CENTRAL AND LOCAL SURVEY ADMINISTRATION THROUGH COMMUNICATING DATA SYSTEMS

SUMMARY

Like any new data processing system, CAPI, which is used for household surveys at INSEE (Institut National de la Statistique et des Etudes Economiques - the French national institute of statistics and economic studies), has significantly changed the work of the various people involved: survey designers, survey administrators and interviewers.
The purpose of this contribution is to present INSEE's reflections regarding the organisation of the work of survey administrators following the reduction in data cleaning tasks linked to the introduction of CAPI. It involved evaluating the possibility of centralising data cleaning at the national level.
Some 150 administrators are employed in INSEE's 18 regional offices (RO) in Metropolitan France who manage a network of household interviewers (the 2 overseas regional offices have a specific organisation of their own). Their task is to organise the collection of data, manage the interviewers and clean the data collected at the questionnaire level.

The two main advantages of CAPI, compared to a questionnaire in hard-copy form, are data capture and verifications at source (on the interviewer's laptop).
The staff who were previously responsible for these two functions in the regional offices have either seen their activity disappear as far as data capture is concerned, or in relation to checking the data collected have seen a significant change in their work.
As the checking process now mainly takes place at the level of interviewers, the survey administrators' task relating to checking and correcting data from the questionnaires – data cleaning - has become extremely limited.

In the case of the administrators employed in the regional offices, data cleaning using CAPI consists in checking the data on the questionnaires collected from the households by the interviewers and correcting the data, where necessary. Each statistician-survey designer defines the checks to be carried out for his survey. They may be carried out mainly by means of:
- non-blocking checks confirmed at the time of collection;
- supplementary checks carried out on the administrator's workstation (with a different data model than the one on the collection device);
- "don't knows" or "refusals to answer";
- average interview time;
- the interviewers' comments attached to the questions, or a general comment attached to the questionnaire.

In the case of certain surveys, a second-level data cleaning, carried out by the statistician who designed the survey, is implemented in respect of certain questionnaires selected on the basis of predefined criteria. A third variant of the data model may be necessary if the checks are different from the ones to be used by the administrators.

Thus INSEE's first idea was to centralise this work at the national level (or in some regional offices), which would only consist of dealing with complex problems that had not been resolved by the checks on the interviewer's laptop. This work would necessitate a limited number of specialised staff. This organisation was expected to engender operational savings and improvements in quality (due to the skill of the personnel and the increased uniformity of the processing). The survey administrators in the regional offices would have at their disposal statistical tables for each interviewer for monitoring the collection and the quality (number of questionnaires answered, refusals to answer, questionnaires out of scope,...; average number of inconsistency messages ignored and unanswered questions, ...; average interview time, ...). They would not have access to the questionnaires, which would be transmitted directly to the national centre by the interviewer.

The heads of the regional household survey departments and statisticians-designers were consulted about this proposed new organisation of the work.

The results of this consultation revealed that the survey administrators' two main tasks - monitoring /evaluating the interviewers and checking data - were closely intermingled: data cleaning is an indissociable component of monitoring the interviewers' work. Through data cleaning, the administrator evaluates and checks the quality of the collection work. The administrator can then authorize the payment of the questionnaire.

The consultation showed that nationally centralised data clean would not spare regional administrators the necessity of checking some questionnaires, and would in fact lead to the duplication of work that is often closely

related: the opening and analysis of questionnaires, at the national level for the checking and possible correction of data, and at the regional level for monitoring the interviewer's work (particularly by means of accompanying interviewers during the survey and by *a posteriori* checking with the interviewees, by post or telephone). If the administrators could not have access to the questionnaires, then a system for returning information to them would be necessary, and this would be complex and expensive. Lastly, the administrator's access to the questionnaires and his knowledge of the interviewer's errors is an essential part of human resource management. The interviewer knows his administrator, who is nearby (thus precluding administration at national level) and he goes to that administrator for support and advice. The administrator therefore needs to be in a position to monitor and know about the whole of his work. INSEE took the view that this was essential for preserving and enhancing the quality of data collection by its network of 850 interviewers.

XXX

The results of the study are set out below. These present firstly the problems raised if nationally centralised data cleaning is implemented in place of the current decentralised system divided between the regional offices. The second section compares the two systems, and shows that centralised data cleaning cannot be brought into general use for all types of surveys.

