

**UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on Statistical Data Collection
(Geneva, Switzerland, 25-27 September 2013)

Topic (i): Centralising data collection

TEN YEARS OF CENTRALISED DATA COLLECTION

Working Paper

Prepared by Tuulikki Sillajõe, Kaja Sõstra, Statistics Estonia

I. Introduction

1. At Statistics Estonia, the data collection tasks of the value-chain of official statistics production were organized into a central functional department in 2004. Since then the central department has provided all specialised statistical departments with data collected from both individuals and economic entities.
2. Centralised data collection has been enabled by the use of office-wide generic software for data collection. A web-based data collection system for economic entities is used since 2006. New software under EUPL for web-based data collection was developed and first used for the 2010 Agricultural Census and for the 2011 Population and Housing Census. The same tool will be implemented for all statistical surveys on individuals starting 2014
3. The paper gives a brief overview of the challenges and problems met during these years. It describes further plans for division of work planned to be implemented in October 2013.

II. Current organization of data collection

4. There is a Data Collection Department which is responsible for supplying all five specialised statistical subject matter departments with all the raw data collected from individuals and economic entities. Raw data are data in the same form as presented to Statistics Estonia by data suppliers. It means that the Data Collection Department does not validate the data they collect. They just have to bring the data into the office. All required checks are made by the subject-matter departments.
5. The Contact Centre is a part of the Data Collection Department. Respondents can call the general telephone number and send e-mails to the general e-mail address set up for questions and enquiries about any questionnaire to be submitted by economic entities irrespective of the domain (e.g. construction, trade, agriculture, wages and salaries, etc.). The Contact Centre handles contacts with the respondents of social surveys as well. The staff of the Contact Centre is able to answer 99% of the respondents' questions.

6. There are two office-wide generic software programs for data collection: the older version is better suited for data collection from economic entities and the newer version is better suited for data collection from individuals. Both have been developed specially for Statistics Estonia.
7. There is a single authorisation point on the Internet called eSTAT. After logging in, each respondent has access to all the questionnaires assigned to him or her as an individual or as the contact person of one or more economic entities. The respondent is not informed whether he or she is using the economic entity's part or individual's part; the respondent might even not realise that he or she is moving between these two parts. The difference lies in the suitability of a particular questionnaire for one of these software programs.
8. CAPI, CATI and CAWI are used.
9. The data collection from individuals has been completely paperless — laptop computers (CAPI) have been used instead.
10. It is possible to present all (188 in 2013) statistical questionnaires of economic entities via eSTAT. The share of electronic data collection from economic entities has reached about 85% of the total number of statistical questionnaires received from economic entities by Statistics Estonia, but there are still questionnaires coming in on paper which have to be keyed into the computer.
11. Each year in December, economic entities are simultaneously informed about all the questionnaires they have to complete in the following year. The information is sent directly to the economic entities, but also published on the website of Statistics Estonia. E-mail reminders are sent to economic entities before (not after) the deadline of each statistical questionnaire.
12. Also, the design of data collection instruments is still decentralised; their building of electronic instruments is done centrally by the Methodology Department. Due to that, a great amount of standardisation and simplification has been done.
13. Altogether, in terms of the Generic Statistical Business Process Model, the phases Collect and Disseminate are centralised, but the phases Specify Needs, Design, Build, Analyse, Archive and Evaluate are still totally or partially decentralised.

III. Development of data collection at Statistics Estonia

A. Economic entities

14. At the beginning of the 2000s, Statistics Estonia made vigorous preparations for joining the European Union. New tasks had to be fulfilled to comply with the Acquis Communautaire in the field of statistics. At the same time, new information technologies and financial resources were at the office's disposal.
15. One of the triggers was the need to prepare for the implementation of the EU harmonised survey Intrastat to collect data on external trade transactions between the EU Member States. After joining the EU, customs borders disappeared and the data about foreign trade had to be collected by surveys. Consequently, the need arose for a new system of data processing, collection, checking and dissemination as well as for an Intrastat enterprise register which would include enterprises engaged in foreign trade. It was decided to develop such a new system that would also be suitable for other surveys on economic entities.
16. The new system was first tested in the framework of the Intrastat project in 2003. It was introduced to the respondents in summer 2004.
17. In autumn 2004, major reorganization took place at Statistics Estonia. Besides the central Dissemination Department, which had operated already since 1993, three new central departments were created: Data Collection Department, Methodology Department and IT Department.

