UNITED
NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2005/8
23 March 2005

Original: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-third plenary session
(Geneva, 13-15 June 2005)

USING THE WEB IN COLLECTING DATA FOR BUSINESS STATISTICS IN FINLAND

Invited paper submitted by Statistics Finland[*]

INTRODUCTION

1. The primary sources of data for statistics in Finland are administrative systems. The Finnish Statistics Act decrees that existing data must be used for statistics in preference to direct data collection. Without going into details, about 94 per cent of the gross data[1] collected by Statistics Finland (SF) are derived from administration. Of the rest, about 2.5 per cent are derived from interviewing persons and around 3.5 per cent are collected directly from government institutions and enterprises themselves. Of this 3.5 per cent, about 1.8 per cent are collected in electronic form (Excel, ASCII and sequential files) and 1.2 per cent using web forms. The share of data collected on paper forms will be about half a per cent in 2005.

2. It must be pointed out that the value of the data collected directly from individuals, institutions and enterprises is far greater than its share of the gross data. In many cases, administrative data and survey data are combined at the statistical unit level to produce the required statistics. Direct collection is necessary for several reasons, the main ones being to obtain data that is not available in the administrative systems, to more rapidly obtain data for particular statistical units and to obtain data directly from large units so as to better control the quality of the data

3. These two types of data are combined, for example, in the population census, income distribution statistics and structural business statistics. For those business statistics where some

[*] Prepared by Ilkka Hyppönen.

data can be obtained from administrative sources, it is typical to collect data directly only from “the largest businesses”. The term “largest” can be used to describe a business “employing 20 or more people”. This is due to the fact that statistics are required at a detailed level of classification, obtainable from enterprises that are very small by international standards.

4. It should also be pointed out that there is no substituting administrative data available for most surveys. This is especially true of local government units, for which many statistics on education and finances are compiled at SF. These are mainly of an administrative nature, and they have mostly been transferred from administrative bodies to SF during the last 15 years in an effort to rationalise the data collection of the central government. However, in spite of these considerations, these statistics are accepted as a part of “official statistics” and are included in this paper.

5. There are two economic entities from which data is collected regularly: businesses and local government institutions. Data collection from these two groups covers about 92 percent of direct data collection by SF (excluding data collection from households). These two groups will be discussed below.

E-GOVERNMENT

6. According to many international comparisons, Finland is considered to be a well-advanced society as regards the use of the Internet. Access to the Internet is very common in Finland and over 95 percent of the enterprises and 100 percent of the local government authorities have access to the Internet (2003). For a number of years now, the Finnish government has run a programme for e-government and advancement of the information society. This has generated a positive atmosphere for using the Internet for transactions with the government.

7. As already mentioned above, the policy of SF is to use existing data when available. This covers both government and private sources. In some cases, SF even buys data from the private sector. In addition, advancement of electronic data collection has been a SF strategy since the 1970s. The result can be seen in the share of direct data collection in file format. The first application using web forms was in 1997, but only in 2001 did progress accelerate. At that time, SF set a target date by which each paper form of data collection should have an electronic alternative. A major project called "Production model" has been under way for two years. One of its sub-projects concerns the data collection process, a primary target of which is the implementation of SF’s strategic goal regarding data collection. This has accelerated the progress considerably over the last year.

SURVEYS OF BUSINESSES AND LOCAL GOVERNMENT

8. There are about 60 different surveys of enterprises. Of these, 15 concern the financial markets and are designed together with the Financial Supervision Authority in cooperation with banks and other financial institutions. They are based, at the moment, on Excel spreadsheets and are very comprehensive in content. These will not be discussed any further in this paper. For many other business surveys as well, it has long been possible to report data in electronic form in addition to paper format.

9. Most of the business surveys in Finland are small, measured by the number of respondents and the number of data items. Over one half of business surveys are annual and the remaining ones are either monthly or quarterly surveys. Monthly and quarterly surveys contain, for the most part, only a few questions/variables per respondent, whereas the annual surveys can consist of hundreds of data items.