The plan used is as follows:

I. - THE PROBLEMS RAISED BY CENTRALISED ORGANISATION OF DATA CLEANING

1. - 1 The need for the regional collection offices (RO) to be able to consult and keep all the questionnaires
1.- 2. The needs for return of information from the data cleaning centre to the collection RO
1.-3. Contacts between the data cleaning centre and interviewers for data cleaning purposes
1. - 4. Contacts between the data cleaning centre and the collection RO

11 - COMPARISON OF THE TWO DATA CLEANING SYSTEMS: CENTRALISED OR DECENTRALISED

II. - 1. Data cleaning as a factor of monitoring the quality
II- 2. Data cleaning at the heart of survey management tasks
II - 3. Quality of the checking and correction of the data collected
II - 4. Diversified organisation of data cleaning depending on the survey concerned

I. - THE PROBLEMS RAISED BY CENTRALISED ORGANISATION OF DATA CLEANING

I. - 1. THE POSSIBILITY FOR THE COLLECTION RO TO CONSULT AND KEEP THE QUESTIONNAIRES

It seems essential for the questionnaires to be transferred to the collection RO, and kept there, in parallel with their transmission to the data cleaning centre, **to ensure that the interviewers' work is monitored** and for getting them to correct their errors during the survey - as soon as possible after the start of data collection. It is necessary for the questionnaires to be available in the collection RO to enable interviewers to be accompanied and the *a posteriori* checks to be carried out. It should be noted that organisation of a centralised data cleaning process would preclude the collection RO from modifying the questionnaires.

IT development work necessary would remain limited, and would be directed towards sequential transmission, with the collection RO receiving the questionnaires several hours after the site RO. Transmission after data cleaning is excluded, which although it would reduce the burden on the circuits would nonetheless result in the loss of a significant factor by preventing the RO from having the questionnaires as soon as they were transmitted by the interviewers.

I - 1.1. **The quality indicators,** necessary for pinpointing the questionnaires that need to be checked, **might not eliminate the need to consult questionnaires** for more detailed analysis. Moreover, the RO should be able to read any comments attached to the questions and the general comment, which generally denote a problem. The comments system in CAPI must be used fully by the interviewers. Its aim is to inform the administrators of their difficulties and their choices, for monitoring and data cleaning purposes.

In the "Assets" survey, the reason for a particular interviewer's short average interview time was explained, after his questionnaires were opened, by the fact that he was conducting his interviewing in HLM (social housing), where there were few positive replies regarding the holding of financial assets. In the "Journeys" survey (regional and paper-based, not in CAPI) an indicator of the mobility rate appeared low in the case of one interviewer in comparison with the average: after consulting the questionnaires, it emerged that a large proportion of retired people with a lower mobility rate than the average of the population, lived in this district. An indicator of a quantity of "don't knows" in reply to important questions makes it essential to consult the questionnaires, as does a quantity of non-blocking ignored checks.

A widespread custom in the regional survey departments - often on the instruction of the statistician-designer - is to analyse each interviewer's first three questionnaires, which he has to transmit rapidly, except where applicable in the case of recurrent surveys. This method is used to ascertain that the training has been assimilated and to correct any errors very quickly, particularly in the case of interviewers who have shown difficulties during the training. This analysis is the (or a) determining factor for deciding which interviewers to accompany at the beginning of the data collection. The experiences derived from actual cases will be kept for use in future training programmes (in the case of periodic surveys).

I 1.2. **The availability of the questionnaires in the collection RO is also important for its relations with the interviewers**. The interviewers need to know that they can obtain advice there. They may need to discuss certain questionnaires or certain problems encountered, over and above the comments procedure. More generally, it is important for the interviewers to know that their collection RO receives the product of their work. It is the ROs that manage them and supervise them. This consideration also has to be taken into account by the administrators.

I.1.3. In an organisation with a centralised data cleaning process, we have shown **the need for the RO to be able to consult the questionnaires. But they would not be able to modify them** - a task that would be carried out by the national centre(s).

Many of the people consulted think that this separation of the tasks is a factor of demotivation for the administrators in the regional survey departments. Opening the questionnaires, possibly spotting errors in them without being able to correct them themselves can be frustrating and may finally discourage them from consulting them. Moreover, the discovery of errors in the questionnaires should be notified to the national data cleaning centre.

