Evolution of Electronic Data Reporting at Statistics Canada

Tony LaBillois, Jacqueline Mayda, Statistics Canada
Tony LaBillois, STC, 1-C26 Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6

1. Introduction

The recent but rapid growth in the number of users of the Internet in Canada, the willingness of the government to offer its services on line and the interest of the organisation to offer more flexibility to respondents in completing surveys are some of the factors that have led to the development of strategies for Electronic Data Reporting (EDR) during the last few years. These strategies are still in evolution and they will be for a long time as the technology is evolving very quickly and as the perceptions and habits of the users are also changing over time.

The advent of EDR can be viewed as a natural extension of the use of technology in data collection activities. Whereas Computer Assisted Interviewing (CAI) effectively combined into one step previously distinct operations such as interviewing, data capture and editing, EDR goes a step further by shifting such activities to the respondent. Any EDR option also has to be integrated in the usual collection process with the rest of the methods offered to respondents.

As a national statistical agency with strict policies related to confidentiality and privacy, technological and security-related issues have had to be addressed at each step along the way to providing EDR. This has had an impact on flexibility, user-friendliness and overall acceptance of the EDR tools by the survey respondents. The conflicting challenges of finding a fast and simple EDR solution for the respondents and implementing totally secure approaches have always been at the center of the thinking for viable alternatives.

The focus of the paper will be to provide a brief background on the early development of EDR at Statistics Canada and on the situation today. It will also present some of the issues that we have encountered while deploying EDR as a collection option. The paper will close with a brief description of Canada's Government On-Line initiative as it applies to statistical data collection, and our conclusions.

2. Background

Electronic forms have been available for some surveys since the early 90s at Statistics Canada. Forms were designed with commercial software packages and were then installed on the respondent's workstation to allow the completion of the forms on their computer screen. We were basically developing desktop applications that we were distributing on diskettes. The Internet technology was then first used to allow the respondent to send back the data electronically using email or the File Transfer Protocol (FTP). Security was addressed through encryption algorithms applied to the source data. The Internet technology was also used to allow the respondent to download the desktop application from the Statistics Canada web site instead of receiving it by mail. This approach has been very successful for a very limited number of respondents in a few surveys.

This approach was then further extended to a larger number of respondents of annual economic surveys and we had very limited success. The download time necessary before the respondents could start using the EDR option was very high, there was a lack of flexibility in the software package used to develop the questionnaires and there were many deployment problems due to the various configurations and products installed on the respondents' machines. Each respondent's computer can use different versions of operating systems, different components, different browsers and different modem speeds. Very few respondents took the opportunity to visit the web site offering the EDR option even if they were encouraged to use it in their information package. This fact suggests that the majority of respondents were not ready to use an EDR option for various reasons. We need to identify these reasons and better market our EDR approaches while taking these reasons into account.

The slow adoption of EDR by respondents is observed not only in Canada but also in many other countries. An interesting path for increasing its use is to focus our efforts on repeated surveys (monthly and quarterly) with a small number of questions. Keeping it simple, at least from the respondent's perspective, appears more important now than ever if we want this to be more successful in the future. The technology changes rapidly and it still cost a lot of money and resources to be able to develop and offer good and secure solutions for EDR. For now, no cost savings can be expected in the near future and many statistical agencies are facing the same situation. To save costs, it would require a situation where a sufficient proportion of respondents use EDR without the need for telephone follow-up and in which the technological solution last for long enough without a need for redesign.

3. Current and future strategies for EDR

We have learned a lot from our first experiences. We now try to look at all technological and non-technological aspects of EDR with our limited but rapidly increasing experience. We stay realistic about what EDR can do in the short term and we explain the situation to the various people involved so that they have reasonable expectations. We try to base our strategies on standards instead of proprietary software. We are also starting to do more in-depth research into the needs of our respondents and on the barriers that they face before using EDR.

The two current strategies are still based on the most secure approach possible, which requires encryption at source. Because of that, they all use the processing power of the respondent's machine and they require that the respondent download software from a web site or a CD-ROM. Software doesn't mean a complex package here but rather refers to either simple tools to help view and process the questionnaire or to encrypt the data and send back the file(s).

