Working Papers

07/2011

Towards a corporate-wide electronic data collection system at the National Statistical Institute of Spain[1]

Pedro Revilla, José Luis Maldonado, José Manuel Bercebal

The views expressed in this working paper are those of the authors and do not necessarily reflect the views of the Instituto Nacional de Estadística of Spain

First draft: May 2011

This draft: May 2011

[1] This document has been published at the Congress Work Session Statistical Data Editing in May 2011

Towards a corporate-wide electronic data collection system at the National Statistical Institute of Spain

Abstract

Electronic collections present new challenges and opportunities in order to improve editing tasks. They offer the possibility of using built-in edits in electronic questionnaires previously not possible in paper or other modes of data collection. This topic covers all issues relating to methods or strategies about editing of data acquired through electronic data collection (CAPI, CATI, CAWI, etc) and the way the respondents can carry out editing when using electronic questionnaires. Other related topics may include comparisons of editing practices between electronic collections and other collection modes, as well as different problems using multimode data collections. Measuring the respondent burden and the quality and reliability of the responses in order to provide valuable information to other survey processes is another issue of interest. Papers describing editing strategies to improve relationship with respondents or the general editing process are also welcome.

Keywords

Electronic questionnaries, electronic data reporting, web questionnaries, CAWI, IRIA, meta-data corporate-wide data base

Authors and Affiliations

Pedro Revilla

D.G.de Metodología, Calidad y Tecnología de la Información y las Comunicaciones, INE

José Luis Maldonado

Subdirección General de Tecnologías del la Información y las Comunicaciones

José Manuel Bercebal

Subdirección General de Tecnologías del la Información y las Comunicaciones

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (ii): Editing of electronic collections

TOWARDS A CORPORATE-WIDE ELECTRONIC DATA COLLECTION SYSTEM AT THE NATIONAL STATISTICAL INSTITUTE OF SPAIN

Invited Paper

Submitted by the National Statistical Institute, Spain[2]

I. INTRODUCTION

1. Traditionally, the production of official statistics has been based on a *stovepipe model*, where statistics of different domains have been developed independently from each other. The *stovepipe model* is the outcome of a long historic process with a well-known number of advantages (Commission of the European Communities, 2009).

2 Changes in circumstances (increasing needs of data-users, excessive respondent burden, budget cuts), put pressure on official statistical offices to redesign the way its statistics are produced in order to improve the efficiency of statistical production processes. In particular, the stovepipe model presents two main drawbacks: the difficulty to reuse procedures that are similar from survey to survey and the difficultly of integrating data from different surveys.

3. A new model, based on a single standardised production line for all surveys, supported by metadata systems and generic and standardised tools, is difficult to perform in the short term. The main difficulty is to address the great diversity of surveys carried out by statistical institutes. Hence, a step-by-step approach may be used in a way the *stovepipe model* would be gradually abandoned in favour of a more integrated one.

4 An important step to get a more integrated production process is to design and implement a corporate-wide data collection system. The system should be flexible enough to accommodate a variety of surveys. Moreover, it should be able to work in different channels (Internet, telephone, mail, personal interview, etc).

5 Electronic methods offer new opportunities to improve the efficiency of statistical processes and get high quality incoming data, reducing costs at the same time. Concerning data editing, they offer the opportunity for new editing strategies. It is generally accepted that moving editing closer to respondent can significantly contribute to improve editing effectiveness. We can go a step further using electronic questionnaires by integrating the respondents in editing processes.

[2] Prepared by and Pedro Revilla, Jose Luis Maldonado and Jose Manuel Bercebal

6. Traditionally in INE most of the data were collected in paper by enumerators or through self-administered mail questionnaires. Nowadays, like many others statistical agencies, INE has a significant interest in electronic data reporting methods, and in particular, in Web-based data reporting. Concerning business surveys, Web questionnaires are offered as a voluntary option. Moreover, electronic data collection is considered a strategic issue and increasing the percentage of collection via the Internet and other electronic means is a goal included in INE Strategic Plan (INE Development Strategies for the Coming Years, 2009).

7. This paper discusses the possibilities of electronic questionnaires in order to reduce editing tasks and presents the INE project to construct a corporate-wide collecting system. The project is inside the goal to move on to a more integrated production model from the current *stovepipe model.* The paper is focused on the electronic methods and in particular on the Web channel. In the following section, the challenges and opportunities of electronic questionnaires are discussed. In section III, the INE corporate-wide data collection system project is presented. In section IV the CAWI channel is described. The paper ends with some final remarks.

