Distr.
GENERAL

CES/AC.71/2004/7
8 March 2004

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Geneva, 17-19 May 2004)

Topic (i): Web technology in statistical information systems

THE USE OF A DATA CAPTURE TOOL IN STATISTICAL SURVEYS ON ENTERPRISES

Contributed Paper

Submitted by the National Institute of Statistics, Italy[1]

I. INTRODUCTION

1. Statistical surveys normally involve three respondent classes: households, institutions and enterprises. Each of them has its own different characteristics, which can influence the operational procedures and the choice of the most appropriate data capture technique.

2. While using web technology in data collection, an analysis of the statistical context is necessary to deal with the methodological matter of *self-selection* of respondent units, which can affect the sample design and the statistical inference. The web cannot be the means of extracting random sample units, nor can these units be representative of the whole population. Nevertheless, experience shows the possibility of applying the web collection in many different situations, if we keep the sample design step separated by the collection phase.

3. Respondent classes will be considered separately.

4. Experience gained from sampling surveys has shown the National Institute of Statistics (ISTAT) that households taking part in sampling surveys are used to interacting with interviewers, so the CATI/CAPI techniques - which don't create any problem with the coverage rates of the population - are used more often than web techniques while innovating the survey process.

5. Public institutions are usually better equipped with advanced computer technologies to allow a more efficient data capture. They are almost always involved in total surveys and their technological profile is well known to the Institute, which keeps in touch regularly with them. In this case, web collection is feasible and only organizational problems can arise.

[1] Prepared by Rossana Balestrino (rossana.balestrino@istat.it); Patrizia Larese (patrizia.larese@istat.it)].

6. Enterprises generally participate in sampling surveys and fill in statistical forms that are sent and returned mainly by mail or by fax. The enterprise's technical equipment may be very heterogeneous, depending on the size of the company. The choice of electronic forms and the use of the Internet cannot be imposed on these users. The best way to renew the data collection phase is the adoption of a mix-mode technique that allows users to choose by themselves the answer method: either traditional (paper forms) or innovative (electronic forms and the Internet). In this respect, it is necessary to communicate to the units, which have been previously extracted with the classical sampling method, the possibility of using a web form or an e-mail form as an alternative to the traditional paper questionnaire. The multi-technique collection, which even implies additional costs to manage and monitor data coming through different channels, is methodologically feasible and contributes to the timeliness of the final statistical data.

II. TELEFORM SOFTWARE

7. Teleform is a CARDIFF product that needs a WINDOWS server platform (NT, 2000, 2003 SERVER). It was chosen by the Data Capturing Laboratory-IT Directorate of ISTAT for the versatility to manage multiple data capture channels. The information capture system is capable of processing thousands of paper forms per day. It generates electronic forms in HTML or PDF, which can be administered by web or by e-mail. It allows the fax-server use for mailing and receiving paper forms, which can be processed digitally by the optical character recognition machines included in the system. Teleform works seamlessly with production-level scanners, fax servers and the Internet to capture, verify, process and index data.

