AFTER 10 YEARS OF CAPI, INSEE EMBARKS ON THE CAPI 3 PROJECT

Christophe Alviset, INSEE
Mario Avi, INSEE
Philippe Meunier, INSEE

Plan:

I – Broad outlines of the CAPI 3 project.
II – The future data collection work-station, a tool for communication
III – Data coding integrated into the data collection work-station ?

I – BROAD OUTLINES OF THE CAPI 3 PROJECT

Following initial tests conducted between 1987 and 1989, it was in 1990 that INSEE (France's National Institute of Statistics and Economic Studies) began work on the CAPI project, with the employment survey. Blaise 2 software was used. The next few years, through to 1999, saw all household surveys gradually transfer over to CAPI. At the end of the last millennium, only surveys for which an electronic questionnaire was not suitable (surveys involving the homeless, regional surveys performed on small samples, etc.) still remained in a paper-questionnaire format.
10 years after, the new CAPI 3 project that INSEE is going to develop, will not create any upheaval in the process of designing and conducting a survey in CAPI. Nevertheless, it is still an ambitious project in terms of the computer-related workload it will generate, as well as the organisational changes involved for statisticians, interviewers and survey managers.
As a result, analysis is already at an advanced stage in terms of the objectives to be achieved, highlighting four aspects of development.
The first aspect will involve providing statisticians and interviewers with an Extranet information system, including an e-mail program, access to the INSEE website and a common forum. The use of electronic cartography and GPS satellite positioning will be tested.
Secondly, the system will incorporate Blaise innovations (Blaise 4 and Windows). Finalisation of the TADEQ project (electronic questionnaire documentation) is eagerly awaited. Survey taking via the Internet will be tested and implemented, together with all the necessary precautions, where relevant.
Thirdly, it will involve introducing developments corresponding to the requirements identified in recent years. We may mention here:

- the possibility, on a statistician-designer work-station, of monitoring data collection at national level;
- organising auditing arrangements, differentiated according to surveys;
- the involvement of statisticians in Blaise programming;
- preparation of a single source data model, for the three different types of use involving the interviewer, the survey manager at regional level and the statistician-designer;
- integration into CAPI, of the management of the samples by the interviewer and supply of the data required for the purpose of their remuneration;
- construction of a standard survey kit for regional surveys.

Lastly, the final theme relates to developments that might be implemented following validation by an in-depth organisational study. This will involve the deployment of CATI, which has not so far been used at INSEE, or the routine introduction of coding at the source, i.e. on the data collection work-station. The idea of coding at the source, which has been much debated among statisticians, has been the subject of various consultations with designers, a summary of which will be found in Part Three of this paper.

II – THE FUTURE DATA COLLECTION WORK-STATION, A TOOL FOR COMMUNICATION

While the vast majority of INSEE interviewers have liked the new method of electronic data collection, using Blaise software, over the course of time, they have not failed to denounce its imperfections and weaknesses, both individually and collectively. This led INSEE, from 1998 onwards, to initiate various studies (audits) aimed at producing an assessment of the transfer of surveys over to CAPI. The interviewers' reaction was a unanimous one: the idea of returning to the paper questionnaire was simply out of the question! This assessment led to the groundwork being laid for a new project, CAPI 3, which is both intended to make up for the weaknesses of the current system and to innovate strongly, by incorporating everything that appears to be sufficiently tried, tested and reliable among the new technologies.

The starting point for this analysis was to allow any INSEE employee and by extension, any interviewer employed on temporary contracts to access the INSEE information system from outside, to enable them to do their job better. This implies both an analysis of the type of work-station and of communication requirements, while bearing in mind constraints governing authorisation for access and security.

If we are to achieve this goal, we shall need a clear vision of the software and hardware systems to be deployed, following an exploratory study of the new information and communication technologies, in order to arrive at a documented statement of requirements.

To prevent us being overtaken by day-to-day events and constraints, it was decided to organise 2 round table sessions, with ten or so interviewers being invited to each session, in order to catalogue their experience, their requirements and their expectations, and to compare them with the Institute's own objectives and the currently available technological possibilities.

The main expectations expressed by CAPI interviewers can be listed under three headings:
- a very strong need for communication with INSEE and with fellow colleagues;
- simplifying the tasks of pinpointing households on the ground and gaining access;
- easy-to-use, tried and tested technologies.