3.1 Data Return Facility with or without a standardized questionnaire

The first approach consists of expanding on the existing data return function build into the desktop applications to receive survey data securely through encryption algorithms applied to the data on the respondents' workstation. The survey data can be sent by a Data Return Facility in several formats including text files and spreadsheets. The respondent has to download and install the module on the workstation. He can then use the module to encrypt and send back a completed questionnaire developed by Statistics Canada in a software he already has (e.g. Excel) or he can also send any other files to us if requested. The module can send the encrypted files back using the File Transfer Protocol (FTP) or email if FTP is not available from that respondent's machine. This simple approach is using the Internet only for transferring the information.

This method is used for an increasing number of economic surveys with a standardized questionnaire. Some examples are the Survey of Suppliers of Business Financing, the Computer Services and Internet Service Providers Surveys, and the Canadian Automated Export Declarations. It is also used without a standardized questionnaire for several large companies (key providers) who send data to more than one Statistics Canada division. Most of the survey managers choosing this method are also providing CD-ROMs to eliminate the issue of the download time.

3.2 Web-based questionnaire

The second method of EDR consists of an interactive questionnaire on the Internet. It involves the creation of our own web-based infrastructure to allow the respondent to complete a questionnaire over the Internet using Secure Socket Layer connections and encryption at source. Under this option, the respondent is required to download and install a small customized browser and the Java Runtime environment before the survey process can begin. The respondent then accesses the survey HTML pages on the Statistics Canada web server. No respondent data is transferred across the Internet while the survey is being completed, simply HTML pages, containing the instructions and questions, and Java scripts, for editing and navigation.

3.3 Automated survey response

So far, when we are offering an electronic questionnaire, the respondents have to enter their information into the form appearing on their computer screen even though it could be already available electronically in their own systems. For businesses, the development by the industry of a common language to map and extract information directly from someone else's systems is still in the early stages. The Canadian Institute of Chartered Accountants is currently working with international partners to develop the eXtensible Business Reporting Language (XBRL) to allow the exchange of financial information in many different software packages. Even if it would reduce the burden of respondents to report in their own way, it has many implications in the processing operations as well as in the consistency of the results. When we will be technically able to offer this type of solution, we will have to carefully consider the possible impact on consistency and quality of the information and we will have to find ways to minimize these impacts. Statistics Canada is following closely the progress of XBRL development and intend to co-operate with the various participants to try to incorporate the functionality for survey reporting in the context of this new eXtensible Mark-up Language (XML).

4. Issues we have encountered while deploying EDR as a collection option

In the past two years, Statistics Canada has placed an emphasis on including EDR as a data collection option for several Business and Agricultural surveys. Both the Data Return Facility and web-based solutions have been implemented with success in many collection operations. We are finding that with each implementation of a new EDR application we learn some new lessons. They serve to continuously remind

us that with more reliance on computers and consequently less on human intervention, we can be at technology's mercy. This section of the paper describes some of the issues we have encountered while introducing EDR as an option for several surveys.

4.1 Issues specific to Data Return Facility option

This method is used for an increasing number of economic and institutional surveys with a standardized questionnaire or file format. Since a majority of business respondents are comfortable with using Microsoft Excel, many survey areas are choosing to use Excel to develop their questionnaire and either make it appear like a questionnaire, or a spreadsheet. Once completed, the file is then sent to Statistics Canada using the Data Return Facility, and forwarded to the survey area for further processing. This has proven to be a good choice, as the respondents do not need to learn how to use other software, they are comfortable with the tool and, in some cases, they can even link or copy the information from their usual files to complete the survey. Editing can be available on the Excel form to simplify processing down the line. However, the disadvantage of using Excel as opposed to a custom-developed application (such as in Visual Basic) is that you are limited to the functions of Excel itself. That being said, such software already has built-in features that would take time to develop, such as sorting and printing facilities.

The Data Return Facility is also being used without a standardized questionnaire for several large companies (key providers) who send data to more than one Statistics Canada division. This facility allows the respondent to send whatever kind of file they have, whenever their time frame permits. Most of the survey managers choosing this method are providing CD-ROMs with the software to eliminate the issue of download time for respondents. This is a good approach for a select number of respondents, as it is the most accommodating to them, and reduces respondent burden. The downside when no standardized questionnaire is used is that it is relatively complex to convert the information when it arrives at Statistics Canada. We are taxed with the burden of entering it or converting it in the common format of the regular collection vehicle for the survey. For this reason, we try to encourage standardized formats in the majority of cases.

