

The Experiences of Web Based Data Collection from Enterprises in Finland

Mr. Rami Peltola
Statistics Finland

Abstract

Advancement of web based data collection has been a strategy of Statistics Finland for the last ten years. A couple of years ago Statistics Finland set a target to offer an electronic option in all data collections from enterprises by the end of 2007.

The development of electronic data reporting has targeted towards not only better customer service in the public sector, but also improvement of cost-efficiency, accuracy of data and timeliness. That said electronic data reporting serves many of the goals now topical in developing statistics in Europe.

Most web surveys are quite user interactive. The respondent usually gets to see a web questionnaire that is customised and provides at least some validation routines. Implementing this kind of system is easy when using traditional web programming techniques. The problem is that it is quite difficult to create such a system so flexible that it can handle different types of data collections without the need for additional programming.

A generic server application has been developed in Statistics Finland which can be used for most of web surveys, typically without the need of any programming. This generic application is heavily XML based and uses XML for the description language of the questionnaires and survey logic. The power of the approach is the way XML and the related techniques are used throughout the application.

Keywords: data collection, electronic data reporting, XML, cost-efficiency, accuracy, timeliness, response burden

1. Background of Developing Electronic Data Reporting in Statistics Finland

The production process of statistics in Statistics Finland has been further developed in a major project called "Production model". One of its subprojects concentrated in developing data collection process. This accelerated the progress of developing electronic data reporting considerably during the last years.

1.1 Strategies towards Electronic Data Reporting

Advancement of electronic data collection has been a strategy of Statistics Finland since the 1970s. The first application using web forms was made in 1997, but only in year 2001 progress accelerated. At that time, Statistics Finland set a target to offer an electronic option in all data collections by 2007 (collection from municipalities and enterprises). However it's always the respondent's choice whether to use electronic data reporting or not.

The strategic target was also agreed with the Ministry of Finance, which decides of Statistics Finland's funding. This target does not include person statistics, even though many of them already offer an electronic alternative to paper forms. And of course computer assisted interviews are used as a main collection method in person statistics.

Over 99% of incoming data to Statistics Finland arrive in electronic format. It can be explained by the fact that Finland is a promised land of administrative data sources. The Finnish Statistics Act decrees that existing data must be used for statistics in preference to direct data collection. Statistics Finland has a right and even an obligation to use registers prior to direct collection. 97% of data are derived from administrative registers.

About 3% of data comes from direct data collection. Of this 3 per cent 2,5 percentage consists of some kind of electronic data. The share of data collected on paper forms is less than half a per cent. Some collection in paper form will remain in the foreseeable future also. This fact somewhat diminishes the potential benefits of electronic data reporting. The paper option remains for many person statistics and also for all those enterprises, that are not willing to use electronic data reporting.

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. 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. There is no substituting administrative data available for most surveys. It must be also mentioned that the

administrative registers are mainly of very high quality in Finland.

Many statistics compile these two types of data (administrative and directly collected), for example, in the population census, income distribution statistics and structural business statistics. For those business statistics where some 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.

There are about 50 surveys to enterprises in Statistics Finland (excluding collections with less than 30 respondents). On July 2006 there were 45 web collections in use and 5 more were under development or in testing phase. By the end of year 2006 all 50 surveys will have an electronic option in use.

1.2 The Finnish Infrastructure for Electronic Data Reporting

The situation is very favourable in Finland for using and developing electronic data reporting. This is the case especially for data collection from enterprises and local government.

In Finland response rates have traditionally been very high in business surveys (in both annual and sub-annual surveys). In the best surveys response rates have remained up to over 99%, for example in the collection of monthly volume index of industrial production. Maybe one of the reasons is the persistent and friendly staff in Statistics Finland. Statistics Finland has good relations with data providers, in many cases continuous personal contacts and it has been possible to remain very high level of trust in data security. It will also be a challenge for electronic data reporting to maintain the high response rates and to invest in data security.

Access to the Internet is very common in Finland and almost every enterprise has access to Internet. In enterprises with employees more than 10 the access rate to Internet is around 98% and in enterprises with more than 100 employees 100% of businesses have access. Because business surveys are mainly aimed at the largest enterprises, the readiness to use electronic data reporting is extremely good. It can be said that there is a positive atmosphere for using the Internet in transactions with the government. Some respondents are even enthusiastic about using the Internet, which makes it feel less burdening to answer to surveys.