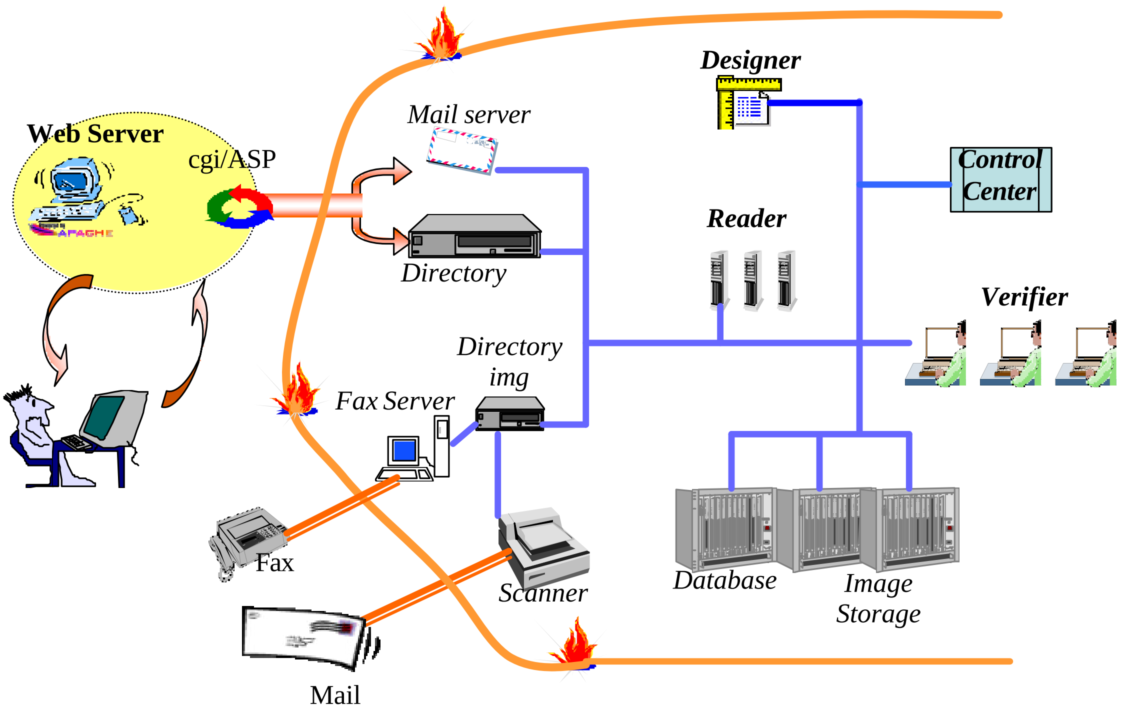
8. Teleform is a modular system which can be expanded according to the requirements imposed by the workload. Three main modules – *Designer, Reader* and *Verifier* – make up its architecture (Fig. 1):

- *The Designer* to define forms;
- *The Reader* to manage the form mailing, receiving and reading;
- *The Verifier* utilized by operators to verify the quality of the data captured.

9. The *Designer* is the *point-and-click* interface that makes it possible to create new forms from scratch or to automate existing forms. It allows all the checks and verifications needed to interpret the documents once they are loaded, to be defined beforehand. Once a form has been defined, it can be produced automatically in various formats, such as traditional paper, fax document, PDF or HTML file. The *Reader* is Teleform's heart. It handles document scanning, capturing images from fax server, importing images from existing directories, any merging of output with data from an existing archive, identification of documents from among those stored in the system, and recognition of data. If electronic forms are used, the *Reader* handles mail server management, document capture, form identification and creation of the output data set. Teleform automatically works out paper forms, although some manual touching-up is required. The *Verifier* interface allows the operator, with the aid of the image, to quickly verify and correct any characters that the OCR/ICR engines are unsure about. The *Verifier* is also used to solve any cases where characters have been recognised, but violate any validation rules set when the form was defined.

10. A PDF module has been developed in partnership with ADOBE Systems. Using it, a digital version of the module, from a definition prepared with the *Designer,* can be generated. The digital module can be published online or sent by e-mail and processed automatically. In either case, once the respondent has filled in the form and submitted it, Teleform takes over the process and converts the output into an automatic e-mail to the producer of the form. Even if the user prints the PDF form and fills it in manually, automatic processing is still possible, simply by scanning the returned form or retrieving it from the fax server. Teleform also includes the *AutoMerge Publisher* for pre-filling forms, costuming documents and delivering them automatically via fax, print or e-mail.

Fig.1 - Teleform Architecture



III. PROCESSED SURVEYS

11. Timeliness in the data collection phase is crucial for enterprise surveys, and especially for short-term statistics, as data have to be available within deadlines which are fixed by the European institutions. The Teleform solution has been considered suitable to optimize the data capture for these surveys and respondent typology.

12. During the last months a few experimental Teleform projects were realized by the statistical units which, in the present organization, are autonomous in developing EDP solutions. These surveys were:

- The monthly survey on retail sales uses two forms for three pages. The theoretical number of the respondents is 2,700.
- The annual survey on telecommunications uses three forms for nine pages. The theoretical number of the users is 534.
- The monthly survey on employment, working time, and salary. It utilizes two forms for three pages. The theoretic number of the users is 1,500.
- The quarterly survey on vacancies and worked hours uses one form for four pages. The theoretical number of the users is 8,000 for every quarter.