We have made every effort to ensure that the process of transferring a file using the Data Return Facility is fully automated once the data are sent to STC. The file decryption, transfers, reports and other related processes are all performed using timed scripts, which we feel is the only way to go. Any manual intervention in this process lessens the efficiency of transmitting data electronically.

The vast majority of our feedback on this method is very positive. Respondents feel this is a simple way to send their data securely to Statistics Canada. Offering this software via the CD-ROM seems to be increasing more than the web-site perhaps due to the time involved in downloading. It seems the number of survey areas requesting this option increases every day. However, some data collection operations are more suited to a more interactive approach, thus we continue to pursue the web-based model as an alternative option.

4.2. Issues specific to the Web-based model

This model was developed for the 2001 Census test (for Population and Agriculture), but has since been used as an option for two quarterly surveys: one business and the other agricultural. For the census, 5% of those offered utilized this option and this was felt to be a good take-up rate. The quarterly surveys have found that between 5 and 25% of those respondents offered the option actually use it. This rate has increased each subsequent quarter the EDR option has been available.

In our experience, the use of a customized browser and Java standards practically eliminates the issues related to variations in the environment and the tools (operating systems, browser versions and types, service packs, etc) already present (or absent) on the respondents' machines. The download and installation occurs the first time the respondent is adopting EDR, and is not necessary for subsequent occasions unless an update to the browser is necessary. This is advantageous for respondents who may want to reply to more than one Statistics Canada survey electronically.

Other positive advantages of the web-based option include the ability for questionnaire navigation, built-in skip patterns, cross-page edits, and a final page edit feature. This feature checks the contents of the HTML form before submission (on the final page) to ensure that at least some data has been entered on the form. If not, a message pops-up to the respondent prompting them to complete at least the key fields. We have adopted an unwritten policy that no edits should be mandatory on the web-based form. That is, the respondent always has the option to ignore an edit message and continue to submit their data.

The customized browser also allows for multiple languages (essential in our country), and the facility for multiple sessions, with save and restore features. This ensures that if the respondent cannot complete the form in one session, or more than one person's input is required, the software will accommodate the respondents needs. We have also found that the use of individualized passwords and confirmation codes upon receipt of the data serve to make the respondents more at ease with the security of their data transmission. In terms of data reception, we have implemented a generalized XML model that allows for one structure or format for receiving the data. This gives us flexibility in terms of easy implementation for multiple surveys using the same browser. The use of a database to receive the incoming data allows for easier reporting and quick integration with subsequent processes such as follow-up using a CATI vehicle.

Although we have cited many positive aspects of our web-based model, we have incurred distinct disadvantages to this approach. As the software (Java Runtime and browser) is approximately 6.4 MB, at the moment the download time is an obstacle in some circumstances, such as respondents with slower modems, less up-to-date equipment. This can be reduced by offering a CD-ROM with the installation, but may not be cost-efficient in circumstances where a small proportion of a large sample will use it. The CD-ROM approach also alleviates the updating, testing and maintenance involved in using a web-site as the source of the collection software. Respondents who are not willing to give up space on their hard drive to satisfy STC have also identified the size of the download as a concern. For one survey, 50% of people who hit the web-site chose not to download after having read the download instructions. Our feeling is that as the availability and use of high-speed Internet connection increases, the download issue will become less significant.

We have also found that the maintenance aspects of a customized browser must not be under-estimated. One must have the resources to support the software, to be able to update it and to make sure it is well documented. Another downside to the customized browser is that from time to time updates must be sent to the respondents if the browser changes. This has an impact as the onus is on the respondent to uninstall the old version and install the new one.

Although cited as an advantage to the web-based model, the database on the receiving end of the data must be set up properly and maintained. This involves different knowledge than the development of the EDR collection tools themselves, such as database administration.

4.3 General issues involved in the introduction of EDR as a collection option

Computer-related: Installation instructions must be clear enough to allow a novice user to utilize them, and not so complicated as to scare a respondent away from using the product. That being said, we have many examples of respondents ignoring the instructions provided and trying to figure things out themselves. For example, detailed information is provided on the web-site related to the application features and survey-specific help for each of our surveys using the web-based model as an option. For one survey occasion, only 25/1000 hits were to those pages.

