This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Data Collection: Techniques and Tools

Contents

General section

1. Summary

This module provides a description of the main techniques and tools used by National Statistical Institute (NSIs) to collect data. Characteristics and peculiarities of each of them will be described together with organisational aspects to build data collection instruments and to set up, run and finalise data collection, in accordance with the sub-processes 3.1, 4.2, 4.3 and 4.4 indicated by the GSBPM model for “Phase 4. Collect”.

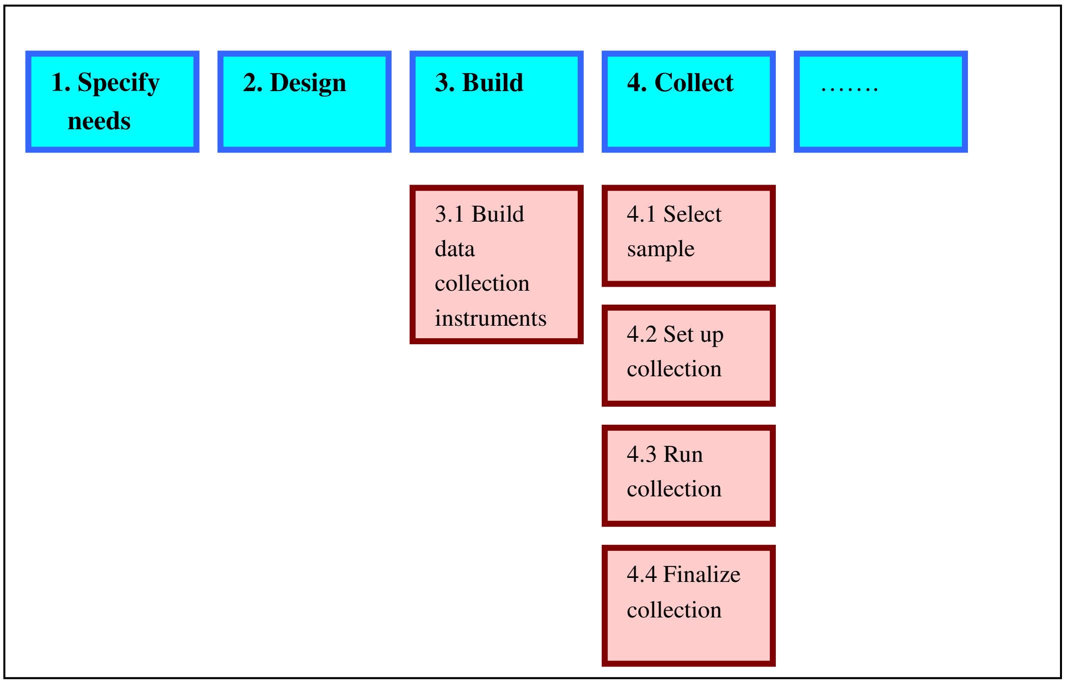

Figure 1. GSBPM – Phase 4: Collect

The choice of the most suitable technique or the way they can be combined is described in the module “Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method” where the reader can also find in section 2.2 a possible classification of all available modes (Table 1).

In this module only the main and most used techniques and software tools to collect data for business surveys are described, dividing them into two main groups: interviewer-administered and self-administered modes. More specifically, the module is organised as follows:

Section 2.1 is about interviewer-administered techniques, CATI (Computers Assisted Telephone Interviewing) and CAPI (Computers Assisted Personal Interviewing). Here the advantages and disadvantages of the presence of interviewers and of the use of an electronic questionnaire will be described. The section will also make hint to the Direct Observation mode.

Section 2.2 is about self-administered modes: Mail and Web surveys. The potentialities of the electronic questionnaires typical of web surveys will be highlighted as well as the importance of good questionnaire layout for mail surveys. The section will also talk about software tools for entering the information collected through the paper questionnaires used in mail surveys.

Section 2.3 talks about the structured electronic exchange of information based on EDI (Electronic Data Interchange) and XBRL (eXtensible Business Reporting Language) and the last section 2.4 is about administrative data, as their use is going to change the way NSIs organise their data collection process.

