A COMPLETE SYSTEM OF DATA CAPTURE TO IMPROVE TIMELINESS OF SHORT-TERM STATISTICS

Rossana Balestrino and Mauro Politi[•]
ISTAT, Italian National Statistical Institute, Rome, Italy

SUMMARY

The usefulness of short-term statistics for analysing the economic cycle and influencing economic and monetary policy is highly dependent on their timeliness. Of the steps that may be taken to ensure timeliness, National Statistical Institutes should focus on technological innovation and on making the tools used by businesses to respond to the statistical surveys both user-friendly and efficient. This paper describes a trial run carried out with a complete system that allows data to be captured whether it has been transmitted by mail, fax, the Web, or any other channel which may be used to return the questionnaires.

I. THE NEED FOR TIMELINESS IN SHORT-TERM STATISTICAL SURVEYS

Council Regulation (EC) No 1165/98 concerning short-term statistics states which statistical indicators each National Statistical Institute must produce and sets deadlines for the dissemination of those indicators. Today, more than two years after the Regulation entered into force, these deadlines are still not being met by certain countries and for certain indicators.

It should not be forgotten that, while the procedure to prepare the Regulation was running its course (activities of technical working parties, meetings of the European Council, approval of legislative bodies, transition period, need to adapt national statistical systems to the new needs for economic statistics) an extremely important event took place: the birth of EMU, Economic and Monetary Union. In addition to bringing forward the full implementation of the Regulation, this event created new needs and purposes for economic statistics. In order to carry out the activities expected of it, the European Central Bank has made a series of requests to which the European Statistical System must give some response. So, while National Statistical Institutes were already experiencing difficulties complying with the new Regulation on short-term statistics, additional demands were arising with regard to the comprehensiveness of the indicator set to be produced and the timeliness of the data's dissemination.

When the Ecofin Council approved the second report on the statistical requirements for monitoring Economic and Monetary Union produced by the Economic and Financial Committee, it asked Eurostat and the ECB to draw up an Action Plan to improve statistical information on the European economy and to thus improve economic and financial policy. A large section of the Action Plan is dedicated to short-term statistics, for which timeliness is considered indispensable.

Most of the short-term statistics referred to in the Regulation are calculated from data collected in statistical surveys of businesses: production, turnover, new orders, occupation, earnings, prices,

[•] Sections I, II, III and VI were written by M. Politi, while sections IV and V were written by R. Balestrino.

etc. Only a small number of sources are administrative. Enterprises are therefore likely to receive an ever-increasing number of requests for information, to which they must reply in an ever-shorter time. Optimising the process of producing of the surveys is one way that Statistical Institutes may meet the new requirements (timeliness and comprehensiveness) brought about by monetary union and avoid making replying to the surveys an excessive burden on enterprises.

II. INNOVATION IN THE PRODUCTION PROCESSES OF SHORT-TERM STATISTICAL SURVEYS OF ENTERPRISES

Respondents to statistical surveys of enterprises (industrial, construction, retail and services) generally share a number of characteristics: the enterprises generally have a cooperative attitude towards surveys carried out by the National Statistical Institute, although, at the same time, they are extremely aware of the increase in the statistical burden insofar as the more requests are made by the national statistical system, the greater the cost to the enterprises of responding to the various surveys. Minimising the statistical burden for enterprises has always been a concern of Istat, which is why it has tried to plan its surveys so that they are carried out as efficiently as possible.

Efficiency, with regard to short-term statistical surveys of enterprises, means the following:

- Updating the reference archive efficiently. In order to construct representative samples, capture the effect of the demographic dynamic of enterprises and contact respondents correctly, the archive must be complete and continuously updated. Enterprises judge harshly any archive of the Statistical Institute which contains old or erroneous information. Short-term statistical surveys therefore require the archive to be updated in real time.
- Sending questionnaires to enterprises on time.
- Making it as easy as possible for enterprises to respond. It is a duty of statisticians to examine every possible way of assisting enterprises in replying to surveys.
- Shortening the time taken to collect data. The results of the survey could then be made available more quickly, and enterprises could receive feedback to the information they supplied and take good advantage of it.
- Make targeted, well-motivated requests. Where there is no response, the enterprises which have not replied, and only those, should quickly be contacted.