We have found that in the EDR environment, software is being developed for use in many survey areas and for various purposes. Under these circumstances, it is important to think of generic programming, such as using functions that may be replicated in different surveys. We are in the process of developing a standardized way to receive specifications from the client for EDR applications in terms of edits, navigation, look and feel, data capture and output file format. We have also deemed it necessary for developers to have a common directory structure, flow diagrams, backup procedures, and log files of changes made to programs to allow backtracking while trouble shooting.

Testing, testing and more testing. Our experience has shown that the more eyes that see the product, the more problems can be identified. We stress internal testing in development, testing in conjunction with the client, full end-to-end simulations with test data, and including the involvement of anyone else within the organization who would like to try an interesting new product. We also ensure that the data are checked once received through the EDR structure to ensure they have been transmitted correctly and that all automated processes are working as planned.

The organization's ability to maintain the hardware, software and infrastructure involved in EDR development in processing should not be underestimated. Continual updates, changing products and volume of responses are factors that should not be ignored, as well as the costs related to them.

Respondent relations: The updating of contact information obtained through EDR collection methods must be addressed within the agency. We have found

that whomever sends the emails announcing the availability of the EDR option is the person who is provided with changes to contact information, email addresses, etc. If this is a developer rather than a subject matter officer or operations specialist, this information must then be relayed to the persons responsible. All of the usual concerns connected with interactions with the respondents must also be addressed for EDR respondents, although the issues are sometimes different and may be conveyed in a different manner.

Editing: The question of editing responses given by EDR respondents has also required careful thinking. We are hesitant to program the same edits that we usually apply to responses in other methods (e.g. CATI, CAPI) in the fear that we might irritate the respondent if edits are triggered too often automatically. However, we have had experiences where we have removed many of the edits in the EDR vehicle, only to have the CATI follow-up be performed due to too many inconsistent values. Obviously the balance must be somewhere in the middle, and we work with each survey area individually to assess and address their needs.

Currently we do not have historical edits built in to the majority of our EDR applications as our security requirements restrict us in how we send data back to respondents. The data must be encrypted before sending, the channel must be secure, and we must ensure that the recipient is the person we intended to send the data to. This is obviously a limitation for repeated surveys where historical edits are a key. To address this issue we have in some instances sent password protected, encrypted diskettes to respondents separately from the software itself. In other cases we rely on the CATI collection vehicle used for telephone follow-up to perform historical edits. Although this is not the way we would like to address this in the long-term, it satisfies our security requirements at the moment without placing too much burden on respondents.

Integration of EDR with usual collection processes:

The effect of the introduction of an EDR option must be considered at every step along the way. From a mailout package contents, to changes to the questionnaire and frames to accommodate email addresses, to the impact on the CATI collection vehicle flows. We must think of the influence that EDR has on the front- and back-ends of data collection, reports and processing down the line. Many survey areas are not willing to immediately change their back-end processing systems (E&I, estimation, etc.) to adjust to EDR returns – even if it is to identify that a record was received by EDR versus telephone or paper.

We have also found that the development (systems) area has the most expertise in recognizing the factors affecting the introduction of EDR to an existing survey program – and must play a key role in such a plan. This is a new experience for everyone, but especially for the collection operations area, methodology and subject matter specialists.

There is a possibility of cost savings over time for repeated surveys if a sufficient number of respondents use EDR for the number of periods that they stay in the sample. When dealing with the same respondents over time, there is also more potential for marketing electronic data collection. The identification and telephone follow-up of non-respondents in a very timely manner is key here, as is integration with the existing data collection tools. With EDR, each respondent sends back a separate file containing the survey responses. A generic processing system to combine all of the EDR responses to various surveys needs to be developed. In a repeated survey with short collection periods, cases will have to be transferred efficiently from one collection method to another and adaptation to our current collection systems also has to occur to allow that to happen. Cases will also have to be easily transferred from one collection method to another over time as respondents change their preferences. The people responsible for telephone follow-up need to have timely information on completed or not completed cases in order to maintain good relations with respondents and to focus on the most important cases.

