of counties started to serve as an additional driver, motivating people to enumerate themselves online.

9. Another effective feature on the web site of Statistics Estonia was a tachometer with a green, yellow and red zone for indicating the workload and traffic on the census page. The green zone indicated the best time for self-enumeration, the yellow zone marked a good time for self-enumeration, and the red zone meant possible disturbance or slowness of the system.

10. In case of PHC 2011, the social media also played an important role. During the data collection period, Statistics Estonia’s PR staff worked on Facebook virtually for 24 hours per day. Information was quickly disseminated. People’s questions were answered when the lines of the Contact Centre were busy or e-mails were not answered as fast as expected.

11. From Statistics Estonia's point of view, the new generic software for data collection has several outstanding features. It supports the data collection process of various surveys and censuses, dividing it into three sub-processes:

(a) Preparation work for data collection;
(b) Data collection on the web (CAWI) and fieldwork (CAPI);
(c) Support and management of the whole process.

Preparation of data collection allows to define questionnaires; to import samples from the statistical register; to assign the people involved in data collection and the hierarchy of management; to divide work tasks within the hierarchy; to grant access to systems and assign relevant roles for statisticians, survey managers, the call centre, interviewers, etc.; to pre-fill questionnaires with data from registers; to communicate with different people involved in preparation.

During CAPI it is possible for interviewers to navigate between different questionnaires, to use map info for easier subject location and work planning; to plan their work and to schedule contacts with interview subjects. The local database in the interviewers' laptops is encrypted; all data for offline work are synchronised over an encrypted channel.

Fieldwork managers have an overview of fieldwork progress and can easily and quickly distribute tasks between interviewers and approve filled-in questionnaires. The staff of the Contact Centre can schedule contacts and assist with questionnaire completion. There is also functionality for data operators who can classify answers in questionnaires and remove duplicate questionnaires. Communication between managers of different levels, statisticians, and other parties is supported. Various activities in the system are centrally logged.

B. Data processing phase

12. As mentioned before, a considerable part of the information systems at Statistics Estonia were modified for PHC 2011. Among the other systems a new data processing system was developed which can be used as a generic system for other statistical domains, but ensures cost-efficiency even if used only for data collection for PHC 2011.

13. Table 2 compares the costs of data processing of PHC 2000 and PHC 2011. The data in the first row of the table are based on real costs of PHC 2000, taking into account the consumer price index of the past ten years (meaning that if in 2011 the same data collection and processing method had been used as in 2000, the cost would have been 3.6 million euros). If in 2011 new generic software and a different method of data collection had been used (compared to 2000) and the online participation rate had been 25%, the cost of data processing would have been 2.3 million euros. The real cost of data processing was 1 million euros.

Table 2. Costs of data processing of the Population and Housing Census 2011

	e-Census participation rate	Number of operators	Cost of data collection, euros
PAPI (prediction)	0%	264	3,600,000
CAWI+CAPI (prediction)	25%	90	2,300,000
CAWI+CAPI	66%	30	1,090,198

14. The main reasons for lower costs of data collection during PHC 2011, compared to PHC 2000, were as follows:

(a) Automatic checks diminished mistakes during interviewing in both cases (CAWI and CAPI);
(b) The first stage of data processing, primary data arrangement, was part of the data collection software and ran as parallel activity during data collection and immediately after collection. As a result the collected data was prepared for the next stages of data processing.

(c) Data processing (coding, amending errors, handling of duplicates, etc) was highly automated. Totally 35.7 million errors were corrected, 20.2 million of them corrected automatically and 15.5 million manually.
(d) Problems with data came out in early phases of processing, because statisticians were able to monitor data during the data processing.
(e) For checks and handling missing data was possible to use data from registries and from PHC 2000.
(f) Management of data collection was much cheaper because much fewer employees were needed, thanks to the automated data processing done by the new software.

15. The new generic software for data processing has several outstanding features. It supports data processing of various surveys and censuses and can be used for much general data integration tasks like:
(a) Extracting data form different sources, including administrative registers;
(b) Preparing data to pre-fill questionnaires;
(c) Load data to data warehouse;
(d) Etc.

16. The new data processing software is a collection of tools and technologies aimed at automating data processing (Phase 5 in GSBPM) related to statistical activities to prepare data records for analysis.
In essence, the task of check, clean, and transforming statistical activity data can be identified as taking the raw data from one or more sources (sub-process 5.1 Integrate data) survey (administrative registry, sample questionnaire, census, etc) and transforming it to analytical system source data input data base structures (sub-process 5.8 Finalize data files).