II. CHALLENGES AND OPPORTUNITIES OF ELECTRONIC QUESTIONNAIRES

8. The quick developments in computer technology have had an important impact in the way survey data is being collected, processed, analyzed and published. Electronic questionnaires offer new challenges and opportunities to improve the efficiency of statistical processes. Concerning data editing, they offer the opportunity for new editing strategies.

9. Electronic data reporting (EDR) methods offer new opportunities for improving editing tasks and getting high quality incoming data reducing costs at the same time. Whereas Computer Assisted Interviewing (CAI) integrates into one stage previously distinct phases such as interviewing, data capture and editing, Computerized Self-Administered Questionnaires (CSAQs) go a step further by shifting such activities to the respondent. Hence, electronic questionnaires offer the opportunity for re-engineering editing processes, in a way the respondents may play a more active role in data editing.

10. Several advantages could be expected from using electronic questionnaires. These include improving accuracy and timeliness, and reducing survey cost and enterprise burden. Improving accuracy results from built-in edits, which allow the respondents to avoid errors at the moment they are made. The elimination of data keying at the statistical agency directly gets rid of a common source of error. Some electronic devices (automatic data fills and calculations, automatic skipping of no applicable questions, etc.) could help the respondent to fill in the questionnaire easier and faster. On the other hand, survey respondents may misinterpret the questions they are asked, potentially undermining the accuracy of their answers. One way to reduce this risk is to make definitions of key question concepts available to the respondents (Peytchev et alt., 2010). Although an improvement on data quality could be expected from electronic questionnaires, it is very difficult to measure the real impact on accuracy, given the self-selective nature of the respondents that choose the electronic option. Another accuracy problem is the introduction of bias through mode effects. People without Internet access will never be able to participate using this mean. Even more problematic is that this access is unevenly distributed over the population. A typical pattern found in many countries is that elderly, low educated and ethnic minorities are under represented in using the Web channel.

11. The elimination of data keying reduces the processing time of the survey. There are other factors that can also contribute to improve timeliness. Data transfer on the Web can be done much faster than using the postal system. The cost for statistical offices to carry out a survey using electronic questionnaires could decrease. Savings could be achieved from reducing storage, packing, postal charges and eliminating data keying and keying verification. Some of the editing task could be reduced from built-in edits.

12. Nevertheless, to get the target of reducing respondent burden using electronic questionnaires is not so straightforward. Even though built-in navigation and some electronic devices (automatic data fills and calculations, automatic skipping of no applicable questions, etc.) could help the respondent to fill in the questionnaire easier and faster the reduction in the respondent burden is not always obvious. The respondents' benefits depend largely on the way metadata support the respondent in filling in the questionnaire (help texts, auto-fill rules, pre-filled data, etc). In any case, the respondents' benefits need to be clearly explained to convince them to use the electronic questionnaire. An important element to improve the acceptance of electronic questionnaires among the respondents is to consider electronic questionnaires in a wider context of all their administrative duties and of all electronic data reporting. It is unlikely that respondents are willing to adapt their systems only for statistical purposes. Hence, statistical offices should be aware of the habits of respondents and try to adapt electronic questionnaires to these trends (for example, e-commerce, e-administration, etc.).

13. Many statistical offices are experimenting with the use of different electronic data reporting options in data collection. Web questionnaires offer some advantages over other more complex EDR methods. The Web is a mature technology for EDR because of widespread public acceptance. The prerequisites are only a PC, access to the Internet, and a browser. There is no need, in principle, to incorporate other software on the respondents. The Web makes it simple to put electronic forms at the disposal of almost every respondent, especially in enterprise surveys. Moreover, Web surveys offer new attractive possibilities, such as the use of multimedia (sound, pictures, animation, etc.).

14. Nevertheless, for most of the surveys, EDR cannot be at the moment the only way of data collection. Paper data collection and associated procedures (like scanning) are probably going to stay with us for some years. Hence, a mixed mode of data collection (partly paper, partly electronic) should be used. Global strategies should be designed, because data editing strategies differ whether using paper or an electronic questionnaire. Corporate-wide data collection systems designed to work in different channels may be a useful tool in order to improve the efficiency of multimode data editing processes.