Istat has taken a number of steps to meet these needs, the most important of which are the following:

- Designing user-friendly questionnaires. Over the years, the questionnaires have been changed a number of times in order to meet enterprises' requests. For example, shuttle questionnaires have been created.
- Using a single post box as the return address for the questionnaires from each survey. Pre-paying postage on the return envelopes, so that enterprises do not have to cover the cost of postage.
- Preparing computerised procedures to make targeted, rapid requests to enterprises which have failed to reply. In this way, repeated reminders will not be sent to those who have already replied.
- Using an automated system to stamp and post questionnaires.
- Using an automated system to send reminder faxes or letters in real time to enterprises that have not replied.

As can be seen from the needs described above and the action taken to meet them, there has been a continuous process aimed at improving two aspects of short-term economic statistics: quality and timeliness. One of the ways this can be achieved is clearly by introducing innovative technologies into process of producing statistical information.

III. A COMPREHENSIVE DATA-CAPTURE SYSTEM

Having enterprises respond to the short-term statistical surveys solely by fax has made a significant contribution to the timeliness of short-term economic statistics. In addition, enterprises reacted very positively to the request.

This approach has many advantages.

For enterprises:

- Rapid fulfilment of the obligation to reply: it takes less time to send a fax from one's own office (or a room nearby) than to post a letter.
- Ease of use: fax machines are not particularly difficult to use.
- Low cost: sending a fax is cheaper than having a messenger post a letter.
- Immediate confirmation of receipt: the transmission report provides immediate confirmation that the questionnaire has been received, which may be useful in case of a dispute with Istat.
- Reduces the irritation caused by mistaken requests: as the responses to the questionnaires arrive at Istat immediately, reminders will take into account the most recent arrivals, and only those who have not replied will be contacted.

For Istat:

- Questionnaires are available very quickly.
- The problems connected with relying on the mail, such as delays and losses, are avoided.

If data are to be received by fax, two conditions must be met: the questionnaires must not be too long (no more than 4-6 pages), and the telephone lines must never be busy. The respondent must not be faced with a bottleneck that wastes his time. The Statistical Institute must therefore have a fax server that can handle several telephone lines.

Starting from the fact that a fax is simply an image file (*.tiff) which has been transmitted over a telephone line, and using recently-developed technologies, Istat has carried out a study on using a fax server with character recognition for statistical surveys [3]. The trial was carried out on questionnaires for the monthly turnover survey, using an application with both hardware and software components. A number of questionnaires with hand-written responses were faxed and then "read" by the recognition engine. In a nutshell, this experiment resulted in a 0.7% rate of incorrectly recognised characters at the end of the recognition process. This has encouraged us to continue this approach.

Over the last few years, use of the Web to exchange information has grown exponentially. At the beginning of 1999, it was estimated that 10% of enterprises had access to the Internet, but it can now be said that practically all medium- and large-sized enterprises have access. An additional incentive was provided by the enterprises themselves, as many showed they were willing to submit the questionnaires over the Web. This motivated us further, and, inspired by the need to improve timeliness, we searched for a comprehensive product that was capable of capturing data from paper (mail or fax) and the Internet (directly from websites or by e-mail).

Of the various products that are commercially available, we settled on one which included all of the following phases in a single software package:

- data capture using every method available to enterprises;
- data registration;
- preliminary data checks and subsequent correction of errors;
- data archiving.
-

IV. TELEFORM - A DATA-GATHERING PROGRAM THAT USES MULTIPLE TECHNOLOGIES

Istat has recently completed a trial run of the monthly surveys of enterprises' industrial production using the Teleform program, a very interesting example of a multi-technology approach [1] [4] to data gathering.

Teleform 7.0, produced by Cardiff (USA), is dedicated to the processing of forms in both paper and digital format. It is therefore highly suitable for gathering data efficiently in a statistical environment where various methods of responding are to be made available. The ability to process both paper and electronic copies is what makes this product stand out.