10. There are 12 surveys directed to local government authorities that use paper and/or web forms. Among these are statistics on local government finances, education and local government personnel. With two exceptions, these are annual collections, one of them being quarterly and one half-yearly. Typically, these surveys are fairly large, in the number of variables or in the number of personnel or students.

TWO TYPES OF SOLUTION

11. Two solutions are used in Internet-based data reporting: one is to use the services of an outside service provider and the other is to develop applications in-house. There are about 50 applications in production, under construction or being planned. 21 use/will use an outside service provider and 20 are/will be developed in-house. For the rest, the decision has not yet been made.

Number of web collections by status and solution at the beginning of 2005

	In-house	Outside	Not yet decided
In production	11	14	
Under construction	3	3	
Planned	6	4	11
Total	20	21	11

12. The outside service provider functions as an intermediary between the respondents and SF. The service provider develops the forms and handles respondent administration such as user IDs and passwords.

13. In-house development of an application began in 2000 and resulted in a model application that has since been refined and generalized. The application used today, developed in-house, is called XCola (XML based Collection Application). In Xcola, the questionnaires are defined as XML documents, and are then customized for each respondent and transformed into web pages at runtime. Annex 1 gives more information about XCola.

DATA SECURITY

14. All data security aspects are handled by the outside service provider. The data are sent to SF as encrypted files.

15. Concerning the in-house solution, the following technical components/arrangements contribute to security:

- all traffic on the Internet is SSL-encrypted (128 bit encryption);
- the XCola engine and the collection database reside in two separate local area network (LAN) segments, which are connected with firewalls both towards the Internet and towards the rest of the SF LAN;
- it is not possible to initiate an Internet connection from these segments; servers on these

segments can only receive/accept connections;
- the segment where the database server is located can receive traffic only from the segment where the XCola server is located;
- the user IDs and passwords are kept on a separate directory services server;
- the user is first authenticated and then, in a separate step, authorized;
- the collected data are transferred separately from the collection database to the production database as requested, typically once or twice a day.

16. Before the first in-house developed collection application was used, an Internet/data security audit was undertaken by an outside consultant. Some modifications in the internal arrangements and the technical solution were made. This level of protection is considered sufficient by SF and by the respondents using the web alternative. No respondent has experienced problems or shown a lack of trust in the data security of SF web collections. It must be said, however, that SF does not really know the reasons why certain respondents are not using the web solution. Doubts about data security on the Internet may be among the reasons.

17. Data security is based on standard procedures and rules as well as on technical solutions. Some examples are as follows:
- new user IDs and passwords are given every year (old ones are discarded);
- user IDs and passwords are initially sent in a letter. If the user so requests, they can be re-sent. Only one of them can be sent by email: the other one must be sent in a letter or given over the telephone. The telephone call must be initiated by an employee of SF;
- only a certain number of our staff have the right to handle user ID/password information (typically 2 persons per survey to have back-up).

SITUATION AT THE BEGINNING OF 2005

18. For the seven monthly business surveys and one quarterly business survey with a web option in production, 4 have been in use for more than a year. For these surveys, the share of respondents using the Internet option is 52 per cent in one and more than 70 per cent in the other three surveys.

19. Of the six yearly business surveys, the web option was used in only one last year and the rest are at present in their first round. In the survey used last year, the share of answers via the Internet was 15 per cent, even if the option was offered only after the initial mailing of the paper forms. In one of the first round surveys, the share of web-form answers is likely to go up to 30 per cent. This is a good estimate for the other surveys, too.

20. For the eleven local government surveys with a web option, one comprises an eighty per cent share of web-answers, and the rest over 95 per cent. This is slightly surprising, considering the fact that they are annual surveys – a respondent can not learn the user interface by heart but has to learn it all over again each year.

EXPERIENCES

Success factors

21. As has been seen, web collections have succeeded very well with the local government

and in short-term business surveys. In annual business surveys, only part of the evidence is here because of the slow repetition cycle and the fact that such surveys are started later than the other two groups.

