

WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

D. Data processing: Effective IT strategies for data processing (including LFS files for Eurostat) and data storage

D3 – IT strategies for Data Processing and Storage of LFS at the Spanish NSO

Jorge Velasco– Spain

IT ACTIVITIES ORIENTED TO THE DATA PROCESS OPTIMIZATION IN THE SPANISH LFS

*Jorge Velasco López
SDG. CIT Directorate
National Statistics Institute, Spain*

April 2012

-
- 1. INTRODUCTION**
 - 2. IT ACTIVITIES IN THE DATA PROCESS**
 - 3. CONCLUSIONS**
-

1.- INTRODUCTION

Spanish National Statistics Institute (INE) publishes the results of the Spanish LFS a month after the end of the quarter (being the reference month the one in the middle of the quarter). Compared to the standards of other countries it is a considerably short time and it is ultimately boosted by the desire of the final users (Government, Unions, Institutions, etc..) to obtain the data as quickly as possible.

There are several factors that contribute to this short time of data dissemination, and the **optimization of the technological processes** is one amongst them.

In particular, we focus on the IT factors and activities undertaken in the **data collection stage** that allow the early disposal of the data at the next stage in the survey process, the data process.

Afterwards, we will describe these factors and activities in the **data process stage** itself, that contribute to this performance.

Regarding the **data collection** and from the technological point of view, we can highlight the use of an electronic questionnaire for pre-debugging the data; a proper management of the technological processes to ensure business continuity; the use of an effective application for the collection management that helps to monitor and perform an effective fieldwork; data consolidation to ensure the quality of the data transferred on a weekly basis and the establishment of a realistic and tight schedule of field work and data delivery.

If the data is collected with the required quality and on time, the continuity of the survey process transfers to having available a contrasted and fault tolerant **data process**.

This stage tries to optimize the technological environment used (database, programming language, etc..) and to have a good coordination among the different agents of the IT environment (Systems, Operations, Development ...). The process also relies on the deadline imposed by the publication schedule and the quality of the auxiliary files created by different units of the INE, involved in the data process.

Another key factor is to carry out the monthly processes before the quarterly one. Although the results are not published in the media, it helps to speed up the quarterly process and it is useful to contrast the programs and the data quality received so far.

Regarding the fault tolerance, INE has a support system which could be used to run the data processes if needed with another infrastructure apart from the headquarters.

The following describes these activities and factors both in the collection and in data process in the INE, which are focused on optimizing the data process in terms of performance, but also in costs and reliability.

2.- IT ACTIVITIES IN THE DATA PROCESS

2.1 Data Collection Stage

In the collection stage there are factors and actions from the technological viewpoint that influence the data process optimization:

- 2.1.1. Using an electronic questionnaire
- 2.1.2. Technological Processes
- 2.1.3. Effective field work application. Organization and monitoring of field work
- 2.1.4. Weekly download
- 2.1.5. Collection and data delivery schedule

2.1.1. Using an electronic questionnaire

In the **data collection** stage (CAPI¹ method for the first interview and CATI² for second and subsequent), the use of an electronic questionnaire ensures the quality of the information collected, because it includes online rules for inconsistency and flow validations while collection. This ensures that data received for data process are already pre-purged; thus it reduces data process time as there are fewer errors to be debugged.

2.1.2. Technological Processes

To adjust as much as possible the data process time when information is available in Central Services (headquarters) and to ensure the deadline to receive the data, there has to be a customized organization of the technological processes in the collection stage.

The system ensures business continuity at a database level, duplicating infrastructure where it deems necessary and establishing an appropriate system of data storage. At an application level, several weeks of dwellings are kept in advance in the Delegation³ servers allowing, if necessary, the continuity of the field work and the deferral of the data transfer. At a communications level, business continuity is ensured allowing the work in degraded mode if communications with headquarters fail, by establishing temporary repositories in a backup server in the Delegation, which will afterwards be synchronized with the central repository.

Security is ensured in several stages:

- To login the collection application, an authentication mechanism with a personalized certificate is required.
- When sending data, a certificate is also needed to ensure confidentiality, integrity and non rejection. Besides, data are encrypted if they come from nodes outside the intranet.

¹ CAPI: Computer Assisted Personal Interviewing.

² CATI: Computer Assisted Telephone Interviewing

³ A Delegation is a regional office. There are 52 in Spain at a NUTS-3 level.