A note for readers: in the rest of this topic the term "respondent(s)" is used. With this term it is intended to represent all the "actors" involved in providing the information to be collected according to the surveys' needs. Respondents can be defined as "*Respondents are businesses, authorities, individual persons, etc., from whom data and associated information are collected for use in compiling statistics*" (OECD Glossary of Statistical Terms). This definition, therefore, includes all the expressions like "reporting units", "observation units", "data provider", etc. whose definition can be found the glossary, thus simplifying the reading.

2. General description

2.1 Interviewer-administered techniques

In this section interviewer-administered techniques are illustrated, focusing on CATI and CAPI modes. The last part is dedicated to Direct Observation that is treated apart since it is a particular interviewer-administered technique as it is not based on the interaction between interviewer and respondent.

Common features of CATI and CAPI are discussed right below, while characteristics of each technique (including also Direct Observation) are treated in the dedicated sub-sections.

The key features of CATI and CAPI are the presence of well-trained interviewers and the use of electronic questionnaires that, put together, allow the management of complex surveys, where complexity is in terms of survey content and structure of the questionnaire (questions, skipping and checking rules). In more details:

- the presence of well-trained interviewers allows to:
    - administer complex interviews in terms of survey content;
    - get in touch or interview difficult targets, like managers of large businesses;
    - find the right respondent especially in case one questionnaire has to be answered from different professional profiles inside the same company ;
    - collect data by directly observing a phenomenon (direct observation);
- the use of electronic questionnaires allows to:
    - have the paper questionnaire, the checking plan and the skipping rules, that are shown on the video screen of a pc, into the same software package;
    - increase data quality since editing can be performed during data collection (see "Questionnaire Design – Editing During Data Collection") and, especially for CATI, very

    complex checking plans can be managed. This aspect can also positively influence the editing and imputation phase making it simpler and faster;

  - implement a set of indicators to monitor in real time data collection, making it possible to take the corrective actions in due time;
  - reduce or avoid, with respect to self-administered techniques, follow-up calls;
  - avoid the data entry, as data are sooner available in an electronic format, thus improving the timeliness of data;
  - use data that are collected in previous waves of the survey or derived from other sources, like administrative data. In this way it is possible: *i)* to reduce the respondent burden since there is no need to ask again some information, *ii)* to improve data quality making comparison between these data and those collected during the survey;
  - use the same electronic questionnaire for mixed mode surveys thus reducing the technique effect and programming costs.

The two features (presence of interviewers and of pc) combined together offer the opportunity to use, in CATI and CAPI questionnaires, any kind of question formats, including also open-ended questions. These can be coded on-line if the software, used for the questionnaire implementation, has got an assisted-coding function. The assisted coding is quite an important feature (see "Coding – Main Module") because, although it might require a specific training session for interviewers, it guarantees high levels of standardisation and quality of the coded data since interviewers can use probing until they find a suitable code together with the respondent.

If the adoption of CATI and CAPI provides advantages in terms of response burden, data quality, amount of data entry and editing and imputation phase, on the other hand it requires considerable financial and organisational efforts. In fact:

- the presence of interviewers implies the setting up of the training and monitoring phases: the first is necessary to make interviewers able to collect data in the most objective way in order to reduce as much as possible the interviewer's effect; the second is fundamental to keep under control the interviewers' activity and to take the correct actions in due time;
- the IT feature implies some extra costs for software and hardware components: an initial cost for the questionnaire implementation and the creation of the entire IT infrastructure, plus a cost for testing and maintaining the applications. In particular for CAPI, there is the hardware cost represented by the laptops/tablets used by interviewers. Anyway these costs can be reduced if these techniques are used for periodic (not ad hoc) surveys.

The presence of interviewers makes it easier, with respect to self-administered techniques, to find the right respondent(s) for the questionnaire compilation. This is because interviewers can use appointments or can talk directly to people inside the enterprise that will address him/her to the correct target unit. Anyway, it is important, for both techniques, to plan a contact strategy to send to all the sampled units a pre-notice letter that will advise them about a future phone call or a visit from authorised personnel, thus creating a climate of trust. How to identify who the letter should be addressed to inside the company is fully described in the module "Data Collection – Design of Data Collection Part 2: Contact Strategies". Here it is important to say that the letter has to specify the content and aim of the survey, the deadline and, if possible, which is(are) the designated role(s) to answer the questionnaire. Sometimes a paper questionnaire is enclosed with the letter, like it happens for mail surveys and sometimes for web surveys, in order to let respondents know in advance which information is required. This is especially useful in case the questionnaire contains questions with very technical concepts only known by experts or when it is necessary to retrieve the information required, that can be contained in documents belonging to different business departments.