18. The centralisation of data collection and IT development was mainly aimed at the improvement of the production process, while the establishment of the Methodology Department also had another objective — to achieve better comparability and cohesion of official statistics through better coordination of classifications, definitions and methodology. It was planned that the reorganized statistical departments should orientate to the production of statistics based on statistical analysis. The change meant a significant reorganization of the work process. Statistics Estonia started to outsource its supporting activities (e.g. posting of questionnaires).
19. As a result of the reorganized work process and the restructuring, the number of employees at the office was reduced by 62 persons, or 16%. However, the main tasks of Statistics Estonia for the year 2004 were successfully accomplished and the quality of statistics was maintained in spite of some reorganization-related difficulties.
20. It took almost two years after the creation of the central Data Collection Department before Statistics Estonia was able to implement the web-based data collection channel eSTAT. Among the other objective and subjective obstacles was the fact that the initial version of the software lacked a respondent-based view. It allowed a perfect overview for each statistical questionnaire, but not by economic entities. So, it was very time-consuming for a respondent to answer a simple question in two different questionnaires.
21. In 2006 Statistics Estonia launched eSTAT, the web-based data collection channel for economic entities, which was declared “The Best Innovation 2006” by the State Chancellery in the same year. Since then, the share of electronic data collection has risen year by year. The software itself has been developed further to make it more convenient for both respondents and data collectors.
22. Some concrete actions had to be taken to canalize respondents into electronic data collection channels. In 2008 Statistics Estonia stopped mailing empty printed questionnaire forms to economic entities. In the same year, partial pre-filling of eSTAT with data from the previous reporting period was introduced. Since 2011 the annual statistical questionnaires of economic entities are pre-filled using administrative data. All of these actions helped to increase the share of data transmitted via eSTAT.
23. However, there was a critical level of mistrust within Statistics Estonia towards central data collection. This was one of the reasons why the checking of data quality, after its arrival to the office, was left to the statistical subject-matter departments. For data checking purposes, each statistical department had to keep specialised staff for calling the respondents and specifying the data. As they were contacting respondents, the same respondents saw them as their contact persons within the statistical office and kept contacting them whenever they had questions. So it happened that there was a considerable amount of second-level user support, as they were called inside the office.

B. Individuals

24. 2004 marked ten years since the interviewers team of Statistics Estonia was formed. At the beginning of 2003, the work of interviewers underwent thorough reorganization. All over Estonia, 58 interview districts were formed with one interviewer working in each district. Previously, interview districts had been formed and interviewers recruited separately for each survey depending on the county sample size. Now, however, the borders of interview districts do not change from survey to survey and they are no longer based on the borders of counties. These changes enabled faster and more flexible data collection as well as a better overview of the interviewers’ work. The new interview districts were formed on the basis of digital Population Census maps.
25. Step by step, electronic data collection was implemented. This accelerated the whole process of data collection and improved data quality. This also enabled the statistical office to save on labour force. 2002 was the year when the electronic data collection was tested; in 2003 the collection of data for the Labour Force Survey (LFS) using laptop computers started. This enabled the immediate transformation of the data into computer-readable form. Preparations started for the electronic

collection of data for household pictures and post-interviews within the Household Budget Survey. The pilot survey took place in 2004.