Further descriptive data of the surveys treated in experimental way are reported below:

Survey on retail sales	
Frequency	Monthly
Number of forms	2
Number of form pages	3
Respondents	Commerce branch enterprises (small, medium and large enterprises)
Theoretic number of respondents	2,700
Technologies for data collection	Mail, Web, Fax-server
Channels used by the respondents:	Web: 4% Mail: 21.2% Fax: 74.8%
Monitoring of returned data	Yes
Security procedures	Access control (User-id and password)
Survey on Telecommunications	
Frequency	Annual
Number of forms	3
Number of form pages	3
Respondents	Telecommunications enterprises
Theoretic number of respondents	534
Technologies for data collection	Mail, Web, E-mail, Fax-server, Fax
Channels used by the respondents	Web 10% Fax and Fax – server 90%
Monitoring of returned data	Yes
Security procedures	Access control (User-id and password)
Survey on employment, working time and salary	
Frequency	Monthly
Number of forms	2
Number of form pages	3
Respondents	Large enterprises (more than 500 employees)
Theoretic number of respondents	1,500
Technologies for data capturing	Mail, Web, E-mail, Fax-server, Fax
Channels used by the respondents	Web 10% Fax and Fax – server 90%
Monitoring of returned data	Yes
Security procedures	Access control (User-id and password)
Survey on vacancies and worked hours	
Frequency	Quarterly
Number of forms	1
Number of form pages	4
Respondents	Commerce branch enterprises (small-medium-large enterprises)
Theoretic number of respondents	8,000
Technologies for data capturing	Mail, Web, E-mail., Fax-server, Fax
Channels used by the respondents	Web 10% Fax and Fax – server 90%
Monitoring of returned data	Yes
Security procedures	Access control (User-id and password); SSL

IV. PERCEPTION OF THE STATISTICAL USERS

13. The results of the experiment were not perceived in the same way by all statistical users. The majority of them appreciated the results, but some were not completely satisfied and confident in the new technique. The IT Directorate, responsible for addressing and coordinating the data processing function and systems, decided to evaluate the software capacity and identify constraints and requirements for a possible massive start in production.

14. The perception of statistical users about Teleform was positive and negative at the same time:
Positive perception:

- the new software reduces time and costs;
- it increases the productivity as it eliminates manual data entry;
- it delivers completed forms and documents into back-end systems for instant storage and retrieval so that the data sent by the respondents are immediately available in the database.

Negative perception:

- lack of product experts to assist statistical users during the procedures;
- long response time using the *Verifier* module*;*
- impossibility of saving partial versions of the PDF completed forms.

V. THE SOFTWARE EVALUATION

15. Local experiences were analyzed to obtain a correct evaluation of the software. The problems met by users were not all due to Teleform itself; many of them were caused by network failures and lack of local specific skill. The analysis confirmed the versatility of Teleform about the use of different channels in mailing and receiving forms. The system is able to generate a standard multi-channel solution for a *thin* form (1-2 pages with a few controls on form fields is the optimum questionnaire type) at low cost. It doesn't require any software development as the executables are automatically produced by the form template created with the Designer module.

16. The treatment of the paper forms by the optical character recognition machines is an important feature of the software. Paper forms can thus undergo the same automated process as the electronic forms. Images, either coming from the fax-server or from the scanner, are automatically passed to OCR/ICR engines and then to operators who validate uncertain data by means of the interactive Verifier module. A test about the quality of the automatic recognition of characters gave good results, aligned with other available optical reading software on the market. The conditions and results of the test are described below.
Fifty real forms from the quarterly survey on employment, working time and salary were scanned at the Institute's data capture laboratory and underwent the following procedure:

- automatic character recognition;
- verification by a human operator;
- export of data;
- analysis of results in comparison to a "perfect" file.