22. The following factors seem to contribute to the success of short-term business surveys:

- short-term surveys have simple forms which are easy and quick to fill in and can be completed in one session. In most cases, the time needed to answer a survey has been greatly reduced, which means a lighter burden;
- the respondents are from large companies. The staff who complete the forms have the technical know-how to use the Internet. The SSL data security solution is also commonplace and requires no extra technical expertise;
- short-term surveys are highly repetitive – so frequent that it is possible to learn them by heart;
- positive feedback has been received from the respondents for feedback statistics, and this has even led to inquiries for more information about the statistics.

23. Concerning the local government, the factors behind the success of their surveys include:

- feedback statistics: a municipality obtains its own statistics from the data which SF has collected from its establishments (such as individual schools);
- some of the data collections directly influence the subsidies the municipality will get from the central government. Accordingly, these data collections are extremely important to them;
- respondents are enthusiastic about using the Internet - some have said it has been FUN to fill in the web forms (as opposed to paper forms);
- for local government finance statistics, pre-filling the form with data from the previous surveys helps the respondent to remember how the figures were arrived at last year. The application logic-like validity checks help the respondent, as does available contextual help.

Development costs

24. The cost of developing web-based applications and running them has dropped by 60-70 percent during the last three years, both in in-house development and in using an outside service provider.

25. Today, the average investment cost paid to the service provider per collection application is in the order of EUR 5000, which corresponds to around 1.5 months’ costs of a member of the clerical staff. The running costs per year starting in the second year are only one fifth of that. One must bear in mind that, for the most part, the number of respondents is in the hundreds and rarely in the thousands, never in the tens of thousands so far. Of course, there is some work for SF’s own staff in both the development and running phases, even if most of the work is done by the service provider.

26. The in-house solutions are already in the third generation phase. In 2000, the first stand-alone application was developed. In 2002-2003, several new ones were built based on the preceding one. The third phase was the development of the XCola application and its use in the implementation of the collection applications. During the first and second phases of the development, the total resource input was about 2.5 person years (IT Designer, Senior

Statistician) including the development of a secure communication environment. The XCola development took about 1 person year, which included 4 implemented collection applications. Additional implementations are estimated to involve about 150 hours of work per application.

Benefits

27. Savings can be gained in material and mail costs and in manual handling of forms and data. In the four monthly/quarterly statistics[2] in which web collections have been in production for more than a year, the average percentage of work saved in the data collection phase is over 40. It makes a saving of altogether 2 person-years in our case. The amount of ground mail in these four surveys has been reduced by 64,000 per year (65 per cent). This corresponds to the cost of about half a year's work for a clerical employee. In total, cost savings correspond to about 2.5 person-years.

28. The average response time of monthly surveys has been reduced in the best case by 7–8 days or 30 per cent. The time benefit is carried over to the next phase of processing, too. For short-term statistics, the time saved in mail delivery is especially important in reminders. In the best case, the number of reminders has gone down by half. In annual business surveys, there is so far no clear evidence of faster response by the respondents. Annex 2 provides detailed information on one of the monthly surveys.

29. The data received are of better quality. "25 per cent less errors" is a common estimate, even if it has not been substantiated with a proper study. This result is true for both the monthly and the annual surveys.

30. For local government finance statistics, it is estimated that a reduction of 25 per cent of work in correcting errors has been achieved because the data received have been less erroneous. In the expert opinion of the person in charge, the quality of the data has also improved in other ways, both technically and in validity. There is no noticeable improvement in the response times.

31. For education statistics, comparing old and new ways of collecting data is very difficult. This is because the contents change when moved to a web-based solution. In one case particularly, the person responsible reported that it would have been impossible to enlarge the contents without using the web form solution at the same time.

32. As manual handling diminishes, it can be replaced by more rewarding tasks in the statistical office.

33. Some benefits for the respondents are:

- to be able to complete the questionnaire rapidly;
- in some cases, pre-filling can help remind respondents of how they answered previously;
- validity checks prevent the sending of erroneous data and additional inquiries as a result of errors;
- in complex forms where the same piece of information appears in more than one place, it needs to be entered only once;
- at least initially, people like to use the Internet because it is a modern way of doing things;
- many respondents have reported that response burden has gone down.