15. Concerning the edits to be implemented, some crucial questions arise: What kind of edits should be implemented on the electronic questionnaires? How many? Only fatal edits or fatal edits and query edits? What kind of edits should be mandatory? What is the different between CATI, CAPI and CSAQs channels? When should the edits be performed? After each data item or after the whole form is processed? On one hand, we need to include some edits. If we do not, then the information collected by CSAQs questionnaires should be treated to the editing procedures in exactly the same way as collected by paper. In that case, we would lose an essential advantage of CSAQs questionnaires: no need to editing again the information with a suitable set of edits implemented in the CSAQs application. On the other hand, we need to be extremely careful in the set of edits to be implemented, because if we implement a big set, then respondents will give up and prefer the freedom they have in paper. Too many edits could even irritate the respondents and increase the burden. In that case we will lose all the advantages of CSAQs questionnaires, as users will prefer the easy way (paper).

16. How to cope with the too few/too many edits dilemma? If we are trying to implement a Web questionnaire in an existing survey, a way is to analyse the current set of edits in order to determine the efficient set of edits to be used in the Web implementation. Hence, the implementation of new procedures obliges to the revision and redesign of the current procedures of the survey. But we should make that revision from the user's point of view. Otherwise, it would be impossible to find out if the users are going to get fed up with the task of filling in a Web form or not. It must be stressed that making that sort of analysis is strictly necessary in order to implement a suitable set of edits that will not discourage users and that will make possible not to edit the Web information in the traditional paper way. In order to achieve this target an analysis similar to that of Martin and Poirier (2002) should be carried out. It is important to have procedures allowing access to versions of data and additional processing metadata that describe how the data were transformed from collection to dissemination.

17. There are a lot of expectations about the role of electronic questionnaires. Nevertheless, until recently, the implementation of Web surveys and other EDR methods in enterprise surveys (and, even more, in household surveys) has often been lower than expected. The take-up of electronic data reporting for statistical data by business providers was generally less than 10%, and often less than 5% (Branson 2002). Other studies also find low rates of response via Internet. For example, Gradjean (2002) finds a rate of 18% for a survey used to construct the Index of Industrial Production in France. In another study, Mayda (2002) finds a rate between 5% and 25% in two quarterly surveys on business and agriculture in Canada. Holmberget alt (2010) finds a rate of 15% in Sweden using a standard strategy but it can be increase to 65% using a web intensive alternative strategy. Even though the usage of the electronic option by respondents has increased lately (for example, Paula Weir, 2005) it still leaves room for improvement.

18 More research is needed to look for the reasons why, up to now, the rate of using EDR is usually quite low, while technical requirements are available for many of the respondents. In the case of business surveys, probably electronic forms have not the same advantages for the reporting enterprises than for the statistical offices. For many of the questionnaires, the most time consuming tasks are to look for the required data and computing the answers. There is no time difference between keying data on a screen and to fill in a questionnaire on paper. The advantages for the reporting enterprises would probably be bigger if the information could be extracted straight from their files. But this procedure may be expensive for both reporting enterprises and statistical agencies, because an initial investment is needed.

19. There are two contradictory targets. On one hand, to implement a single point of entry for all agency surveys, with a uniform security model and a common look across the entire site. And, on the other hand, to allow decentralised applications to cope surveys singularities. One aspect where the difference among surveys has to be taken into account is data editing. Combining the two targets (i.e. integrating a centralised platform with decentralised applications) is a non-trivial task.

20 Encouraging the use of Web questionnaires by respondents is a key issue. Several methods can be used. For example, explaining the benefits to the respondents or considering statistical Web questionnaires in a wider context of all administrative duties and all electronic data reporting (e-commerce, e-administration, etc.). Giving incentives (temporary access to information, free deliveries of tailored data) is another method to increase the take-up of Web questionnaires. In the case of INE Web forms are being offered to reporting enterprises as a voluntary option. In some surveys, we offer tailored data in order to improve the relationships with them (Gonzalez and Revilla, 2002). Using Web questionnaires, when an enterprise sends a valid form (i.e. passing the mandatory edits), it immediately receives tailored data from the server. These tailored data consist of tables and graphs showing the enterprise trend and its position in relation with its sector. Offering this data through the Web has some advantages (speed, possibility to edit the file) over sending this same data on paper by mail. Taking these advantages into account, we expect more enterprises to use the Web channel. Extending this action to more surveys is another goal included in our Strategic Plan.