Research: To date, we have not focused on the evaluation of the impact of EDR in conjunction with other methods on data quality. Typically the EDR vehicle is a duplicate of the paper version of the questionnaire which allows for a direct comparison of the results and makes it less essential to study variation. We do have plans to study areas such as mode effects, self-enumeration studies, etc. Time and resources will be dedicated for this activity especially on the social/household surveys side. It appears to us that certain types of surveys would not necessarily benefit from EDR. For example, surveys where the role of the interviewer is considered very important for the quality of the data do not appear to be good candidates for self-completed questionnaires. At Statistics Canada, this encompasses a large number of our social, household-based surveys. For the foreseeable future, only research into the use of EDR for these types of surveys is envisioned.

5. Government On-Line

An initiative has been launched in Canada that proposes to offer all key government services on-line by 2004 – commonly referred to as Government On-line. As part of this government-wide project, Statistics Canada has embarked on a journey that will result in the majority of our Business and Agricultural surveys offering an EDR option. A recent survey indicated that 56% of Canadians said they want to use the Internet to report to surveys. 85% of businesses expressed the same interest. By April 2002, 11 monthly business or agriculture surveys will have an electronic data reporting option available in addition to their usual method of collection. The formation of an internal multidisciplinary team to work on the integration of electronic data collection in statistical surveys is the next step in the promotion of EDR as a viable option for data collection. We have noticed very early on in this process that a key to our success will be co-operation from the client sponsoring the survey – to address the development of the new system, separate from the production environment of the existing system. We must also have input from the operations area conducting the data collection as well as methodology input on the impact the introduction of EDR may have on the survey results. What has also become apparent is the importance of the seamless introduction of EDR to the entire process – from reports on progress of collection to data editing and the level of follow-up required, especially for monthly surveys where the turnaround time is virtually days.

The coverage of our largest businesses, which are essential for business surveys to produce high quality economic indicators, is another aspect touched by Government On-line. These businesses complete a large number of surveys of various frequencies and it is thus important to develop collection approaches that will reduce their reporting burden to maintain effective relationships with them. We are already implementing the Data Return Facility for some key respondents. It becomes more and more popular as they need to send back to Statistics Canada all kinds of information in a fast, simple and secure manner. As part of Government on-line, we also have a plan for the implementation of the hardware and software to maintain a secure environment to communicate and transmit information in both directions between a small number of large business respondents and Statistics Canada with Public Key Infrastructure and certificates. In some cases, the two-way communication is required or desirable but it will remain limited for the time being because of priorities and available budget. A personalized reporting web-site for our largest businesses is also part of this project.

6. Conclusion

Our intention is eventually to offer an EDR option for all the appropriate surveys. However, it is necessary to determine the appropriateness and the order of priority for the various surveys. It may depend of the requests from respondents, the cost to add the EDR option, the level of security or the technical difficulties associated with each collection strategies. Some types of surveys would seem less appropriate for EDR collection. Even if we develop EDR options for respondents, the success of the new approaches depends largely on the access and acceptance of new technologies by these respondents and on their willingness to use such mechanisms instead of the other methods to respond to our surveys.

It is important to recognize changes in the technology that can be used for EDR, but it is essential to stay aware of the emerging perceptions and habits of respondents and to be able to adapt quickly to these trends of our society. Respondents and the general public are still in their infancy in their approach for all of these new tools. As the speed of the connections and of the machines improve or as the new versions of operating systems and software packages are being released, we have to consider that not all the potential users of EDR are adopting the state-of-the-art equipment at the same pace. It is always a dilemma for developers to choose between the new and more efficient solution and a solution usable by the greater amount of respondents but not necessarily as convenient to them. The reliability, robustness and flexibility of an EDR solution always have to be taken into account.

The biggest challenges related to EDR are not necessarily technological but also consist of a better understanding of the issues related to security, risks or convenience by the respondents and by the statistical agency representatives. Even if a technology seems appealing, it doesn't guarantee that it will be more efficient or more appreciated than the usual collection methods. The notion of measuring the respondent burden is not only applicable to the time and effort to respond to a questionnaire but also has to include all those other aspects associated to the collection method. These aspects are sometimes less quantifiable but they are also very important especially when the respondents can choose their preferred collection method and when a statistical agency tries to implement widely a new one. If a respondent would ask "What's in it for me?", we need to be able to show clearly the advantages of EDR from his perspective in order to increase the popularity of this collection method over time.