26. Since 2005 laptop computers are used for interviewing for all surveys on individuals. Using laptop computers has shortened the time spent on interviews, and the logical controls in data entry programs help to solve most of the problems already during the interviews.
27. This has given rise to great changes in the data quality and in the survey-related work organization. On the one hand, the amount of time used for insertion of data during interviews has decreased. On the other hand, the initial check-up of data to be collected is done already in the course of the interview, which saves the time spent on further data editing. Electronic data collection enables the survey managers and coordinators to receive instant feedback already during the data collection process — in the past, they had to wait for feedback for weeks or even for months.
28. Use of laptop computers has also helped to enhance the security. Before, an interviewer used to fill in a paper questionnaire which had to be sent for data entry and control. Now, the interviewer fills in the questionnaire in the computer and transmits it electronically to the central database. The data are more secure in the computer than on paper and data controls, which improve the data quality, can be added to the questionnaire. If previously the information about the quality of the collected data reached the manager of the survey only after several months, then now this information is transmitted within hours and the overview of the interviewers' work and data quality is much better.
29. So, Statistics Estonia used web-based interviewing (CAWI-methodology) when collecting data in the framework of the 2007 career-related survey of doctorate holders. The result surpassed preliminary estimates: almost 40% of the interviewees preferred to fill out a web-based questionnaire instead of meeting with the interviewer.
30. The real breakthrough in the use of CAWi for data collection from individuals was made in 2011 when 68% of the whole population participated in the web-based Population and Housing Census. Starting 2014 the same mode will be available for all surveys on individuals.
31. Statistics Estonia has used CATI several times by outsourcing. Since 2013 it has its own CATI-team, used for collecting data from both economic entities and individuals.

IV. Impact on quality

32. The main indicator for measuring the quality of the data collection process in case of both household and business surveys is the response rate. Information about response rates is available in the public database for some surveys. Our aim is to make this information available for all surveys whenever it is relevant.
33. In case of economic entities surveys, the response rate is calculated separately for small and large entities. Response rates have been quite high and stable over the past ten years. For large entities (50 and more employees), the overall response rate has been over 99% in the survey "Wages and salaries". The response rate of small entities is lower — between 80–85% since 2004. It should be noted that the response rate for small entities increased remarkably between 2000 and 2004, from 55% to 82%. Also, the response rate depends on the economic activity of business entities. In 2012 it was the lowest for construction and wholesale and retail trade (78%) and the highest for human health and social work activities (93%) (for small business entities, in the survey "Wages and salaries").
34. In household surveys we follow response rates for different subgroups: by sex and age, by county, by interviewer etc. In case of longitudinal surveys, the response rates are calculated separately for the first wave and subsequent waves. In the Labour Force Survey (LFS), the response rate for the first wave has decreased from 74% in 2003 to 53% in 2009. After that, the response rate has been stable with a slight increase last year.

35. Reasons for non-response are also registered in case of household surveys. The main reason for not participating in the survey is refusal and inability to participate (40% of total non-response in LFS in 2004–2012). Also a very common reason for non-response is non-contact (34%). We analyse the reasons of non-response and try to find solutions for avoiding it. There are several actions taken: training of interviewers, higher compensation for interviewers with better results, better information for interviewers (e.g. the contacts of housing associations to avoid non-contact due to the fact that the general entrance door of the building is locked).
36. Some survey-specific quality indicators are also in place. For example, in LFS we follow the time lag between the reference week and the first contact with the respondent. Several questions in the LFS questionnaire concern the reference week and if the time lag increases the respondent is not able to give correct answers.
37. In cooperation between the Data Collection Department, Methodology Department and Social Statistics Department, the Eurostat-financed grant “Implementation of quality improvement actions for the Labour Force Survey” is carried out. This grant includes different actions for improving the data collection process (e.g. training, continuous monitoring of interviewers). The first results showed improvement in both quality indicators – response rate and time of the first contact. As the result of the project: the share of timely first contacts with sampled households increased from 77% in 2011 to 95% in the 1st quarter of 2013; the share of interviews contacted during the 1st week following the reference week increased from 44% in 2011 to 63% in the 1st quarter of 2013; the response rate increased from 62.5% in 2011 to 76% in the 1st quarter of 2013.