Paper questionnaires can be returned by mail or fax, as usual. In the first case, the questionnaire must be scanned. In the second, if a fax server is available at the place of production, the scanning step is unnecessary. In both cases, the scanner and the fax server can be incorporated in the Teleform system. This makes it possible to build up a single archive containing the images of the forms, whether the source is the scanner or the fax server. In a subsequent step, OCR/ICR is used to automatically recognise type-written or hand-written characters, marksense is used for tick boxes, and bar codes are read. Pre-existing paper forms can be processed, or, in an ideal situation, they can be designed from the ground up. Digital forms are delivered to users by e-mail, or made accessible on the Web. The data received electronically can be fed into the same data set as data from automatic character recognition.

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 (Figure 1).

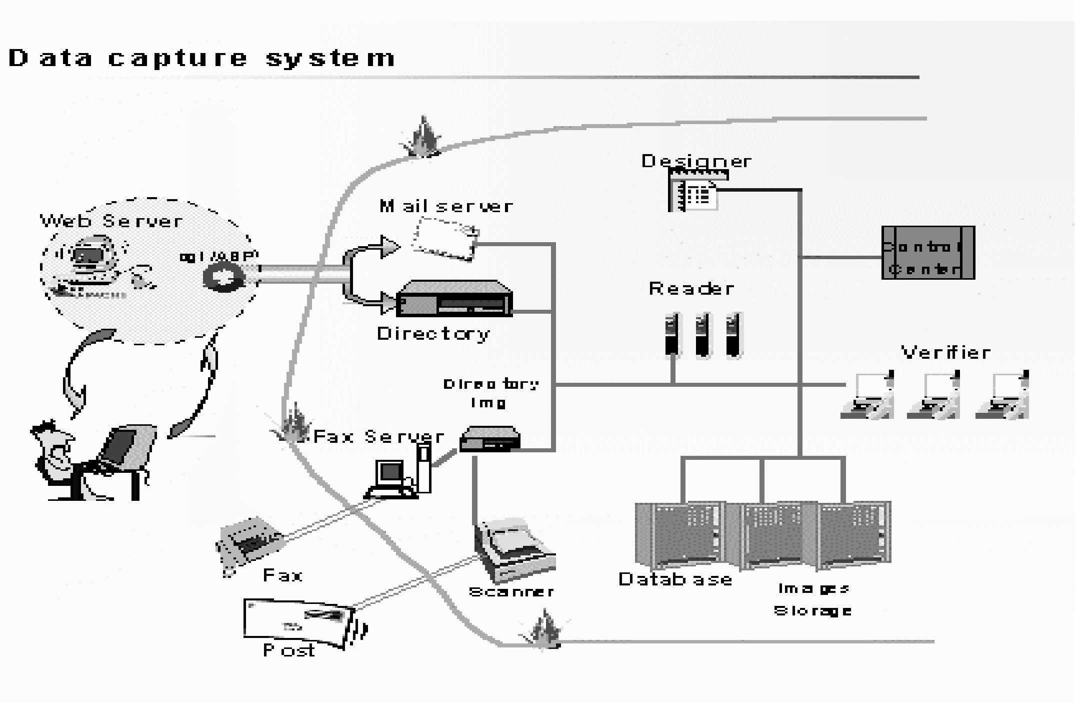
Designer is the point-and-click interface that makes it possible to create new forms from scratch or automate existing forms. It allows all the checks and verifications needed to interpret the documents once they are loaded to be defined beforehand. Once a form has been defined, it can be produced automatically in various formats, such as traditional paper, as a fax document, or as a PDF or HTML file.

Reader is Teleform's heart. It handles document scanning, capturing images from fax servers, importing images from existing directories, any merging of output with data in the archive, identification of documents from among those stored in the system, and recognition of data. If electronic forms are used, Reader handles mail server management, document capture, form identification and creating 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. Verifier also solves any cases where characters have been recognised, but violate any validation rules set when the form was defined.

A PDF module has been developed in partnership with ADOBE Systems. Using it, a digital version of the module from a definition prepared in Designer can be generated. The digital module can be published online [2] or sent by e-mail and processed automatically. In either case, once the respondent has filled out 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 out manually, automatic processing is still possible, simply by scanning in the form or retrieving it from the fax server.