In fact we are faced with what may appear to be duplication: the collection RO and the national data cleaning centre will each open questionnaires that raise problems, the former for monitoring the interviewers, and the latter for the

possible correction of data. One person consulted has devised a substantial range of monitoring and quality indicators to preclude the collection RO from opening the questionnaires. But would not this amount to installing a parallel system, using up equivalent resources?

Another suggestion is to consider that the consultation of questionnaires is different in character: general in the case of the collection RO, and highly targeted (on the variables to be checked) in the case of the national centre.

I. - 2 RETURN OF INFORMATION FROM THE DATA CLEANING CENTRE TO THE COLLECTION RO

I. - 2.1. The return of information is considered to be **just as essential as the transmission of the questionnaires** and for the same reasons, which **relate to monitoring the interviewers' work and quality**. It must comply with strict management rules establishing regular returns, especially at the start of collection, to enable reaction with the interviewers. The accumulation of knowledge resulting from the data cleaning will also be used for subsequent training programmes: in the case of the "Employment" survey, the collection ROs use lists of anomalies from the previous year (imprecise profession, incorrect adress of the firm, etc.) for the current year's training programmes. The administrators rely on cases encountered during data cleaning. One RO estimates that a quarter of the training time is devoted to examining and correcting errors committed the previous year.

This return of information might be done in the form of standardised detailed tables, each having a single objective (rate of data cleaning per interviewer, questions raising problems, etc.) and enabling comparison with other external data (e.g. division of the population by sex). But if the information is subtle, the administrative workload would be heavy for the national centre. Lastly, the fact that the RO would only have the original of the questionnaires (before any modifications made in the course of the data cleaning process) may be a disadvantage, for example for *a posteriori* checks.

The definition of these specifications would be a difficult task (the quality monitoring indicators would certainly have a subjective component), as would the IT developments necessary. The data transfer circuits would be encumbered by these returns to the RO, especially if there are several national data cleaning centres.

Lastly it would also be necessary to take into account in the RO's workload, the transmission to the national centre of errors that the RO have discovered during the consultation of the questionnaires, since they will not have been able to modify them.

The notion of data cleaning shared between the national centre and the collection RO, for example by having easier cases processed locally and complex cases by the centre, has not been envisaged. Apart from the difficulty of assessing the nature and importance of each of the cases, the complexity of the IT circuits to be implemented would rule this idea out altogether.

I. - 2.2. It should also be possible to standardise **the necessary, final information return phase to the interviewers** to a certain extent. The present situation differs from one RO to another. Some publish tables detailing each interviewer's errors. Further, the centralised organisation of the data cleaning process and the necessary return of information to the local centre would lead to increased saturation of the circuit from national centre - via collection RO – to interviewer, and this problem would have to be solved.

I.- 3. CONTACTS BETWEEN THE NATIONAL DATA CLEANING CENTRE AND THE INTERVIEWERS FOR DATA CLEANING PURPOSES

The consultations reveal that direct dealings between the national centre and the interviewers would be detrimental to the smooth running of the survey network, which is based on the unique relationship between the collection RO and the interviewer.

The collection RO should be the interviewer's one and only contact for relationship-related reasons on the one hand, and for reasons of organisation and efficiency of the work, on the other hand.

I. - 3.1 ASPECTS CONCERNING CONTACTS WITH THE INTERVIEWERS

It is essential that it is the individuals who have usual direct personal contacts with the interviewers who get in touch with them for the needs of data cleaning, in other words the person in charge of the survey or their direct

administrator, that is the unit that manages them. An interviewer can only receive instructions from someone near, that he knows. At present problems with interviewers result from lack of proximity. And proximity procures better quality.

Only the collection RO can ensure that requests for information are accepted by the interviewers, since the RO knows them, it trained them and it pays them. It is their immediate superior. Not to abide by this principle would represent a risk for quality, by weakening the team spirit built up between the collection RO and its interviewers. There is a strong psychological component in the relationship with the interviewers. It is necessary to know how to talk to them, to know their character and personality so as to avoid upsetting them. The national centre would simply look like a controller, and receiving criticism from such a person would be a bad experience for the interviewer. He would also be disconcerted by these two levels of administration and would wonder by whom and how his work is to be judged.

Moreover, **direct contacts by the national centre with the household are not considered desirable**. It has to be the interviewer who calls the household because they know each other already and relationships of trust have been established. Otherwise, it would have to be the RO to which the interviewer was attached, because the national centre is distant and can engender suspicions on the part of the household. Moreover, the interviewer as well as the collection RO would be bypassed.