V. Further steps

38. The working time of the employees of Statistics Estonia has been analysed since 2007, in the framework of the Generic Statistical Business Process Model (GSBPM). These analyses have showed quite a high level of specialisation in the central departments and a lot of irregularity and variety within the statistical departments.
39. As Statistics Estonia is still looking to increase efficiency, it was realised that the constant changing of activities over time is less effective than specialising on a particular activity. And so, on the 1st of October 2013, a new structure of Statistics Estonia will come into force. Two keywords of the reorganization are specialisation and efficiency.
40. The new workflow will be more process-oriented, dividing the office into two parts: one concentrating on the activities before the data warehouse, and the other concentrating on the activities after the data warehouse. The aim of the input team will be to put clean microdata into the data warehouse and the aim of the output team will be to make attractive and user-friendly products based on the data warehouse.
41. The biggest organizational change behind this will be that all data collection activities and all processing activities under the GSBPM are consolidated from the current specialised statistical subject-matter departments.
42. In practice, the Data Collection Department will get additional responsibilities of handling all contacts with respondents as well as making all manual corrections in their data. In addition, a totally new department — Data Processing and Statistical Registers Department — will be established. In this department, former employees of statistical departments will do their job in a more standardised and efficient way based on best practices. Statistical departments will be able to concentrate on analysis and creation of products. Also, the number of statistical departments will decrease from five to three, further enhancing standardisation.
43. In the brand new organizational structure, there will be five departments on the input side: Data Collection Department, Data Processing and Statistical Registers Department, Data Warehouse Department, Metadata Department and Development Department. On the output side, there will also

be five departments: Methodology and Analysis Department, Population and Social Statistics Department, Business and Agricultural Statistics Department, Economic and Environment Statistics Department, Dissemination Department.

44. This will be a huge change of mentality for statisticians and probably manageable due to the fact that the office will move to a new building on the 23rd of September 2013.

VI. Conclusion

45. Concerning the speed of changing respondents' behaviour, incl. towards new functionalities of a data collection channel, the product assimilation curve, a term originated in marketing theory, applies. There are about 2.5% of innovators who immediately start to use, 13.5% of early followers, 34% of early majority, 34% of late majority, and 16% lag behind. But still, communication is essential. It is essential to listen to users of the system and communicate their needs to the developers of the system, and afterwards communicate new features back to the users and persuade them to use the developed features. And coordination between different actors around the system is of no less importance.
46. The new software might influence the working routines by initiating dramatic and rapid changes within the office, but not so rapid changes in the habits of external users. But at the same time, the software itself is not enough; usually, organizational changes are also needed.
47. An office-wide generic software could be developed, but hardly implemented, if there is no central department acting as the owner and main user of that software.
48. 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 2014 onwards; μ -Argus and τ -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.
49. Centralisation has enhanced efficiency. The currently centralised processes are dissemination (1993), data collection (2004), methodology (2004) and IT (2004). In 2013 the annual expenditure from state revenues (excl. expenditures of the Population and Housing Census) of Statistics Estonia is roughly on the level of 2006, but the office has not noticeably reduced the output. The quality of statistics has been maintained.

VII. References

Annual reports of Statistics Estonia 2003–2012. [www] <http://www.stat.ee/annual-report> (22.09.2013)

Estonian Labour Force Survey. Methodology. [www] <http://www.stat.ee/dokumendid/65482> (22.09.2013)

Sillajõe, T. (2012). Statistics Estonia on its way to improving efficiency. Paper presented at the Seminar on New Frontiers for Statistical Data Collection in Geneva in 2012. [www] <http://www.unec.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP31.pdf> (22.09.2013)

Statistics Estonia (2013). Quality improvement actions for LFS. Final Technical Report. Grant agreement No 10201.2011.001-2012.180.

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>
(22.09.2013).