2. Electronic Data Reporting Solutions Developed or Used in Statistics Finland

At Statistics Finland there are two different kind of solutions of web based data collection in use: either solutions developed by an outside service provider or in-house developed applications. By the end of July 2006 Statistics Finland has 30 in-house applications and 11 applications by an outside service provider. In addition there are around 5 electronic data collections in fixed format Excel file.

2.1 The Three Generations of In-house Electronic Data Reporting Solutions

During last few years Statistics Finland has already gone through three generations of in-house applications. The first generation started in 2001 with development of individual electronic data reporting system for building cost index. It was built using Microsoft Windows DNA (Distributed Internet Application Architecture). All the first generation applications have been renewed to third generation applications.

During the second generation using VB.NET technology 7 EDR solutions were built in 2002-2005. This new generation offered some new features, such as individually tailored feedback to respondents. The second generation applications already aimed towards a general framework of electronic data reporting solutions.

The third generation applications have been built since 2005 and there are already (by the end of July 2006) 23 new electronic data reporting solutions using XCola - format.

The outside service provider has been used during lack of resources, and also in the starting phase, when Statistics Finland has not yet acquired enough knowledge in-house. It has been used especially in annual data collection, where the amount of collected variables is large. Earlier the price of outside service provider was rather competitive. Now that the prices of outside service has risen remarkably and Statistics Finland has developed a general application, which enables very cost-efficient development in-house, the outside providers are rarely used in business surveys.

Statistics Finland has also piloted an integrated data collection during summer - autumn 2005. Integrated data collection means here data collected straight from enterprise's own management systems. This will be explained with more details in the forthcoming chapters.

2.2 What is The XCola -Data Collection Solution?

The name XCola is derived from XML based collection application. Xcola is a generic application for web surveys - this is the answer to the demand: "one application fits all". Previously each statistical area had its own solution for Electronic Data Collection. The questionnaires are defined as XML documents, and are then customized for each respondent and transformed into web pages at runtime.

XCola is executed on the server side and it does not require any installation on the respondent's side. It works on every modern browser. The application supports client and server side validations of entered data (such as checks for logical errors and comparison to previous data). The application is extremely simple to use for a skilful developer and new questionnaires can be implemented in just hours depending on the length of the questionnaire. The XCola also enables sending secured Excel files.

3. Main Benefits of Electronic Data Reporting

The development of electronic data reporting has targeted towards not only better customer service in the public sector, but also improvement of cost-efficiency, accuracy of data and timeliness. That said electronic data reporting serves many of the goals now topical in developing statistics in Europe.

Cost-efficiency can be improved by simplifying data collection process by means of automation, reducing need for human resources and by reducing other data collection costs, for example the amount of ground mail or printing costs of paper forms.

Some experiences have been gained on improved data accuracy and also decrease of non-response. It has been seen that, data accumulation may speed up with the use of Internet, for the simple reason, that it takes a few days extra to use ground mail.

Regarding customer service electronic data reporting can help nurturing data provider relations. Many respondents have reported a sense of reduce in response burden after starting to use electronic data reporting. However, it has not been systematically measured. Also the enabled direct individual feedback for respondents has evoked very positive comments. For example in the monthly sale inquiry businesses can compare their own development in sales to other businesses in their branch. Browsing of previously submitted data helps the respondents to answer the questionnaires. Assuring high level data security is of course a requirement of maintaining good data provider relations.

3.1 Achieved Cost-Efficiency During The 2nd Generation Solutions

Four second generation solutions have now been in production for 3 years. They have all together 3300 respondents per month plus 800 per quarter. Average per cent of work saved in the data collection phase is over 40 in these data collections. This adds up to 2 person years all together. The amount of ground mail has been reduced by 64 000 or 65%. That corresponds cost of 0.5 person years.

While the average response time has reduced, the number of reminders sent has gone down by half. A "mass e-mailer" is used for all kinds of collections and it has reduced remarkably the workload of sending reminders. The investment of developing one second generation solution has paid off in about a year. Improvement in the quality of the data, the reduced collection time, the higher quality of work of staff and the reduction of the perceived response burden come as extra benefits. These second generation solutions have not been updated to the third generation and therefore there is a need for individual maintenance in these data collections.

3.2 Cost-Efficiency Continues to Improve During The 3rd Generation Solutions of Today

The third generation brought a common framework (one engine) for similar data collection systems to be built up and maintained. This created a very effective build-up phase.

The XCola includes also a simple method for transferring data between collection and production data-bases, so that manual work is not necessary. There is only one application to maintain and support the different data collections, which saves a substantial amount of work after development phase also. This makes support and development knowledge easier to acquire and spread.