I. - 3.2. ORGANISATION AND EFFICIENCY OF THE WORK IN THE NATIONAL CENTRE AND IN THE COLLECTION RO

Independently of the reasons set out above, direct contacts by the national centre with the interviewers would result in the national centre achieving lower productivity in its work of collecting information from the interviewers than the collection RO achieve:

- as the centre does not know the interviewer, it would have greater difficulty in obtaining satisfactory replies;
- it does not know the interviewer's timetable, hence when he can be contacted - the RO already have difficulty in contacting their own interviewers;
- as interviewers often work on several surveys at a time, they would experience an increase in the number of people contacting them that they would not know;
- the answers to the questions asked of the interviewer might perhaps have been obtainable from the collection RO.

The interviewers' relationships with the collection RO would also be distorted, thus damaging the quality of their joint work: on the one hand, the RO would lose their "power" over the interviewers, while on the other hand, the interviewers would tend to consider the national centre as their only contact.

The interviewers would accordingly be naturally prompted to build up a single relationship, as was seen in the RO when they applied directly to the RO's computer services department for loading the applications: they stopped visiting the survey department. The collection RO would then run the risk of having no return of information. It would have no knowledge of conversations between the interviewer and the national centre, and would no longer be able to monitor the interviewer correctly.

I.- 4. RELATIONSHIP BETWEEN THE NATIONAL DATA CLEANING CENTRE AND THE COLLECTION RO

We have just seen the negative consequences of possible direct dealings between the national centre and the interviewers. Unless it is decided that data cleaning must be done solely in the office, in other words autonomously within the national centre, the centre will have to obtain the necessary information from the collection RO - we have already seen the disadvantages of direct contacts with the interviewers, or even the households - , with the RO having the task of contacting the interviewer (or the household) where necessary.

The consultation has shown that **such relationships would be complex, posing the question of the benefit procured by centralised organisation of data cleaning, and that they would run the risk of rapidly ROying up, culminating in fact in data cleaning in the office at the national centre.**

The data cleaning process in CAPI is significantly reduced in comparison with traditional data cleaning, because the checks are integrated into the collection procedure. However, while the problems posed at the time of data cleaning are few, they are more complex. The national centre will have to find the relevant contact in the RO, who will not always be available (absent, part time, busy, etc.).

II - COMPARISON OF THE TWO DATA CLEANING SYSTEMS, CENTRALISED OR DECENTRALISED

As has been already pointed out in the previous section, ***the tasks grouped together under the word "data cleaning» have two functions: on the one hand to check and correct the data; and on the other hand, to monitor, evaluate and correct the work of the interviewers. These two functions are aimed at a single objective, quality.***

The monitoring of interviewers can only be done at the level of the collection RO, because it calls into play human relationships that call for reciprocal knowledge and proximity. This point has also been referred to above.

With centralised data cleaning, the national centre provides for the first function alone, sometimes with recourse to the collection RO. The regional office provides the second function, if there is a return of information.

It seems that we end up with an alternative: time saving versus quality. Indeed, in order to guarantee quality, the RO must be able to consult the questionnaires and exploit returns of information. In other words virtually do data cleaning work.

But data cleaning can be organised in more than one way and be adapted to the particular features of each survey, while at the same time an arrangement can be set in place that is as simple as possible, for reasons of rationality, in particular concerning IT developments and maintenance.

II. - 1. DATA CLEANING AS A FACTOR OF QUALITY MONITORING

For the collection RO, consulting the questionnaires pinpointed by the monitoring table indicators and comments, in fact boils down - at least in part - to carrying out a data cleaning task, except that they cannot modify the questionnaires. This work would continue during the exploitation of the returns of information in the case of centralised data cleaning.

In CAPI traditional data cleaning is reduced. But at the level of the collection RO, good interviewer monitoring necessitates consultation of the comments and confirmed checks, which will also be used as part of the content of subsequent training programmes (periodic surveys). Data cleaning enables malfunctions to be spotted in the course of collection so that the interviewers can then correct them. "*One becomes aware of the quality of the interviewers when one cleans data* ". When an interviewer is being accompanied by an administrator from the survey department (for the purpose of advice and evaluation), the administrator already has an impression of the interviewer's work, which he has been able to evaluate during the data cleaning process.