II - 1 – A VERY STRONG NEED FOR COMMUNICATION WITH INSEE AND WITH FELLOW COLLEAGUES

This is not a new requirement. Internal communication at INSEE, as within other public-sector organisations and the private sector, involves the widespread use of Intranet systems. Thanks to the extranet, these systems are gradually being extended to allow their use by external partners. They could be opened up to freelance interviewers. This would allow the installation of an open e-mail program functioning with INSEE and other interviewer colleagues, globalisable and programmable information transfers, and forums.
At the moment, interviewers have an e-mail program integrated into the CAPI application on their work-station,. However, this is a very restrictive and cumbersome method of communication: it only allows each interviewer to communicate with the establishment of the region they work in, and moreover, within the framework of a single survey (there are as many e-mail programs as there are surveys, being managed simultaneously on a single work-station).
The new project involves providing interviewers with a universal e-mail program, similar to Microsoft Outlook, enabling them to communicate both with one another and with their managers, independently of any work in progress and allowing them to log on to the central systems (e-mail program, databases, applications) in real time, from wherever they happen to be at a particular time.

Communication between interviewers, in addition to the social aspect, should contribute to the exchange of information, assistance and advice, and thus consolidate the feeling of being involved in a team effort. Interviewers should be able to access an indicator showing the overall amount of progress made by the survey in which they are taking part, so that they can judge their own progress. Having a direct link with regional management assumes that the latter will be available and able to react promptly. Management should certainly be capable of meeting an urgent need, and providing logistical support by remedying any technical problems within 24 hours, thanks to contracts signed with suppliers. No doubt this calls for harmonisation of regional staff on-call policies, or maybe even for support to be provided at national level, rather than having local staff on call. This is the appropriate response to the very strong need for communication with INSEE and with fellow colleagues. It has to help remove the feeling of isolation encountered in situations involving difficult surveys (e.g. sociological surveys such as disability-invalidity-dependence). The widespread provision of mobile phones to interviewers, which should go hand-in-hand with comprehensive geographical cover in terms of accessibility, will round off these arrangements.
This communication flow will also help to strengthen the bonds of trust with INSEE. It will carry offers of work from regional management and transmit urgent information. The data disseminated by interviewers will include expenses claims, any supporting documentation requested and feedback involving follow-up and difficulties encountered. Among interviewers, it will allow information and feedback to be shared quickly, and provide help and assistance for new interviewers in the field (guidance and mentoring), as well as accompanying them in the field for a day, in order to understand the nature of the area concerned. It will be a way to provide release from stress and tension, by sharing experiences.
These arrangements will go hand-in-hand with the organisation of periodic meetings, systematic survey review meetings, annual meetings, sharing of information and exchanging experience among small groups (useful tips and advice), internal cohesion seminars, and pleasant moments shared with fellow colleagues. It will be supplemented by the provision of different media for interviewers: INSEE documentation, brochures and leaflets.

The same applies to communication in general. Nevertheless, interviewers have to download their own programs covering surveys and special procedures, such as anti-virus software updates, and transmit their completed electronic questionnaires. This involves the globalisable and programmable transfer of information. At present, this transfer takes place survey by survey, from the interviewer's

own home, via their telephone line, manually (using a non-programmable method). Hence the complaints made by interviewers, who see this as a constraint and are deprived of the use of their own telephone line during the transfer period, even though the cost of the call itself is, of course, paid for by INSEE. It would therefore be desirable initially to be able to program these transmissions so that they can take place at night for example, and to globalise them, e.g. by automatically interlinking any transfers of completed questionnaires related to different surveys and program downloads. At the same time, it may be possible to develop wireless (WAP-type) transmissions sent from mobile phones, as soon as the performance of this technology reaches an acceptable level.
A forum would involve both creating a place for exchanges and discussion between colleagues (where they can exchange useful tips and advice, for example) and providing all the data that interviewers may need (programs, procedures, updates). To some extent, it is therefore a variation of the systems described previously.
Finally, within the framework of the survey system and the intermittent working pattern followed by most interviewers, via appropriate identification we need to be able to limit the access rights of every interviewer to the periods defined by the contract of employment binding them to INSEE.