- A certificate is also required to login the interviewers' tablets and their stored information is encrypted.
- In the intranet, all offices have their own perimetral security network based in firewall.
- Finally, management, operation and support lines for remote maintenance are safe and they are based on the use of a virtual private network (VPN) and remote connections are made via secure shell.

At this stage there is a strict policy of backups, both in Delegation servers and tablets and in Central Services (headquarters) in the data consolidation process. Finally, process and infrastructure are continuously monitored and there is a maintenance support.

2.1.3. Field work Application: effective field organization and monitoring

The application used assists in the collection stage to perform its primary function of collecting and monitoring these data, as well as all other associated features such as sample management, resource management, monitoring listings, etc.

Thus, it allows continuous monitoring of field work and in case of CATI collection method, the interviewers work in real time, so the collection process can be adjusted closely to the fieldwork.

Furthermore, the application has a test and training environment that optimize the development of new functionalities and new interviewers' training.

2.1.4. Weekly download

Once consolidated from all regional offices, information is transferred weekly to the Central Services server. Thus, data process stage initiates.

2.1.5. Collection and data delivery schedule

This weekly download, as established in the schedule, allows to receive information quickly regarding the reference period of the collection and it allows the evaluation of the key features of this data received and the potential need for changes at the stage of the data collection.

The calendar includes deadlines for fieldwork and data transfers according to the reference week.

2.2. Data process stage

The main actions at this stage are:

- 2.2.1. Availability of auxiliary files
- 2.2.2. Optimization of the procedures used
- 2.2.3. Monthly processes
- 2.2.4. Support systems
- 2.2.5. Availability of resources and coordination
- 2.2.6. Publication Schedule

2.2.1. Availability of auxiliary files

Data process requires some files that take part in different phases, like the Geographic Dictionary (with the sections in the sample), the Delivery Schedule, the File of Sections to Repeat and the Population File. The last two files are used to calculate the raising factors.

These files must be provided on time by the other INE units to carry out the process on time.

2.2.2. Optimization of the procedures used

Related to the technological environment, the programming was improved so that processes and database performance were optimized, using an unique key, reorganizing DB2 tables, indexes, etc.

The data process has also been configured to include an adaptation for the LFS of the Automatic Debugging Software (DIA⁴) used in several statistics in the INE.

In addition, the process has been adapted to be more dynamic, reducing the phases and therefore the tasks that other units perform.

2.2.3. Monthly processes

Monthly Processing is a keypoint to achieve very good timing when calculating quarterly results, because there is an early detection of potential problems or inconsistencies in data input.

Thanks to the monthly process part of the quarterly registers are already depurated when quarterly process begins. In particular, there has already been edited and checked 60% of the registers involved in that process.

Finally, apart from providing a file for the Labour Market unit to be used in monthly estimations, programs are checked out every month and problems detected in the 1-8 weeks are already solved.

2.2.4. Support systems

A support system has been implemented in case communications or computer systems in the INE, fail.

This system, which is reviewed quarterly, would ensure business continuity by operating against a mainframe in another environment, in which processes to be undertaken in Central Services are replicated.

2.2.5. Availability of resources and coordination

⁴ DIA stands for Depuración e Imputación Automática (Automatic Debug and Imputation).

Human resources are available for application maintenance and update, either to develop new functionalities that have to be included in the process, or to troubleshoot process issues.

Finally, there is a coordination in the different IT units involved in the processes, in order to ensure the realization and no overlapping of all the tasks to be performed, such as backup processes, database maintenance, etc., relying in an IT working schedule.

2.1.5. Publication Schedule

This schedule shows the dissemination dates for the quarterly results, the press release, the Spanish LFS files, microdata files, tailored tables, etc.

These results are published two weeks after the closing date of the quarter. This deadline requires the data process optimization.

3.-CONCLUSIONS

1. INE disseminates the results of the Spanish LFS a month after the end of the quarter that is published.
2. There are several factors that contribute to this short time of data dissemination. The **optimization of the technological processes** is one of them.
3. The IT factors and actions undertaken in the **data collection** stage, permit an early disposal of the entry data at the data process stage. Some of these are the use of an electronic questionnaire, the optimization of technological processes, an effective fieldwork application, the weekly download and finally the collection and data delivery schedule.
4. In the **data process** stage, different factors and activities contribute to obtain in a short period of time the final results with a standard of quality and reliability. Some of them are the availability and quality of the auxiliary files, the process optimization, the monthly process, the availability of support systems, resources and coordination, and the dissemination schedule.