Benefits/costs in summary

34. Concerning the in-house application development, it appears that the investment in web collection, measured purely in euros, has paid off in about a year in the case of short-term statistics. Improvement in the quality of the data, the reduced collection time, the higher quality of work of the clerical staff and the reduction of the perceived response burden come as extra benefits.

35. Overall, it seems that web form data collections have paid off the investment in under two years in all areas of appliance. The fact that the paper form option must be maintained for the foreseeable future is the one thing that diminishes the potential benefits.

DESIGN AND OTHER CONSIDERATIONS

36. The in-house developed application XCola with its "one application, one database" provides the opportunity to use one system for tasks where, previously, each statistical area had its own solution. This is a step in the right direction, because maintenance and running costs can be cut.

37. There has been no fundamental change in the division of work between the departments of SF. The basic principle is that a team of statisticians and clerical workers is responsible for all phases of production of one or more statistical products. These teams are supported by ADP experts and experts in the publication area.

38. In the experience of SF, no particular pitfalls have emerged. Sound application development methods and care in designing security arrangements have been sufficient. Some design considerations are given below, because they stress the points where SF made errors or could have done better in the past:

- in designing a web-based data collection system, it is not sufficient just to copy a paper form into an electronic form. Because of the limitations of display screens, the layout of the form and the flow of questions must be re-thought. Additionally, advantage should be taken of new possibilities like online validation, context sensitive help, pre-filling and feedback to the respondents. The experience of the (clerical) personnel who have processed the forms should always be taken into account;
- online validation should be as profound as possible. However, checks should be designed in such a way that they do not impede response unnecessarily;
- in testing a data collection application, it is useful to use a small pilot group of respondents (in addition to own in-house testing). For monthly surveys, a convenient period of testing is 2 to 3 months, and in quarterly surveys, 1 quarter. In less repetitive surveys, testing with respondents is not viable;
- the persons administering user IDs and passwords must be trained properly. Special attention should be placed on the design of the administration of user IDs, because each month there will be 1 to 3 per cent of the respondents who displace their user IDs and passwords;
- the collection application should allow work to stop and continue later on. When maintaining the collection application and closing it down for the duration, care must be taken that no respondent loses the data already keyed in;

- when mass-mailing messages to respondents, care should be taken that the mailing does not clutter the email server. It should be ensured that all messages have been sent, for example by sending the last message to oneself;
- in the case of an outside operator, all good procedures for a provider–client relationship should be established. Among these are the rules of communication and cooperation, and receipt and acknowledgement of maintenance tasks.

CONCLUSIONS

39. As has been described above, it is possible, by using a service provider or developing an in-house solution, to accomplish cost-effective data collection applications using the Internet. It would appear that the development costs of both alternatives are about the same and they can be recovered quite quickly, in a year or two. For highly repetitive, i.e. monthly or quarterly, surveys, an in–house solution offers a more robust solution – self-reliance in time-critical applications. These same surveys seem to profit most from the Internet reporting option.

40. There is a consensus at SF that the quality of the data received via the Internet is better than that of the data received on paper forms. This is true for all types of surveys. It also seems that respondents are in favour of the new technologies. It is important to take advantage of this attitude in the constant struggle for respondents' goodwill.

41. It is possible to satisfy confidentiality and data security requirements without undue cost. The security level of our solutions on the Internet is the same as with any advanced Internet services that presuppose the use of electronic money, and the internal security is assured through careful planning, common sense and some advice from a telecommunication security consultant. SF did not experience any special infrastructure requirements in addition to those which in any case had to be met in order to protect SF's computing environment from the evils of the Internet. As a matter of fact, developing data collection applications based on the web increased SF’s understanding of the Internet security technology.

42. A clear strategy is needed for good results. A target can only be reached once it has been set.

[1] Measured in “statistical units” times “number of variables”.