Fig. 1 – Teleform Architecture



In order to fill out a PDF form, the user must have access to the Internet and have Acrobat Reader 4.0 (available at no charge on the Web). The standard approach, using e-mail, is for the user, using his browser, to open the PDF form that Teleform has sent as an attachment, fill it out, and submit it. If the user wishes to fill out the form off-line, he must save the attached file. He may then open it and fill it out, although certain guidelines must be followed. The data is returned to Teleform within the body of an automatically generated message rather than as an attachment, thus keeping the security risk to the receiver low.

HTML forms are also generated automatically on the basis of the definition prepared in Designer. The two distribution methods – Web and e-mail – are again available. However there is currently one drawback for users who wish to use Netscape, as they must have version 6.0 or above.

Teleform also includes a module that allows forms to be personalised by including different information for each user. The information must be stored in specific tables that the system recognises and dynamically associates with the forms. This makes it possible to print personalised forms, or send personalised e-mails and faxes.

A programming language is also available. Called “BasicScript”, it resembles Visual Basic, and allows image evaluation to be customised using ad hoc checks, arithmetic calculations, field comparisons, and calls to external applications. However, BasicScript code cannot be used in the digital versions (PDF and HTML) of the form. In this case, ad hoc checking modules need to be implemented in JavaScript.

In addition to the ASCII format, Teleform is able to provide output in the formats used by the most common Windows databases (e.g. CSV, Excel, Access, Paradox, Sybase, etc.) as well as more sophisticated programs (e.g. Oracle) via an ODBC interface. In addition, scripts can be used to personalise output, thus extending the connectivity of the product to the production environment.

The network version of Teleform includes the control centre module, which makes it possible to coordinate the system’s activity and control the access and functions available on each workstation. It also provides statistics on each phase: scanning, preparation, checks, etc. and enables the overall productivity of the environment to be monitored, as well as that of each station and worker, thus bringing to light any bottlenecks in the process.

V. THE TRIAL-RUN AND ITS RESULTS

A trail-run was carried out to check whether it would be possible to use Teleform for the short-term statistical surveys of enterprises. The Teleform data capture system was installed on an intranet running on the Windows NT platform. However, in order to increase the number of data capture channels available to it, the system interacted with other Windows environments and Unix platforms (mail server, web server, etc.), both within the intranet and on the Internet.

As regards responses on paper, 200 forms from the monthly survey of industrial production were processed using both of the available means of data capture: scanner and fax server [3]. The form was designed from scratch using Designer. 200 copies were printed out and filled out in different handwriting. The completed forms were scanned in at the Institute’s data capture laboratory, then sent by different fax machines of varying quality to the Institute’s central fax server as a way of simulating fax returns from the responding enterprises. Thus, two image files were generated: one from direct scanning of the paper forms, the other from the fax server. The latter method comes closest to reproducing real-world conditions, as currently enterprises most often return the filled-out forms by fax. It is therefore necessary to check that the inevitable “noise” added to the image by the fax machine does not lower the quality of the final result in terms of the data captured. The following procedures were applied to each of the two files:

- automatic character recognition (varying the “confidence” threshold for the character recognition, set by default at 80%, but which may be customised);
- verification by a human operator;
- export of data;
- analysis of results in comparison to a “perfect” file.

All automatic character recognition 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. Once it is set, the recognition step produces one of three results:

1. the character is identified correctly
2. the character is identified incorrectly (false positive or ambiguous)
3. the character is rejected (and will be verified manually)

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.

The results were analysed using a program developed specifically to compare the files exported by Teleform with the "perfect" files obtained by conventional double entry of the 200 forms being processed and reconciliation of the discrepancies. The output of the conventional and automatic methods was compared, byte by byte, to produce the number of correct characters at the end of the process (following manual correction of unclear characters). Table 1 gives the statistics of the form. Tables 2 and 3 give the results obtained using images from the scanner and the fax server respectively.

The forms fit on one A4 page. There were 28 473 non-blank characters on the forms, which is sufficient to ensure the results of the recognition tests are significant. The average number of characters per form was 143, which conforms to real-world conditions. The information to identify consisted solely of numbers, so no indicators for character type were calculated. The average time required for automatic interpretation was two seconds per image.