The important goal achieved are the benefits for the personnel at Statistics Finland. While reducing need for human resources as manual handling diminishes, more rewarding tasks can be offered to people working with data collection. They have more time to deal with data provider relations, and the work with electronic data reporting and in respondents electronic data reporting support alone has been very rewarding to many workers at Statistics Finland. Also more challenging and varied tasks have been introduced while manual handling is disappearing in data collection process.

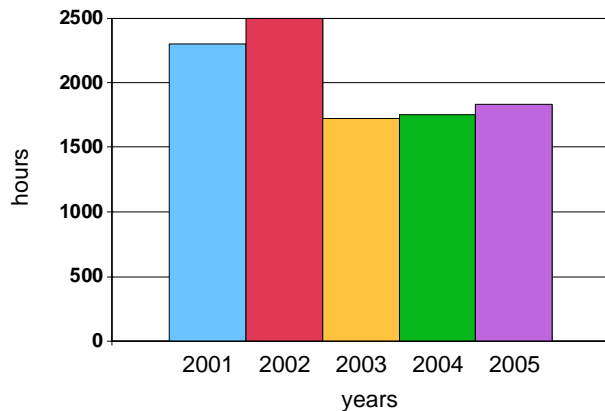


Figure 1. Working Hours Used in Data Collection and Validation in Sale Inquiry

This figure clearly shows that the amount of time used in day-to-day data collection has dropped in monthly sale inquiry after the introduction of electronic data reporting in the beginning of 2003.

The amount of work has gradually risen a bit. One reason is the faster European sample of retail trade (30 days lag in publishing), which was started in June 2003. It has required two phone reminders to businesses that have not yet answered. Also the amount of time used in validation of data has increased - some work has shifted from calculation of statistics to the data collection phase because of the speeding up of statistics production. At the time from the start of 2003 to today the publication of monthly statistics made from this data has accelerated by 15 days.

3.3 Benefits to Accuracy and Timeliness

The data received through electronic data collection are mostly 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 monthly and annual surveys. The automatic validations help respondents in avoiding errors.

Response rates have remained on a high level. The average response time of monthly surveys has reduced in the best case by 8-10 days or 30%. The always available questionnaire makes it possible to answer before the due date. The opening of a new month is immediately informed to respondents by e-mail. While the number of reminders sent has decreased in half, the data is usable earlier than before electronic data reporting. The share of the respondents using electronic data reporting solution has in most cases reached high level. In sub-annual surveys around 60% (in the best case 99%) of respondents use electronic data reporting and in annual surveys around 30% (in the best case 75%).

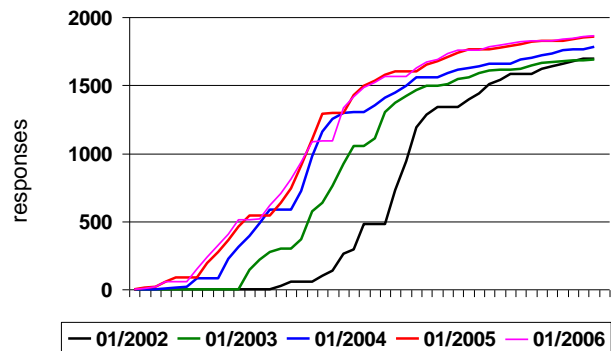


Figure 2. Accumulation of Data in the Sale Inquiry.

According to figure 2, the accumulation of data has accelerated after introducing electronic data reporting. The data accumulation has accelerated by 8-10 days compared to January 2002. The electronic data reporting was introduced in January 2003. The first month of using electronic data reporting immediately accelerated data accumulation by 5-6 days. In a year the present speed of data accumulation was achieved and it has also been maintained.

3.4 Benefits to Data Provider Relations

The perceived response burden has gone down. A lot of feed-back of decreased burden have been received, even if in many firms the data needs to be collected from different sources and written down on a paper sheet and only after that entered into the internet form, checked and sent. The only explanations are that answering is not tied to a paper form, but it is possible to answer anywhere with internet connection. Also there is not a risk of losing your paper form and not the effort of mailing or faxing it. Sometimes the fax numbers may have got jammed in spite of constant follow-up. The validity checks, online help as well as multilingual support of web-based data collection make it more comfortable to fill-in the form. Additional inquiries are rarely needed by Statistics Finland. The immediate, individual feed back may be one of the perceived benefits.