[2] Monthly industrial production survey, monthly turnover survey, monthly construction cost survey, quarterly survey of inventories in industry and trade. These have a total of about 3300 respondents per month and 800 per quarter

ANNEX 1

The XCola application, database and segment structure

1. XCola is an application engine for web-based surveys. In Xcola, the questionnaires are defined as XML documents, and are then customised for each respondent and transformed into web pages at runtime. All this is implemented in a very generic way that would suit for most surveys without additional programming, etc. The application is written using the visual.net (microsoft) and MS SQL server.

2. The collection database, where the data input by the respondent are stored, is at SF. It resides on a server that is connected to a separate segment of the SF local area network (LAN). The segment is separated from the Internet by a firewall and from the rest of the SF LAN by another firewall. The database software is the MS SQL server.

3. Information on respondents, such as names and addresses, resides in the same MS SQL server database as the collection database.

4. User IDs and passwords are kept on a directory services server in the same segment where the collection database is located.

5. User IDs and passwords are stored on a separate directory services server. When the respondent connects to surveys, he/she is first authenticated (let in through login) and then authorized. In the authorization, XCola decides what the role of the respondent is in the survey and then customizes the session accordingly.

6. The XCola application resides on a separate server in a separate segment. These two segments are connected with a firewall. No machine on these segments can initiate a connection to the Internet, and the segment with the collection database can only accept messages from the segment where XCola resides that, in turn, can accept messages from outside.

7. The XML format used in XCola is a refinement of XML4DR, and also includes some parts of W3C's XForms definition.

8. Technically, the user interface is generated as an ASP.NET application, which is sent over the Internet to the respondent's workstation. The XML document defining the questionnaire does not include any user interface specific code. Instead, all aspects of the user interface are handled by the XCola engine and ASP.NET.

9. When generating the user interface, XCola uses information from the XML form, from the collection database and respondent database. From the latter it is possible, for instance, to deduce which rows to put on the form (e.g. on which commodities the respondent provides data to SF). In addition to generating the user interface (the form and its format), the XCola can pre-fill questionnaires with information specific to the respondent in question. Examples are data from a preceding period (periods) for reference and for potential correction.

10. The XCola engine also generates application logic to check the data the user inputs, it generates lists (e.g. classifications like commodities or country lists/codes) and it generates

help.

11. The user can choose from 3 languages: Finnish, Swedish and English.

12. An additional functionality is individualised feedback to each respondent. For instance, the respondent may be shown his/her own figures for the previous period compared with those of the kind of activity the respondent is engaged in.

13. Reminders are sent as email messages. There is a Mass eMailer application to which reminder requests can be sent from an application (recipients, the message). The Mass eMailer then sends the individual emails with appropriate intervals so as not to burden the email server too much.

ANNEX 2

Case study

Case study: monthly sales (turnover) survey, 2000 enterprises, 6 variables:

- in 4 months after its introduction in May 2003, the web-form alternative was used by 70 per cent of the respondents (close to 80 per cent today);
- working hours per year spent in the data collection/data imputing phase were reduced from 2,500 hours in 2002 to 1,500 hours in 2004. This number is still declining;
- the percentage of answers received by the due date has risen from 49 to 76 per cent (at the same time, the due date was brought forward from 20 to 17 days after the end of the month) and the number of reminders have gone down from 1,000 to 500 (from 51 to 24 per cent);
- due to a request by some respondents, the web-application is nowadays opened for respondents on the 1st of the next month instead of the 5th – some of them are anxious to answer!
- the average response time has been reduced by 7–8 days as well as the time needed to get the survey data ready for the next phase of processing;
- the quality of data has improved. For instance, "reasons" are requested if the change from the previous month/year is very large and this facilitates the making of necessary adjustments;
- at least some respondents report a reduction in the response burden (as so many respondents use the system, they must consider it more effective than the paper form);
- the personnel involved in data collection have been happy to do away with everyday tasks. Manual data handling has been reduced by 50 per cent;
- this was the first of the second generation applications developed in-house. Many features/technical solutions were renewed and it took about 850 hours of work (120 days). The work done was paid back in a little more than a year. Maintenance has taken about 100 hours per year;
- with Xcola, this application would need only about 150 hours of work and maintenance would be reduced by half.

* * * * *