2.1.1 CATI – Computer Assisted Telephone Interviews

Data collection with CATI requires a central location were a call centre is settled. Interviewers work with desktop computers equipped with microphoned earphones. Each pc is a client connected to a central server that delivers the units to be interviewed according to the parameters that have been set in the scheduling system.

Being an interviewer-administered technique, a good management of the training phase and a constant monitoring of the interviewers' work during the entire data collection period are quite important for a successful data collection.

- The training phase is very important for data quality, because the communication with respondents is via telephone and it is therefore necessary that interviewers are able to create a climate of trust. In general, in CATI surveys, interviewers' training has to focus on:
    - how to assure respondents about the confidentiality of the information they provide;
    - the importance of finding the right moment to administer the questionnaire, by fixing appointments for days and times suitable for respondents;
    - how to probe for questions that are not immediately clear to respondents without influencing the answer in any way;
    - how to solve potential inconsistencies between answers through the use of error windows that contain messages whose text has to be customised according to the type of error.
- **Monitoring** the interviewers' job plays also a fundamental role on data quality and, in CATI surveys, this activity can be easily carried out because interviews are conducted in a centralised facility that allows for a daily storage of data on a central server always at disposal of the NSI. In this way a set of monitoring indicators can be implemented to indicate, day by day, if interviewers work respecting the instructions provided during the training session or, if they don't, which are the correct actions to be taken to improve their work. The set of indicators is designed by methodologists according to the survey's needs. Anyway it is advisable to use at least the following indicators:
    - number of completed interviews total and per interviewer;
    - number of other definitive contacts results (refusal and definitive interruption) total and per interviewer;
    - number of not definitive contact results (no-answer, busy, answering machine);
    - distribution of appointments per day and hours, total and per interviewer;
    - number of out of target units and reasons why;

- number of units with no contact.

  This common set of indicators generally corresponds to the default one provided by the software itself and, therefore, methodologists have the opportunity to concentrate on the design and implementation of ad-hoc indicators about fundamental or very important survey variables that have to be strictly monitored. An example of an ad-hoc indicator can be the measurement of the time spent in assisted coding of open-ended variables, since in case it is too long methodologists can decide to train interviewers again or not to use this function as it can be too burdening for both respondents and interviewers, with negative effect on response rate. An ad-hoc indicator can be based, for example, on Control Charts that show how the values of the monitored variables varies along the time axis, indicating if variations depend on casualty or on systematic errors due to wrong interviewers' behaviour (Murgia et al. 2005).

Being a CAI (Computer Assisted Interviewing) technique, CATI can exploit the software potentialities to manage different aspects of the questionnaire, like different paths or branching or different types of controls (coherence controls, range controls and skipping rules) or texts' customisation for questions' wording and error messages (Blanke et al., 2006).

The questionnaire layout (see also "Questionnaire Design – Electronic Questionnaire Design") has to be designed in order to avoid the segmentation effect (one question per screen) and a too dense video screen to make the interviewer able to sooner find the information he needs, like explanatory texts, helps on line etc. Different colours and fonts should be used according to the different roles played by texts, for example, red for interviewer instructions, black for texts to be read to respondents, etc.

An important and peculiar feature of CATI is the call scheduler that allows methodologists to plan the contact strategy. Generally speaking, this means that methodologists have to establish in advance, on the basis of previous surveys or of pilot surveys, the time period of the day most suitable to run the interviews and how many times a unit (business) with no definitive contact results (no-answer, busy, appointments) has to be tried before assigning it a definitive contact result (interviewed, refusal, definitive interruption, not reachable) and then to substitute it with another one. Besides, thanks to the management of appointments, the scheduling system allows respondents to plan the interviewing time according to their needs thus improving the response rate. The scheduling system is described in detail in the module "Data Collection – CATI Allocation".