III. THE CORPORATE-WIDE DATA COLLECTION SYSTEM PROJECT

21. As many other national statistical institutes, INE has started the transition from the numerous stovepipe-like chains of production to more integrated production processes. The Generic Statistical Business Process Model (GSBPM) provides a framework for the developtment of this goal. This new model, based on a single standardised production line for all surveys, supported by metadata systems and generic and standardised tools, is difficult to perform in the short term. The main difficulty is to address the great diversity of surveys carried out by INE. Another difficulty is the conflict between the modernization and the continuous production compromises. Hence, a step-by-step approach is used in a way the *stovepipe model* would be gradually abandoned in favour of a more integrated one.

An important step to get a more integrated production process is to design and implement a corporate-wide data collection system. The system should be flexible enough to accommodate a variety of surveys. In 2010 we have started the development of a parameterized tool (IRIA) that will allow the data collection of all INE surveys, whether they are households or businesses, short-term or structural surveys, and by all the established collection channels (telephone, personal interview, mail, internet, etc.).

22. The parameterization will allow the tool to be actually an application of applications. By means of a “generator” module, the different units responsible for the statistical operation programming will be able to decide the properties that they wish to apply to each survey collection. Another of the basic aspects of this tool will be the reusability of the information. It will allow to design and implement surveys in a simple way when its components are common with others, by reusing the information stored within the same database.

23. Each of the properties, parameters, documentation, etc, that define the production phases, will be reflected in a metadata corporate-wide system, standardized and integrated for all statistical operations, and allowing its reusability whenever it is necessary to perform a new operation. Such metadata corporate-wide system will include structured metadata (variables, concepts, categories,...), reference metadata (associated with the survey methodology), process metadata and quality metadata (quantitative and qualitative descriptions of the quality of each statistical operation).

24. The data that this system will accumulate for each statistical unit at microdata level will come from surveys and administrative data. This repository will allow the use of the information already obtained, either for its confirmation or correction request, as support for the collection or validation of other surveys, etc. The system feeds back to itself allowing the reusability of existing information, either to define the properties of a new survey collection, either as support for its own data collection or its validation. The new collection tool is designed to store historical information about the sampling units collaboration, as well as to update its identification information and the hierarchical relationship among units within the business surveys.

25. When a new survey is projected a cycle similar to this one will be followed

1. The Meta-Data Corporate-wide Data Base (MDDB) is the initiator of the generation of surveys. The MDDB will store metadata standard questions about the different surveys and their relationship to variables and concepts, and process metadata, allowing for reusing that information in other surveys.

2. With this information and by means of the Surveys Generator (SuGe), we will assign all the necessary properties for the collection of information (collection channels, each channel properties, etc.). The questionnaires will be generated based on the information of the MDDB, but it will be required the incorporation of the logic functionality (flows, complex validations, etc.). The generated metadata will be incorporated into MDDB as metadata about the process.

3. The data collection will be carried out on a corporate database where original data will remain unchanged. Each micro-data of each statistical operation will be associated to the question used during the information collection and therefore the variables and concepts used (i.e. each microdata will be associated with its meta-data).

4. Once the data collection is ended, from the corporate database with original data, an editing process and the following business processes will be carried out.

IV. THE CAWI CHANNEL

26. INE has a significant interest in electronic data reporting methods, and in particular, in Web-based data reporting. Concerning business surveys, Web questionnaires are offered as a voluntary option. Moreover, electronic data collection is considered a strategic issue and increasing the percentage of collection via the Internet and other electronic means is a goal included in INE Strategic Plan (INE Development Strategies for the Coming Years, 2009).

27. Up to now INE collects data from reporting enterprises by means of the CAWI channel using a system (ARCE: Storage & Collection of Economic Questionnaires) which will be integrated into the corporate-wide data collection system (IRIA). In 2010 INE collected more than 30% of the questionnaires sent to enterprises, using the CAWI channel. The INE sets up its strategic plans to increase this collection in the coming years. It has already made progress in this direction eliminating the paper questionnaire in business surveys, significantly increasing the former percentage In short term indicators, the take-up in the first quarter of 2011 is about 70%.

28. We think CAWI, due to its availability, its friendliness, preloaded data and editing possibilities, is the safest, the fastest, the highest quality and the cheaper channel collection. Moreover, data from 2010 reveal that in Spain:

1. Companies:

    - 97.2% of Spanish companies with 10 or more employees have Internet access. The 98.2% of them are connected by fixed Broadband
    - 70.1% of the enterprises interact with the Government via Internet, two points more than last year
    - Nearly one in four companies makes purchases through electronic commerce.