It also makes it easier to answer, when e-mail informs of the survey, reminds to answer and supplies a direct link to the questionnaire. E-mail reminders has also been experienced less annoying than phone or letter reminders. The questionnaire is "always" available and fast to fill-in. It can also be filled in separate sessions. All the questionnaires have been tried to design as simple and easy as possible. Access to all the previously submitted data and pre-filled questionnaires do make the impression of us not wanting to burden the respondent with anything extra.

4. Reassuring High Level of Data Security

Data security is a very important and sensitive issue to us at statistical offices and to data providers. In Statistics Finland a data security audit was done by an outside consult, when Statistics Finland developed its second generation applications. The audit gave Statistics Finland an excellent evaluation of data security. For data security reasons there are for example separated production and collection data bases. These are connected through firewalls both towards the Internet and towards the rest of the Statistics Finland LAN (Local Area Network). Regarding outside service providers, they are responsible of data security in their applications.

All data collection traffic on the Internet is SSL - encrypted. An authentication / authorisation -process is always needed. New user IDs and passwords are given every year. User IDs and passwords are initially sent in a letter. Ground mail is considered to have high level of security in Finland. Only one password or user ID can be sent by email, the other one must always be sent in a letter or be given over by telephone. Only a few of Statistics Finland's staff have access to user IDs and passwords (usually two persons per survey). Only exceptionally has came up a respondent, that has shown a lack of trust in the data security of Statistics Finland web collections. A bit funny example of distrust towards the Internet was one respondent which rather sent the data to Statistics Finland attached to an e-mail, which uses an unsecured connection.

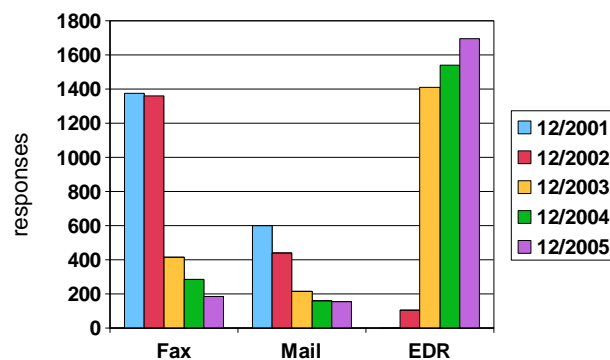


Figure 3. Change in Response Media During years 2001-2005 in Sale Inquiry.

The data providers have welcomed electronic data reporting and have trusted it to be secure. The figure 3 gives an example of sale inquiry, how the response media has shifted in just a few years, almost immediately after introducing electronic data reporting. Whereas 66 % of respondent used fax in 2001 and 2002, in 2003 70 per cent of all respondents had shifted to electronic data reporting. There had already been a 30 per cent decrease

in responding through mail from 2001 to 2002, since the first 100 businesses piloted as voluntary pilots of electronic data reporting. Nowadays around 85 per cent of respondents of sale inquiry use electronic data reporting. Furthermore a little less than 10 per cent uses fax and the rest uses mail as a response media.

5. Costs of Investment and Maintenance

The cost of developing web-based applications and running them has dropped by 60-70 percent during the last three years in in-house development. The direction seemed to be the same for the outside providers, until the price of an outside service provider has went up this year.

Average investment cost per new electronic data reporting solution from an outside service provider that have been made until 2005 was around 5 000 euros. Nowadays building up one new data collection by the same outside service provider would cost about 15 000 euros.

This is why Statistics Finland rather applies in-house solutions even with limited development resources. This question of resources has been taken into account by reorganising IT-services in Statistics Finland. IT-services inside Statistics Finland are now organised by the phases of production process of statistics (collection, production, distribution). The XCola development took about 1 person year, which included 4 implemented collection applications. Additional in-house solutions (XCola) require approximately 150 hours of work.

Maintenance costs of EDR solution from an outside service provider were around 1000 euros per year, but have recently risen. The maintenance cost of an in-house solution (XCola) is around 50 hours of work. During the first and second generations the total resource input was about 2,5 person years ("learning by doing") including the development of a secure communication environment. This included the implementation of 7 solutions, so it took around 450 working hours per data collection.

For example the development of sale inquiry required over 800 working hours, since it was one of the first second generation electronic data reporting -solutions and many general methods were developed (including secure communication environment). The next second generation solutions required much less work, as the average time per 7 applications was 450 hours.

As figure 4 describes the time used in maintenance of electronic data reporting solutions has dropped a bit every year. Now the time used to maintenance tasks is around one tenth of the time used to the original development of the solution.