It has to be said that, among the CAI techniques, CATI allows for the implementation of the most complex electronic questionnaires because of the presence of well-trained interviewers and because interviewers work in call centres together with field supervisors that can provide immediate help in case of any doubt on how to proceed with the interview (while CAPI interviewers work alone) (Capparucci et al., 2009). Therefore, it is possible to make a "heavy" use of editing during data collection that can also manage many blocking edits (edits to be solved to proceed with the interview) more easily than other techniques. Anyway, the number and type of checking rules to be implemented in the electronic questionnaire have to be established by methodologist maintaining a good balance between data quality and response burden.

This unique feature of CATI must always be kept in mind when choosing the most suitable technique for a survey, but, anyway, it is not always usable in business surveys because, in general, complexity is synonymous of lengthy questionnaires that cannot be administered by telephone because long interviews surely decrease the response rate.

For this reason, the CATI technique is especially suitable only for specific types of business surveys or specific situations: *i)* for short interviews, *ii)* when data collection needs to be run in a very short period of time, like it happens for some agricultural surveys that have to produce timely estimates and preferably when *iii)* the compilation of questionnaires does not need the retrieval of the information from the respondent and *iv)* there is no need of different respondents to administer different sections of the questionnaire,.

To finalise data collection is quite simple in CATI surveys, because, at the end of the phase, data are sooner ready for the editing and imputation phase, that can be simplified because part of the data have already being checked during the data collection.

Nowadays CATI is mostly used for follow-up calls to non-respondent units (Parent et al., 1999) or to probe for answers that failed the edit procedure, or in mixed mode with other techniques like CAPI or CAWI (Computer Assisted Web Interviewing) Its use for business surveys is decreasing leaving room to Web surveys because, as explained later, web surveys are less expensive and are becoming more and more suitable to manage business surveys due to the increasing use of internet combined with the availability of administrative data and, in general, of other secondary source of information.

2.1.2 CAPI – Computer Assisted Personal Interviews

Data collection with CAPI requires a laptop or a tablet for each interviewer: data are stored in each pc and then periodically sent via LAN to the NSI’s central server. As described in the following pages, CAPI requires a very good organisation to coordinate interviewers and to assist them both in terms of survey content and of software/hardware instruments. It is therefore a quite expensive data collection technique, that is generally used to administer interviews to the top management of large enterprises, that are difficult to reach by phone, or in mixed mode surveys, with CATI or CAWI, to cover those strata that have a response rate lower than the average one.

Like the other computer assisted techniques, CAPI presents the advantage of allowing the management of complex questionnaires (in terms of skipping and checking rules) and, in addition, the administration of long interviews. This is because interviews are run face-to-face and, in general, after having taken an appointment with respondents.

Being a CAI technique, the electronic questionnaire constitutes the core of the CAPI system. Anyway to run CAPI several other functions must be implemented (Budano, 2008). These are:

a) **interviewers database**: it is a centralised database necessary to organise at best the interviewers’ job and to reduce their work burden. This database, therefore, has to be used to keep under control which are the active interviewers and which are not active (for holidays, illness, etc.) in order to organise their eventual substitutions with other interviewers;

b) **management of the laptops**: this aspect requires the organisation of local assistance to be provided on the territory in case of any software or hardware problems arise (i.e., possible pc substitution). Besides, a uniform configuration must be given to all the PCs that have to use the same strong authentication procedure - to inhibit the use by others – (Parent et al., 1999) and the same encryption program to guarantee a secure exchange of data. Finally, the organisation has to consider the management of the software package installed on the pc to keep it updated with respect to new software releases or to the software application in case of questionnaire changes. In general, a database containing all the events concerning the PCs is advisable;

c) **interviewers training**: it is a very important aspect for the success of the data collection because, differently from CATI, interviewers work alone and cannot rely on supervisors neither for software nor for content problems. The training phase must treat many different aspects (similar to those mentioned for CATI) like presentation of the interview and questionnaire content, electronic questionnaire management, how to face critical problems during the interview, technical aspects concerning the laptops management, how to manage the different functions of CAPI applications. A local contact reference with functions of supervisor is advised;

d) **allocation of sample units to interviewers:** the distribution of sample units among interviewers must be done according to the information contained in the interviewers database;