17. All OCR software is based on one or more recognition engines which interpret the characters by comparing them with character models stored in internal libraries. The confidence threshold for this recognition can be changed by the user but we kept it at the standard value of 80%. Once the threshold is set, the recognition step produces one of three results:

(i) the character is identified correctly;
(ii) the character is identified incorrectly (false positive) ;
(iii) the character is rejected (and must be verified manually).

18. The most sensitive part of the recognition phase concerns the second case - keeping false positives to a minimum is the main quality objective. The third situation is equally important, but for reasons of cost. Rejected characters can be positively identified, but only using costly human intervention. The confidence threshold must therefore be set at the level which best balances the level of quality desired with the intervention required to ensure it is achieved.

19. The selected form fits on 2 A4 pages. There were 8333 non-blank characters on the forms, which is a sufficient number to ensure the results of the recognition tests are significant. The average number of characters per form was 166.66. The average time required for automatic interpretation was 7.04 seconds per form. As for the validation step, the average time required for a human operator to verify the forms was 44.12 seconds per form. There were 833 rejected characters (10% of significant characters), of which 526 (63.14%) needed to be corrected (the others were correct, although the recognition engines were "unsure"). At the end of the process, 1.03% of the total number of significant characters were false positive (errors).

Table 1 – Form Structure

Pages	**Characters**			
	Pre-printed	**Mark-sense box**	**Other Characters**	**Total**
2	42	6	448	496

Table 2 – General statistics on treated forms via Scanner

Forms	**Significant characters**	**Blank characters**	**Total characters**	**Average number of characters per form**	**Average scanning time per form**	**Average interpretation time per form**
50	8333	5893	14226	166.66	4.44 sec	7.04 sec

Table 3 - Quality of character interpretation

Confidence level	**Forms needing verification**	**Characters to be verified**		**Characters to be corrected**		**Wrong characters at the end of the process (false positive)**		**Average correction time**
		N.	**% of significant**	**N.**	**% of verif.ied**	**N.**	**% of total characters**	
80%	50	833	10%	526	63.14%	147	1.03%	44.12 sec

20. In conclusion, the trial run of using Teleform to process paper forms was satisfactory. We must also consider the possibility, if necessary, of further reducing the final error rate - under 1% - by raising the confidence threshold and accepting a heavier consequent manual validation activity. Nevertheless, in our experiences, the use of Teleform with paper forms was affected by too long response times while using the

Verifier module. This was caused by the network infrastructure. Figure 2 illustrates the ISTAT network; the speed of the link between the central site - where Teleform server was installed - and the peripheral site of via Tuscolana - where the surveys processes are managed and the Verifier should have worked - doesn't respect the recommended parameters. For the time being we advise our users to adopt Teleform for dealing with electronic forms which don't ask for the use of the Verifer module.

21. On the other hand, Teleform, which is an effective data capture software, was not designed for statistical targets. Among its limits, it is not able to process complex forms, such as for instance those for investigating social phenomena; it is also not able to save partial versions of the written out forms, which seems to represent a respondent need, especially when the form is not short. Moreover, in order to use electronic forms, some software requirements are mandatory on the PC client side, concerning the browser version and the availability of Acrobat, if the PDF format is adopted. In a network, Teleform, as we directly experienced, must work with a high quality network card at 100 Mbps in the server and in the client.

22. To sum up, Teleform can be considered an interesting software system capable of excellently dealing with simple forms which don't require special conditions of treatment. If more sophisticated checks are necessary on input fields, or intermediate versions of the forms must be stored and presented again to the respondents, an extra EDP activity should be planned, so that it could become more convenient to develop the whole solution in an autonomous way. It must be said, that many of the regular surveys managed by our Institute, especially on enterprises and institutions, actually adopt *thin* questionnaires.

VI. MANAGEMENT AND ORGANIZATION COSTS

23. The complex architecture of Teleform needs minimum requirements to obtain good performances. The network requirements at our premises are the most critical ones, as they have to grant the communication between the different modules. Teleform versatility requires high management costs to maintain suitable service levels. The complete application needs 6 servers (Teleform server, WEB server, Input Mail server, Output Pop server, FAX server, DB server) linked in a network.

24. To renew the data collection process, ISTAT has to support the costs of the installation and maintenance of a more adequate infrastructure. Besides, it has to enrich the network with monitor tools and the management of the service levels.