2. Households:

    - 57.4% of Spanish homes have wide-broadband connection to the Internet, 11.6% more than in 2009.
    - The number of Internet users grows by 7.1% in the last year and is more than 22.2 million people.
    - 17.4% of the population uses e-commerce.

Due to the INE experience in the CAWI channel and the evolution in the use of Internet, we think of CAWI channel as the primary collection channel of surveys both for companies and households.

29. In IRIA project, it has been planned to develop a tool that allows respondents not only to fill in questionnaires and edit them, but also to have additional information of all the statistical collaboration they make**.** This fact has great importance in business surveys, where companies, according to their size or activity may be included in the samples of various statistical operations and for various periods of time. We also must take into account the hierarchical organization inside enterprises, where the company may have been selected in a survey and their local units in other surveys.

30. It is therefore desirable to exploit the Web channel to provide users information about the status of their statistics collaboration in each survey and in each period, which will also be updated with information from other collection channels. That is, the web channel will not only be an instrument for filling in questionnaires, but will also provide information on the exact status of each questionnaire for each statistical unit.

31. Users will be able to use one of the following access methods to the web channel:

- By means of user name and password which give access to a single questionnaire.
- By means of user registration. The user must indicate the statistical units on which he wants to get information, requesting specific data that only he and the INE know.
- By means of electronic certificate, allowing the display of information about all the questionnaires owning the certificate.
- By means of concerted keys, adding an additional safety feature to the user and password, so after logging he is asked for a fact that only the user and the INE know, accessing to a single questionnaire.

32. Once the access is made, the users can surf through different screens that display the different statistics units, the statistics operations, the questionnaires related to them and their status, and specific information of each survey (methodology, published information, tailor made data, etc.). They will also provide information on the people who can assist them to solve any question or problem, contact phone number, email, etc. Users can also have access to the questionnaires they wish to complete, and in case they pass all the validation rules, they will be sent a receipt.

33. Regarding the editing process, the CAWI channel will use full potential of the new collection tool. It will include specific questions validations, screens, groups of screens or the entire questionnaire. Each validation rule will have properties on the type of severity (major or minor errors), error message, message error parameterized aid, asked to come back, etc. Information system errors detected will be stored, allowing subsequent quality studies.

34. Respect to the completion of the questionnaires, it will be available information about the navigation followed by users, access and entry and exit times for each question or screen, etc. On the web channel it will be possible to transmit the information from one or more questionnaires using Web Services or sending files in XML format. The user response will be the same as if they had completed the questionnaires on-screen.

35. In summary, the web channel will be used to enable respondents to fill out the questionnaires (by three ways of communication), and to provide information on the surveys, as well as about the collaboration of the reporting units and their questionnaires state.

V. FINAL REMARKS

36. Nowadays, public statistical offices are under continuous pressure from society, which demands more and more data, to be produced at a lower cost and with a lower respondent burden. New IT tools and statistical methodologies offer the opportunity for re-engineering statistical production processes in a way the stovepipe model would be gradually abandoned in favour of a more integrated one.

37. According with our experience, the combination of Web questionnaires and selective editing strategy appears very promising. Our goal is that, after implementing correct Web edits, no traditional microediting will be needed. A selective editing approach based on stochastic optimization (Arbues et alt., 2010) would be used in a way that the most influential suspicious values could be detected. Hence, all fatal errors and the most important query errors could be corrected before the survey is disseminated.

REFERENCES

Anders Holmberg, Boris Lorenc, Peter Werner (2010)“*Contact Strategies to Improve Participation via the Web in a Mixed-Mode Mail and Web Survey”* Journal of Official Statistics

Andy Peytchev, Frederick G. Conrad, Mick P. Couper, Roger Tourangeau (2010) “*Increasing Respondents' Use of Definitions in Web Surveys*. Journal of Official Statistics

Arbués I., González M. and Revilla (2010). “*A Class of stochastic optimization problems with application to selective data editing”*. Optimization. Taylor & Francis

Commission of the European Communities (2009). “*Communication from the Commission to the European Parliament and the Council on the production method of EU statistics: a vision for the next decade”*

INE (2009) “*Development Strategies for the Coming Years* “

Gonzalez, M. and Revilla, P. (2002). “*Encouraging respondents in Spain*”. The Statistic Newsletter. OECD. Issue No. 12.

Weir, P. (2005), *“The Movement to Electronic Reporting and the Impact on Editing”*. 55th Session of the International Statistical Institute. Sydney. April, 2005