II - 2 – SIMPLIFYING THE TASKS OF PINPOINTING HOUSEHOLDS IN THE FIELD AND GAINING ACCESS

Paradoxically, the problems linked to pinpointing the dwellings surveyed did not give rise to any particular comments in terms of the pinpointing techniques used. These were not formally requested and a number of new difficulties will undoubtedly emerge with the implementation over the next few years of a new method of conducting an ongoing census of the population. Part of this census will be conducted exhaustively, and part by conducting a survey (municipalities of 10,000 inhabitants or more).
Pinpointing of the dwellings to be surveyed relies on two sources and is based on two sampling techniques. The two sources are the survey frame constituted by the last census performed and the administrative file of new dwellings. Both techniques consist of a sampling involving several degrees known as the master sample (applied to most households surveys other than employment surveys) and an areolar sampling made up of areas containing about 20 dwellings (in the case of employment surveys). The problems arising differ in terms of their nature and their scale in both cases. The first case involves pinpointing a specific dwelling, while the other case involves accurately circumscribing the contours of a geographical area containing about twenty dwellings and locating them all within this area.
Until quite recently, regional management was in possession of all the paper documents related to the last census (1990). These documents enabled the search for a dwelling that was being surveyed but was not clearly identified, to be refined, as needed, via the consultation of adjacent documents. From now on, this will no longer be possible. Paper documents are being centralised and transcribed onto magnetic media using optical character-reading techniques. The information required for pinpointing dwellings to be surveyed (address cards) is printed out centrally and sent to regional management. A certain loss of clarity may therefore be associated with this transcription (which we are endeavouring to reduce as far as possible) and in the event of a difficulty, it will not be possible to search the surrounding area.
Assistance with pinpointing using the GPS (global positioning system) or any equivalent system, associated with electronic cartography, would no doubt merit consideration. This has not been the case in the past, since no genuine requirement has ever been ascertained and maybe also because the pinpointing involved was not regarded as sufficiently accurate. Indeed, GPS is probably not accurate enough for urban use. A test conducted in the department of La Réunion for the purpose of the census was inconclusive and the idea was abandoned. However, this department uses géographic information system software, loaded onto a laptop.

Finally, in the future, and as has just been stated, the censuses conducted in France will employ a new methodology (the updated population census or UPC), functioning continuously, and involving a technique of surveying 40% of the dwellings in an urban environment (towns with over 10,000 inhabitants). GPS-type pinpointing would also be able to offer assistance here, that would merit analysis, provided sufficient accuracy can be achieved.
On the other hand, interviewers complain quite rightly about the inaccuracy of the pinpointing information provided to them and are requesting "up-to-date" address cards. They would like to have a software program that can search for addresses without any other details, lists of residents living in a particular block of flats (a study to be performed in connection with the future updated population census), information that will help them optimise their rounds of household visits, provide street maps and passes (of the type used by post-office employees when they distribute the mail) allowing them access to buildings with door entry codes. They would like an electronic phonebook function on their work-station, and street maps of large municipalities, initially on their laptops, and later on their personal assistants (see below).
In order to facilitate access to the household being surveyed, interviewers will be given mobile phones for the purpose of making contact, particularly where they find themselves on the other side of a closed door.
They are also requesting a system to help them optimise their rounds of household visits: a study is envisaged with a view to examining a solution based on a personal assistant, if this can incorporate address cards and visit record books.
The mobile phone could be used to number the series of pre-recorded calls made from the laptop (e.g. for those types of surveys, such as the ongoing employment survey, which involve an initial face-to-face interview, followed by several telephone interviews, one after the other).
A notification letter is systematically sent out, as part of the arrangements for making contact with the household surveyed,. The Post office is putting in place a new system aimed at offering traceability for mail - "the tracked letter". This system relies on each postman entering the barcode of the letter he is distributing, and it should be operational in 2003. Hence the possibility of accessing this information i.e. tracking the progress of any letter posted. It will be tested in order to assess its benefits and effectiveness.

II - 3 – EASY-TO-USE, TRIED AND TESTED TECHNOLOGIES

Among all the difficulties reported by interviewers in carrying out their work there is one that is linked primarily to the hardware used, i.e. the laptop: it weighs too much, the batteries don't last very long before they need recharging, it is not always reliable, it takes an excessive amount of time to power up and load questionnaires, it is difficult to change from one survey to another, and data transmissions take a long time and are subject to frequent incidents. However, as mentioned above, no interviewers want to return to the paper questionnaire.
Advances in the Blaise software and the advent of Windows 2000 have considerably improved the questioning process. The progress made over the past 10 years has also improved the performance of laptops, and allowed data transmissions to be optimised and become more reliable. However, weight is still a problem, because in addition to the laptop itself, there is the mass of paper documents that interviewers have to carry around with them - address cards, documentation linked to the survey, etc.
New personal assistant-type tools, which will be also tested within the framework of preparations for the new updated population census, might be able to incorporate address cards and visit record books into personal assistants and thus considerably reduce the weight of paperwork that interviewers have to carry around.
It will be advantageous to study their convergence with mobile telephony, in order to simplify the hardware provided to interviewers.