25. To monitor the Teleform system there is no product that covers completely the process. We have to use different tools that allow us to control partially the hardware and software resources. ISTAT will develop personalized tools to control the activities of the servers involved in the process. The most important commercial tools used for monitoring are reported below:

(i) ***Teleform Control Center*** is a module of the Teleform product. It gives system administrators the power to monitor and manage the data collection process. Administrators can report and react to every activity, from form input and recognition to data output across a network. Based on the Microsoft Management Console (MMC), the Teleform Control Center ensures that throughput goals are achieved by providing real-time reporting on individual, group and system-wide performance. The Control Center allows system administrators to:

- view work in process and overall system performance;
- view and monitor operator performance and throughput in real time;
- make changes to system-wide and operator/workstation settings;
- control the priority of work and escalate specific batches for immediate processing;
- maintain system security, assigning operator logins and passwords;
- configure alerts that will highlight specific throughput or processing issues for immediate attention;
- generate extensive reports on system and operator performance;

- alert notificatio n via email, command line execution, and/or NT event log .

(ii) ***Webalizer*** is a free web server log file analysis program. It produces highly detailed, easily configurable usage reports in HTML format, for viewing with a standard web browser.

(iii) ***The Multi Router Traffic Grapher (MRTG)*** is a tool to monitor the traffic load on network-links. MRTG generates HTML pages containing graphical images that provide a *live* visual representation of this traffic.

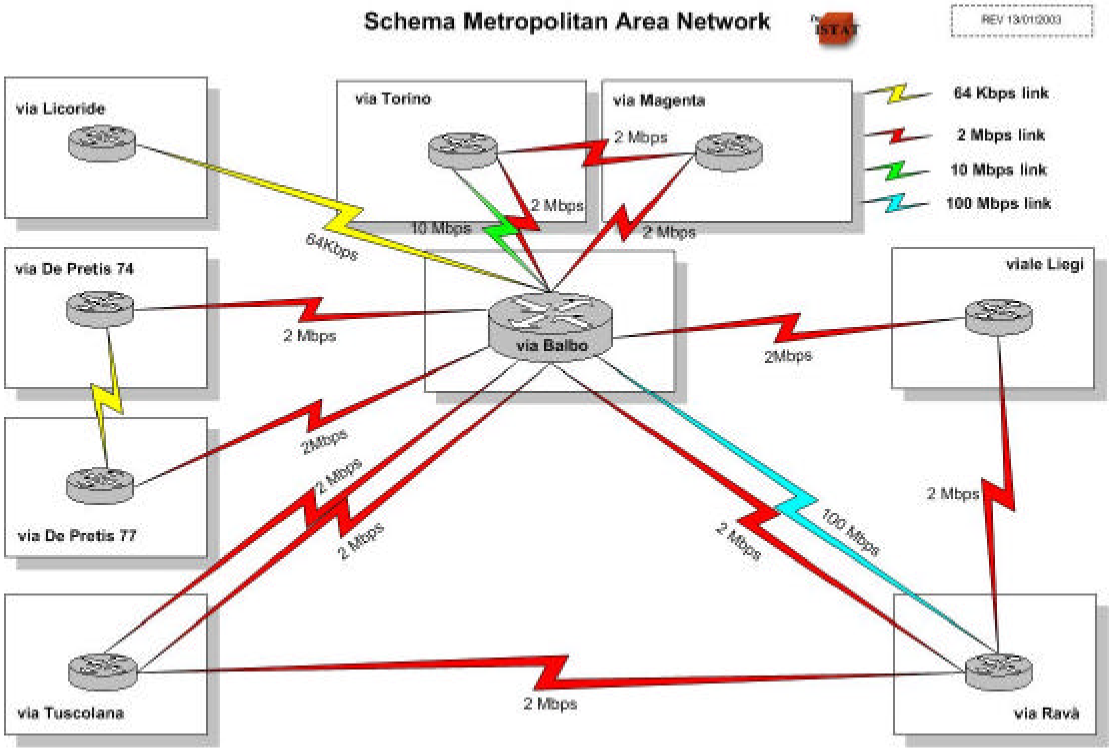
26. Statistical units responsible for managing surveys, should accept changing the human resources organization to use Teleform. Most of them will have to learn the product to manage the electronic forms. It means additional costs and time to have new professionals to operate with the new software.