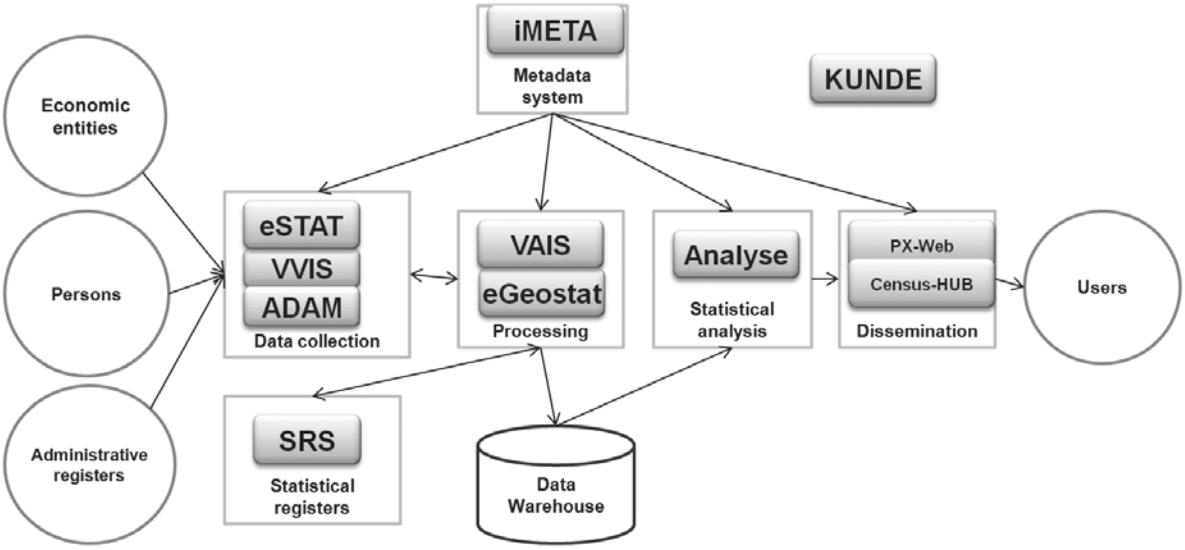
17. Besides pure data transformation from one structure to another, there are a number of operations that need to be carried out:
(a) Classifying and coding the input data. Including automatic and clerical coding routines which assign numeric codes to text responses according to a pre-determined classification scheme (sub-process 5.2 Classify and code).
(b) Validation of raw data for technical and logical errors, correction of errors and amendment of raw data (sub-process 5.3 Review, validate and edit).
(c) Imputation of missing data fields and/or statistical units using a rule-based approach (sub-process 5.4 Impute).
(d) Derive variables and statistical units that are not explicitly provided in the collection, but are needed to deliver the required outputs (sub-process 5.5 Derive new variables and statistical units).
(e) Create weights for unit data records according pre-defined methodology (sub-process 5.6 Calculate weights).
(f) Create aggregate data and population totals from micro-data (sub-process 5.7 Calculate aggregates).
This software is designed to provide tool framework for all these tasks.

18. Data processing managers have an overview of processing progress and can easily and quickly distribute tasks between operators and add additional rules or change existing ones. The staff of the Contact Centre can schedule contacts and assist with questionnaire completion. There is also functionality for data operators who can classify answers in questionnaires and remove duplicate questionnaires. Communication between managers of different levels, statisticians, and other parties is supported. Various activities in the system are centrally logged.
Statisticians can design and build different data processing workflows using re-usable components and filling in necessary metadata, like checking rules, imputation method to be used, calculation algorithms for calculated variables, etc.
Programmers can add or modify re-usable components, monitor running jobs, run jobs manually or automatically.

III. The production system as a whole

19. The data collection and processing phase cannot be approached separately from the seven other phases of the statistical business process. At Statistics Estonia, generic office-wide software has been developed or is under development for other phases as well.

Figure 1. Architecture of the information system



20. The architecture of the information system is presented in Figure 1, where the acronyms stand for the following information systems:

(a) iMETA – integrated metadata system (2011);
(b) SRS – system of statistical registers (2012);
(c) Kunde – customer relationship management system (2006);
(d) eSTAT – data collection system for economic entities (2006);
(e) VVIS – data collection system for private individuals (2011);
(f) ADAM – data collection system for administrative data (2011);
(g) VAIS – template-based data processing system (2012);
(h) Analysis system (2013);
(i) PX-Web – output database (2001), will be replaced by .Stat (2012 and onwards);
(j) Census Hub – dissemination tool for European statistics (2013–2014).