Later on, a new generation of ultra-light laptops with a removable flash memory will no doubt become accessible, and its benefits should include enhanced data security, since the flash card can be removed from the computer after use.
Advances in investigation methods linked to an expansion of the scope of issues dealt with by INSEE to encompass social affairs, will require new functions to be available on laptops. One such example is data acquisition via an optical barcode reader, which is already possible with Blaise 4 and is undergoing testing at INSEE with the future Health survey in mind. However, the main function involved here is audio recording and playback, for some types of interviews, to allow high-quality listening (playing various sentences or texts to people and asking them to respond with their own interpretation). While observing safety and reliability requirements, it will also be necessary to gain partial access to the inside of the laptop for the purpose of carrying out exercises, supporting other applications, playing self-tuition CDs, etc.
While considerably reducing the burden of paperwork to be carried around and handled, the multiplicity of functions that can be supported by laptops and their associated hardware reinforces the need for interviewers to be able to print out from their work-station.
Finally, some interviewers are advocating the benefit of a webcam to facilitate the sharing of information with colleagues and/or their regional management (though not in connection with surveys).
It is interesting to note that interviewers make the point that the technologies implemented must be tried and tested, simple, easy to use, flexible and open-ended, so that they can accommodate progress. They must be accompanied by appropriate training, with logistics and support facilities consistent with the development of the technological resources themselves (timely repairs in the event of a technical problem). For the purpose of diagnostic analysis of any incidents, we should have a black box that could retrieve all the manipulations that have taken place.

III – DATA CODING INCLUDED ON THE DATA COLLECTION WORK-STATION ?

This involves implementing a standardised coding system on the data collection work-station, which will either be assisted or automatic, for textual answers provided by survey subjects during the course of the interview. This is therefore a system of coding "at the source". The aim is to cut costs, by removing the coding stage, which at present is carried out after data collection, and to cut lead-times by providing the survey designer with the results as soon as the survey is completed.
In fact coding is already in use on the interviewer's work-station in the case of certain questions that involve highly standardised answers (type of road, code of the municipality, department, country of birth, etc.). However, the CAPI 3 project has raised the question of a standardised coding system for all textual answers, which amounts to entrusting the entire task of coding to the interviewer, using the coding software on his or her data collection work-station.
It will be seen that after considering the benefits and disadvantages of this system with survey designers, it does not appear to be necessary for full coding at the source to be routinely included on interviewers' work-stations. Its implementation can only be partial, based on the degree of complexity of each response contained in the survey.

III - 1 – ARE PRODUCTIVITY BENEFITS REALLY ACHIEVED OR IS THE WORKLOAD MERELY TRANSFERRED UPSTREAM ?

The routine introduction of coding at the source would abolish the tasks of monitoring and coding, which are performed after interviewers have dispatched their survey data to INSEE's regional management offices. However, the workload of the interviewer during the course of the interview, on the other hand, would be increased.

First of all, before conducting the survey, he will need to be trained more thoroughly than is the case today in the various classifications. These may be complex, and their hierarchical structure and detailed contents have to be understood. Admittedly, fully automatic coding would allow all answers to be processed, though at the price of rejects that would have be recycled, as they are at present, downstream from data collection.

The interview will also take longer, leading to a higher cost of data collection. Either coding will simply be assisted, and the interviewer will have to finalise the choices on offer, or it will be fully automated, with software that will necessarily have to be powerful plus a vast reference system, taking up machine-time and increasing the length of idle periods during the interview.

Moreover, any productivity benefits due to the expertise of coding specialists at regional-management level would be lost. Conversely, there would be an improvement in quality because the interviewer could immediately specify the household's answers in cases where several options are on offer – provided that he does not influence the answers given by the survey subject too strongly.

III - 2 – SHORTER LEAD-TIMES FOR PRODUCING THE RESULTS, BUT TO THE DETRIMENT OF QUALITY ?

The second objective of standardised coding at the source, i.e. on the data collection work-station, would be a reduction in the lead-times required to provide the survey designer with the data collected. This objective should indeed be achieved - except in cases that involve the processing of automatic coding rejects, which is carried out downstream from data collection. Whether or not the data is checked subsequently to data collection, the centralised coding phase following the survey would be considerably reduced.
Consequently, there is a reduction in lead-times, but undoubtedly to the detriment of quality.