Analysis of the results of the scanned data capture shows that, at an 80% confidence threshold (Teleform default), the average time required for a human operator to verify the forms was nine seconds per form, or a total correction time of 29 minutes. There were 1 488 ambiguous characters (5.2% of significant characters), of which 850 (55.4%) needed to be corrected (the others were correct, although the recognition engines were "unsure"). At the end of the process, 99.47% of the total number of significant characters were correctly recognised.
When the confidence threshold was raised from 80% to 90%, which marginally increases the manual correction time, the already-high quality of the exported data increased (to 99.79%). Conversely, when the confidence threshold was lowered to 70%, the number of characters which had to be double-checked dropped to 2.8%, and correction time decreases to 19 minutes. Unsurprisingly, however, the number of false positives increased, leading to a final recognition rate of 99.26%.

The results for the forms on the fax server produced no surprises. Overall, the level of quality was still very high and comparable to the level for the Scanner, but, as projected, slightly more time was required to verify the images that were not directly scanned in. Confidence threshold settings of 70%, 80% and 90% resulted in data quality at the end of the process of 99%, 99.29% and 99.67%, and total correction times of 26, 32 and 57 minutes, respectively.

In conclusion, the trial run of using Teleform to process the paper forms was highly satisfactory, and the data quality which can be achieved is fully in line with the quality standards for the production of statistics applied by Istat.

Table 1 – Industrial Production Form: general statistics

Forms	Significant characters	Blank characters	Total characters	Average number of characters per form	Average variables per form	Average interpretation time
199	28473	27363	55836	143	21	2 sec

Table 2 - Industrial Production Form: Scanner

Confidence level	Forms needing verification	Characters verified		Characters changed		Correct characters at the end of the process		Average correction time	Total correction time
		No	**% of signif.**	**No**	**% of verif.**	**No**	**% of signif.**		
70	158	797	2.8	790	99.1	28269	99.26	7 sec	19 min
80	190	1488	5.2	850	55.4	28322	99.47	9 sec	29 min
90	199	3421	12.0	940	26.0	28412	99.79	14 sec	47 min

Table 3 - Industrial Production Form: Fax server

Confidence level	Forms needing verification	Characters verified		Characters changed		Correct characters at the end of the process		Average correction time	Total correction time
		No	% of signif.	No	% of verif.	No	% of signif.		
70	191	948	3.3	826	87.0	28188	99.0	8 sec	26 min
80	191	1723	6.1	902	52.4	28270	99.29	10 sec	32 min
90	199	4162	14.6	1018	24.5	28380	99.67	17 sec	57 min

The electronic forms were tested in PDF format, as this format is very common in business environments, and because the reproduction is faithful to the original both on screen and in print. What is more, the viewer for the forms is available free of charge, and does not allow the forms themselves to be modified. Using the simple program Acrobat Reader 4.0, the forms on the Istat website (or provided by e-mail) can be filled out and returned to the data capture system in a way that is completely transparent to the user. Teleform can be directly integrated into a POP server on any platform (we used an SMTP/POP3 server), from which the data e-mailed by the respondent can be extracted. The transmission procedure is completely transparent – all the respondent need do is click on the "send" button. The electronic questionnaire produced for the test includes personal information and a large number of checks on fields. We checked that the control tables could be accessed and that the program could be integrated in the production environment. We also checked that it was possible to use an existing data capture website running on Unix (Apache 1.2.4 server under AIX 4.2), as well as our recently-purchased Topcall fax server.

In sum, the trial run was satisfactory and leads us to conclude that it would be feasible to use Teleform to improve the data gathering process.

VI. CONCLUSIONS

The trial run demonstrated the potential that could be freed up by using a complete system of data capture. Such a system would simplify the response procedure, thus making it easier to meet the requirements of the survey rapidly, while minimising the statistical burden and cost. The most significant advantages for Statistical Institutes are:

- surveys are completed considerably more quickly;
- efficiency is increased via the automation of a series of phases, freeing up resources for other activities.

For these reasons, and having completed the trial run, Istat has decided to progressively implement this system for use on short-term statistical surveys from 2001. This, in conjunction with other action, will improve the timeliness of short-term statistics in the near future.

References

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

[2] Balestrino, R. (1998), Data Capturing on the Web, Proceedings of the Seminar NTTS'98, Sorrento.

[3] Politi, M. (1999). The Fax Server as tool for statistical surveys. Proceedings of ETK-99 International Seminar on Exchange of Technology and Know-how, Eurostat, Prague.

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