21. In the context of data collection, the data collection system for administrative data should be described in more detail. At Statistics Estonia, the Data Collection Department is not responsible for that type of data collection. This responsibility is divided between two other central departments of the office: the Methodology Department (established in 2004) and the Data Processing Systems Department (also established in 2004, but reorganised on the basis of the IT Department in 2011). By the way, in case of Statistics Estonia, the bulk of traditional IT-functions are supplied by another institution – the IT Centre of the Ministry of Finance (established in 2012). It means that the Ministry of Finance has centralised IT-functions in its area of administration.

22. The Methodology Department consolidates the needs of statistical domains within Statistics Estonia for data available in registers; conducts negotiations with the holders of administrative and other registers, and organises the conclusion of agreements with these holders. The Methodology Department is also in charge of the description of data in a central metadata system. Technically, the data from registers enter Statistics Estonia via a single entry point which is under the responsibility of the Data Processing Systems Department. This

department runs pre-agreed data processing and makes the data available for in-house applications. This ensures that there is no duplicate collection of data and the data are ready for statistical analysis.

23. In 2012 Statistics Estonia used about 100 different administrative registers (Population Register, Estonian Education Information System, Register of Construction Works, Health Insurance Database, Commercial Register, Register of Taxable Persons, State Register of State and Local Government Institutions, etc.) for the production of official statistics. In case of those registers which are more widely used within the office or from which data is taken more often, automatic extraction of detailed personalized data takes place using X-road (national data exchange layer). A special data acquisition application (ADAM) has been created for transition to register-based statistics production.

24. Statistical registers have been an important prerequisite for standardisation and improved efficiency of processes at Statistics Estonia. In 2011 the office started to develop the System of Statistical Registers (SRS). SRS will integrate existing statistical registers (economic entities and agricultural holdings) and new statistical registers (persons, and buildings and dwellings) into a common system.

25. The current strategic approach of Statistics Estonia has been to develop only generic office-wide software, but whenever possible to use commercial statistical software (SAS, SPSS, etc.) or software developed by other members of the international statistical community. So, for the output database, PC-Axis has been used for many years and .Stat will be used from 2012 onwards; µ-Argus and Ʈ-Argus have been used for disclosure control, CLAN for variance estimation, IVEware for imputation, and Demetra for time series analyses. Statistics Estonia believes that this is an example of how scarce resources have been used efficiently. More statistical domains have benefitted, even though at the same time great efforts have been made for the standardisation of methods and working routines.

IV. Conclusions and further developments

26. Statistics Estonia has observed economies of scale based on centralisation of functions. The currently centralised processes are dissemination (1993), data collection (2004), methodology (2004) and IT (2004). Centralisation of data processing is under serious consideration.

27. Inside the function of data collection, increases in efficiency have been achieved through the integration of data collection from economic entities and individuals. Efficiency has been found with the help of close cooperation between the data collection and dissemination functions. In the future, the stronger leadership of methodologists, supported by systematic implementation of common software for all statistical domains, could lead to a level of efficiency which has not been reached ever before. These gains are imbedded not only in standardisation, but also in bringing in new sources of data and using more complicated methods of data collection (e.g. data mining from existing sources, etc.).

28. Statistics Estonia was able to successfully conduct PHC 2011 at a two times lower cost per enumerated person than the use of the method of the previous census would have allowed. It was achieved due to the change of method and implementation of newly developed software.

29. Based on the software developed for PHC 2011, Statistics Estonia plans to start introducing CATI for data collection from both types of respondents (economic entities and individuals) step by step from 2012, but also to introduce CAWI for other surveys on individuals starting 2013. Pilot surveys using CAWI are scheduled for the 4th quarter of 2012.

30. There are also plans to increase the training offered to data suppliers (economic entities, individuals, registers, etc.), underlining their personal and public gains from official statistics.

31. There is still room for further simplification of statistical questionnaires (harmonisation of concepts, deadlines, practices, etc.). A specialised laboratory could find a way through the existing problems.

32. And last but not least, Statistics Estonia has plans to develop infrastructure for selling data collection services. This infrastructure should include questionnaire design, sample design, interviewing, management of interviewers, and processing and primary analysis of collected data.

V. References

Annual report 2011. (2012). Statistics Estonia. [www] http://www.stat.ee/annual-report (26.10.2012)

Statistical Database of Statistics Estonia. [online database] http://pub.stat.ee/px-web.2001/dialog/statfile1.asp (26.10.2012)

The Generic Statistical Business Process Model. Version 4.0, approved by the METIS Steering Group in April 2009. [www] http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model (26.10.2012)