The main objection, from a statistical point of view, to coding at the source is distortion of the truthfulness of answers by interviewers, via their own prism through which they apprehend reality. They would appear to direct the answer to match their own vision and knowledge. For example, as regards a household's occupational activity, they would tend to distort this by coding it to match their own social category, according to the jobs they are familiar with. With assisted coding, they would tend to force this issue, thus introducing a bias and wasting time. This tendency might be reinforced by the survey subject's own hesitations, as he does not always know, for example, which sector of activity his employer is involved in (it would be possible to introduce a non-blocking control onto the work-station, activating a directory of all establishments and their business code, but this would burden the machine-time considerably). Whether it is automatic or assisted, coding by interviewers harbours the danger that they will truncate the answers given by survey subjects. Interviewers tend to standardise the answers given, which in turn tends to impoverish the variety of different situations. If new jobs and new diplomas come into being, interviewers will tend to code them within the current nomenclature. Thus, the task of determining the occupation of the survey subject would seem to correspond to their own knowledge of socio-professional categories, without being able to apprehend either new jobs or distinctions calling for a technical knowledge of the sector of activity in which the survey subject is involved. What tends to happen is a "smoothing-out" of the answers, which distorts the reality. Nor shall we discuss here special questions, e.g. those concerning diseases, the answers to which can only be coded by health experts.

The result of "smoothing out" answers would seem to be the absence of a gradual enrichment of the learning base, due to the interviewer's lack of expertise in these areas. For example, as far as the classification of socio-professional categories is concerned, we would see changes in occupations escaping our grasp, and we would no longer identify new occupations or defects in our own classifications. At present, we are safe as regards this situation, since the departments responsible for coding enrich the database with any changes they detect in the answers provided by survey subjects. However, one of the leading objectives of many surveys is precisely to determine transformations taking place in society, in its activities, behaviour and opinions.

Without using standardised coding at the source, one very useful tool on the data collection work-station - and fairly inexpensive to install - would be a spell-checker, which would not include any blocking control. This would be used mainly for the purpose of clearly entering the occupation of the survey subject and would reduce the failures in the automatic coding process currently applied downstream from data collection. It would also be used for clearly entering the nature of the business of the employer of the survey subject (or a household member). Lastly, it could be tested on the surname and first name of the address, but the database would require a substantial capacity.

Another objection to coding at the source is that the time involved in coding increases the length of the interview, either in order to obtain detailed information – related to the progress of the software - or as a result of idle periods while the software is functioning. Clearly, this is harmful to quality. Interview length is a key factor in any survey, the aim being to limit this to a reasonable time. Increasing the length of the interview is likely to tax the patience of the household before the interview is over, idle periods are not used by the interviewer for discussions with the household,

causing him to lose concentration, and the information required by the software might be other more important questions for the purpose of the survey.

III - 3 – CONCLUSION: THE INTERVIEWER MUST CONCENTRATE ON HIS KEY TASK: CONDUCTING AN INTERVIEW LEADING TO THE RIGHT ANSWERS

Full coding at the source, which is thought to generate productivity benefits and shorter lead-times for making the data collected available, adds an extra task onto the interviewer's key function, which is to obtain satisfactory answers, i.e. answers that do not distort reality. There would be idle periods during the coding operation, during which the interviewer would not be engaged in talking to the household. This situation is unacceptable: any time spent with the household must be devoted exclusively to them, as the interviewer's task is to conduct a dialogue with the household. The interviewer is too tense to gather information during the interview because he is concentrating on other things apart from conducting the interview. Automatic coding – which is a cumbersome task - can therefore only be envisaged away from the household.

If benefits are to be gained, it is on this focal point, reducing the survey time while maintaining the quality of data collection. Coding is also somewhat removed from the interviewer's own job. A good interviewer may not necessarily have an aptitude for coding. This would involve combining two jobs, and certainly to the detriment of overall performance.

Having said that, a certain amount of coding already exists on the work-stations of INSEE interviewers. It is straightforward, and well accepted. It is based on straightforward, known classifications. Empirical rules are suggested by survey designers:

- During the course of the interview, do not introduce coding exercises that take up a large amount of time (e.g. socio-professional category). No variable containing over 60 headings should be coded, otherwise, the coding operation would take too long, introducing an idle period into the interview. From a practical point of view, coding must take no more than 15 seconds; beyond this time, it breaks up the interview.
- Types of nomenclatures that the interviewer cannot itemise:
- Too many headings (see supra), involving too much complexity (numerous root structures) and thus time;
- Phenomena that the interviewer is unable to code (e.g. diseases, pharmaceuticals).

Seeking to make matters any more complex when visiting the household would distort the meaning of data collection. In fact, the problem is similar to that of the controls located on the data collection work-station: they are effective if they are limited in terms of their number and complexity.