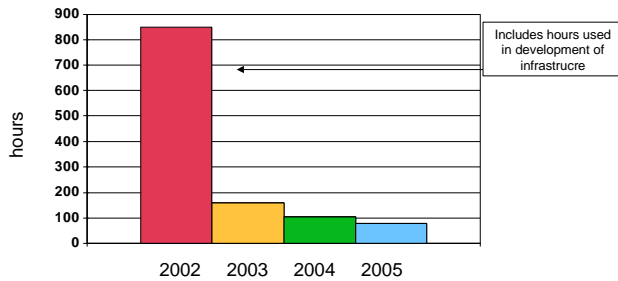


Figure 4. Work Done in Development and Maintenance of An Electronic Data Reporting Solution of Sale Inquiry.

6. Challenges of In-house Development and Maintenance

The development of surveys can be very fast if the IT - personnel have good skills in XML and related techniques. At the moment the number of very skilled survey developers is limited. The whole production environment around XCola is not yet finished. There are only two very technically skilful developers and around ten more are learning by developing their first solutions. In development and maintaining electronic data reporting solutions Statistics Finland is somewhat dependent on certain named persons.

The statistics departments typically have a lot of requirements for electronic data reporting solutions for their surveys. Therefore some minor development in XCola is needed all the time.

7. Future Development Challenges and Directions in Statistics Finland

In integrated data collection data are derived directly from management systems into Statistics Finland's database. This is one of the future directions of developing data collection. Some experiences have already been gained today.

7.1 A Pilot on Integrated Data Collection

Statistics Finland came up with integrated data collection when developing electronic data collection for tourism statistics. The idea is to collect the data directly from Hotels' management systems. Statistics Finland made a decision to check up on the possibility to induce a few software vendors in the sector.

Software vendors implement a module for the hotel's management software using Statistics Finland's definitions for data and service interface. Data collection solution is implemented by using typical business to business integration technique: XML Web Services. No manual work is needed, except to initiate the transfer of

data from a hotel. Data is entered into the enterprise's own system for their own use and the data needed are sent to Statistics Finland by pressing a button. Then the data are submitted to the standard validation process at Statistics Finland.

Integrated data collection will be a major possibility to decrease response burden and increase data quality, because statistics can use the same information as the firm itself. This type of system only fits collections with very clearly assimilated concepts such as the amount of overnight visitors in hotels and other boarding houses etc. Also some amount of centralization of data system providers in the sector in question is needed.

An important point is that the role of statistical office should be to provide definitions and requirements of the data system to the software vendors. National statistical offices should not provide readymade modules for data transmission, otherwise also the responsibility of maintenance would be at national statistical offices.

7.2 Productisation and Integration as Future Challenges

Finally some considerations of the near future (2-4 years) of electronic data reporting at Statistics Finland. There remains the question whether Statistics Finland is able to create more co-operation with data system suppliers in order to increase the amount of integrated data collections. It would mean intensive co-operation and negotiations. These kind of steps should only be taken when cost-efficiency can be reassured.

There is an ongoing project for productisation of XCola (started just at June 2006) at Statistics Finland. Some further developments to XCola application v. 3.1 have already been made: a developer's manual (only in english, about 60 pages), finalised administration tools, routines for transfers between collection and production databases and an XCola version for outside evaluation has been built. This XCola version can be sent to any statistical office interested in testing XCola's adequacy to its' data collections. During the project of productisation of XCola Statistics Finland will develop a graphical editor for building questionnaires and create links to metadata. The final target in developing XCola further is that the definitions for the creation of an XCola solution for a survey can be readymade in the statistic unit in question.

Statistics Finland started a project for co-ordination of business surveys in 2005. The purpose of this project is to create and disseminate a common knowledge of different business surveys. Also a tool is needed to have a control of what different surveys one business has to answer in total. This project started from a data providers initiative

and then from Statistics Finland's director general's decision. Maybe in the more distant future Statistics Finland has more co-ordinated surveys - instead of many independent surveys targeted towards the same businesses. But because of the nature of Finland's economy, dependence of the largest enterprises is high and therefore they will always be the primary target of business surveys.

Acknowledgements

I would like to thank all those colleagues at Statistics Finland who have given me valuable comments and inputs during the preparation of this paper. Special thanks to Mr. Ilkka Hyppönen who has brought together different aspects of electronic data reporting at Statistics Finland by preparing his paper on this topic last year and to Mr. Toni Räikkönen for his contribution concerning technical details of XCola.

References

Mr. Ilkka Hyppönen, "Using The Web In Collecting Data for Business Statistics in Finland", United Nations, Economic and Social Council, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians, Geneva, 13-15 June 2005