II. - 2. DATA CLEANING AT THE HEART OF SURVEY MANAGEMENT TASKS

Data cleaning is the last phase in the process of preparation, carrying out and monitoring a survey entrusted to an administrator. This process is a comprehensive whole which makes it possible to follow a survey operation from A to Z, right up to the supply of the data files to the designer. **For the administrator, centralised data cleaning would represent an amputation of part of his work**, which would mean that he would no longer be able to clearly appreciate the whole picture, and which would dispossess him, deprive him of responsibility and demotivate him.

The reduction of the data cleaning tasks linked to CAPI might lead to increased diversity of the administrators' tasks, which would be enriching and in the long run beneficial for the RO. Conversely, centralised organisation would be like assembly line work, with repetitiveness of tasks - especially in the case of ongoing surveys.

So far as training programmes in data cleaning are concerned, centralisation would reduce the costs of this. However, should one not take the view that training for a survey is a whole, and that data cleaning aspects answer questions that all the survey's administrators need to know, for advising and monitoring the interviewers?

II. - 3 QUALITY OF THE CHECKING AND CORRECTION OF THE DATA COLLECTED

Some of the people consulted consider that, in the case of those surveys that are still in the form of paper questionnaires, decentralised **data cleaning outside CAPI** in the various RO (for most surveys), results in **disparity in the quality of corrections**, and in saying this they are not calling into question the competence of the

personnel concerned. Centralised data cleaning would guarantee uniformity of processing, and would perhaps ensure a more rigorous follow-up of the data cleaning instructions. It should concern only difficult cases: significant number of inconsistency messages ignored, where a variable far exceeds the relevant threshold (e.g. rent > FRF 10,000), etc. Such cases might call for a very specialised skill, which all the RO might not have for all types of surveys.

Conversely, **data cleaning in CAPI automatically imposes a strong uniformity of processing**, hence fewer discrepancies in quality between the RO. On the other hand, a higher ranking administrator would be needed for supervising the work of the RO, for example the statistician-designer or a specialised RO, who oversees several hunROeds of questionnaires during the collection process (e.g. looking at how the RO processed the "don't knows" during data cleaning).

The section relating to the consultation of questionnaires by the RO (§I.-1) has shown by means of examples, the importance of proximity for better monitoring of the interviewer's work. This also applies to data cleaning as such. In a survey on household living conditions, where an interviewer has collected data on attacks and thefts incorrectly - and even if he appeared reliable when accompanied on the survey - this is only likely to be discovered by the local administrators, because they know the district where the interviewer operates. Likewise in the case of data relating to rent.

In return, the RO should pass on any problems of correction they have encountered to the designer, so that everyone benefits from the experiences of cases already dealt with.

II. - 4. DIVERSIFIED ORGANISATION OF DATA CLEANING DEPENDING ON THE SURVEY CONCERNED

Beyond the general views on data cleaning presented above, the consultations with certain survey designers have highlighted differing opinions, which can be explained by the particular characteristics of their surveys.

This leads to the concept of different data cleaning arrangements depending on the survey concerned.

In the opinion of the designers of the "Household assets" surveys and ongoing surveys of living conditions, the data cleaning process should be done at local RO level.

The designer of the "Rents and charges" survey is in favour of centralised data cleaning. His reasons seemed linked to the specific characteristics of that survey:

- Depending on the regional office concerned, the survey is run either by the survey department or by the prices department, which have different constraints. This leads to lack of uniformity in data cleaning which would be eliminated with centralisation.

- The deadlines are short, linked to the production of the rents - and prices - index every three months, leaving no time to go back to the interviewer to inform him of his errors, at least for the survey wave in question. The interviewer takes the comments into account for the following quarters. Occasional telephone calls are made by the administrators to ask for details. Moreover centralised data cleaning would undoubtedly save time, a fundamental aspect for the designer concerned.

- The (quarterly) survey takes place over one month, with 300 interviewers. Half of the interviewers have fewer than twenty questionnaires. **Hence the monitoring is only done on a small number of questionnaires**. It should be noted that provision would be made for consultation of the questionnaires by the RO.

- In small regional offices there is only one administrator for the survey, on account of the small number of questionnaires and the light workload that each represents. This poses the problem of the **minimum staff and workload required.**

- Lastly, data cleaning is humdrum work, of less interest to administrators than, for example, the ongoing survey of living conditions, part of which is different each quarter. Moreover, after the interviewers' CAPI apprenticeship period (rent bill posts), they will do very little data cleaning, and the designer will take charge of it.

XXXXXXX