VII. PROSPECTS

27. The use of Teleform is recommended with *thin* forms and a computer network at 100 Mbps. The central IT Directorate is renewing the intranet network capacity raising it at 100 Mbps for all the workstations wherever located. In order to favour the optimal use of Teleform, it is planned to keep centrally an expert team to analyse together with the statistical users the requirements of the process to be treated by Teleform in terms of structure of the questionnaire and check activity on the input data. Moreover, an estimate of workloads on the involved servers and network should be provided to be sure that the whole environment will be able to support the new process. To monitor the Teleform processes once started, new procedures will be set up to measure service levels and guarantee the availability of the resources employed. At the same time, as we have already noticed, Teleform cannot cover the treatment of every kind of form: the more complex forms, with many pages, *filter* questions and automatic control of the sequence of sections to be filled in, require a different treatment. Ad hoc dedicated web applications can be developed or, if we look for a generalized solution, further testing activity of specific software products will be necessary, both commercial and available from the statistical official community. In particular the BLAISE software is a standard product already used in ISTAT for CATI/CAPI data capture. It is scheduled a test of the Internet BLAISE module.

28. In other words, we do not yet have a specified standard for web forms, and maybe the standard will not be a unique tool but possibly a set of tools and solutions. What we are mostly interested in would be to fix an operational model to realize data collection through web, which implies dealing with topics such as authentication of the parts who exchange data, security rules for transferring and storing data, monitoring procedures, etc. At the moment we are planning on setting up a unique web site, in UNIX or LINUX environment system, dedicated to data acquisition. This website, compliant with the above-mentioned operational model, should host the executable electronic forms – possibly developed with different systems and languages – to be accessed by respondents to any survey.

Fig. 2 – ISTAT Network



References

Balestrino, R. and Barcaroli G. (1998), “The Introduction of CASIC Technologies in an Institute Producing Official Statistics”, *Documenti Istat*, 3, National Statistical Institute, Rome.

Balestrino, R. and Politi M. (2001), “A complete system of Data Capture to improve timeliness of short-term statistics”, *Seminar on short-term statistics-Improving timeliness and co-operation*, Luxembourg, 8-9 March 2001.

Bergamasco S., Rotondi G., Serbassi S., Toma A, “A complex software architecture for a Nation wide continuous longitudinal sample survey”, *2002 International Conference on Software Engineering Research and Practice, SERP'02*, Las Vegas, USA, June 24-27, 2002.

Biffignandi, Silvia and Pratesi, Monica (2002): *Modelling the respondents' profile in a web survey on firms in Italy*. Metodološki zvezki 18. Faculty of Social Sciences, Ljubljana.

Couper, Mick. P. (2000): "Web survey: a review of issues and approaches", *Public Opinion Quarterly*, 64: 464-94, 2000.

Couper, Mick. P. (2002): "New Technologies and Survey Data Collection: Challenges and Opportunities", *ICIS 2002 - International Conference on Improving Surveys*, Copenhagen, Denmark, August 25-28, 2002.

Fabbris, Luigi (2002): "Satisfaction Scales in a CAWI Survey on University Teaching Evaluation", *International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET),* Charleston, South Carolina, November 14-17, 2002.

House, Carol C. (2002): "Integrating paper and web instrument development to enhance efficiency and standardization", *ICIS 2002 - International Conference on Improving Surveys*, Copenhagen, Denmark, August 25-28, 2002.

Romano, Maria Francesca and Himmelmann Maurizio (2002): "Determinants of Web mode choice in a "Web and paper" survey in a high education population", *ICIS 2002 - International Conference on Improving Surveys,* Copenhagen, Denmark, August 25-28, 2002 .

Tremblay, L. (2000), Integrating Business Survey Systems, *Proceedings of the IFD & TC Conference 2000,* Portland, Oregon.

www.cardiff.com: Teleform documentation.

- - - - -