e) **contacts management:** respondents can be contacted by telephone to get an appointment or through visits at their location. In both cases, in order to guarantee all sample units the same chances to be contacted for an interview, it is necessary to plan a protocol of contacts that defines which are the possible sequences and amount of contact attempts to be done and which actions have to be taken before assigning a definitive non-contact result to the unit. Let's suppose contacts are made by telephone: in case, for example, of a sequence of "nobody answers" the telephone, then the action to be taken is "a visit at respondent's place of work"; or in case the unit has moved, then the action is "find the new address". At the end of the contacts sequence, the eligibility of the unit can be verified and it can be asked for an appointment to make the interview or it can be abandoned because *i)* it was not possible to contact it, *ii)* it was not possible to find the new address or *iii)* the unit refused the collaboration. All these contact results must be stored in the sample unit database and managed in a **contacts report** that can be electronically sent to the centre to monitor the interviewing phase;

f) **interviewers' agenda:** it is an important application in CAPI surveys to manage appointments and in general contacts with respondents. Therefore, it must be related to the contacts report chart because interviewers report in it all the events relative to the contacts with respondents like: changes of addresses, appointments to start or continue the interview, definitive contact results. The agenda is also related to the electronic questionnaire (in the same fashion as the scheduler for CATI surveys), because during the interview (when filling the electronic questionnaire) it is possible to register some contact results, like completed interview, appointment to complete the interview on another day, refusal to complete the interview. All these results will update the contacts report and the sample unit database;

g) **interview:** during the interview, interviewers put in practice what they have learned during the training phase. They have to read the question wording as it is, to give explanations when the respondent asks for them, to read carefully the error messages to try to solve consistency errors together with the respondent, to manage critic situations which could compromise the completion of the interview, to take notes of any problems/difficulties encountered;

For all these reasons, the electronic questionnaire design is fundamental (see also "Questionnaire Design – Electronic Questionnaire Design"). It must be made so as to reduce the so-called segmentation effect (Blanke et al., 2006), to make clear to the interviewer which parts are to be read to respondents and which not and where he/she can find help functions concerning both technical problems or variables definitions. A good electronic questionnaire design can be useful to reduce the interviewer effect. It must be easy to assign the contact results from the electronic questionnaire, so as to update automatically the contacts report, the interviewer's agenda and the sample unit database.

h) **exchange of data from/to the centre:** it necessary to manage the exchange of information from and to the centre (NSI): the NSI sends interviewers data about units they have to contact and receives from them data concerning completed interviews and contacts results (possibly after each working day). In this way it is possible to monitor the interviewing phase and to take in due time the right actions in case of problems. All electronic data exchange must be done guaranteeing security in terms of data integrity and privacy requirements. These features are obtained through the use of secure protocols and data encryption. The electronic data exchange function must be easily called by the interviewers, possibly by the agenda;

i) **monitoring of interviewing phase:** a monitoring system must be defined, to be daily analysed by the survey manager. It should be updated with data sent by interviewers and should process these data automatically to produce synthetic indicators to take under control different aspects of data collection. In particular, it should monitor: the state of the art of the interviewing phase (units to be contacted, not eligible, to be substituted, etc), the respect of contacts protocol by the interviewers, the interviewers productivity and any odd behaviours of interviewers, like, for instance, interviews too long or too short, too high units substitution rates, etc.;

j) **finalised data collection:** at the end of data collection, data are ready for processing. Like in CATI, the editing and imputation phase can be simplified as part of the data had already being checked during the data collection. Anyway, as interviewers work alone, it is advisable not to implement a "heavy" editing during data collection, leaving the correction of inconsistencies to the revision phase.

2.1.3 Direct observation

Direct observation is another way to perform data collection. With respect to the other modes, data are directly "observed" by the interviewers with no need of asking the information to respondents and therefore no response burden exists. It can be conducted with or without the support of computer and therefore all the elements relative to the support of the computer can be found in the previous sections.

The organisation of the data collection is similar to that of CAPI, since interviewers are spread over the territory, but the role played by the interviewer is even more delicate as he/she is the only observer of the phenomenon. Therefore, skilled interviewers on statistical methodologies and IT tools have to be used for this kind of survey and consequently a very deep and detailed training phase has to be managed by the NSI. For these reasons, Direct Observation is a very expensive technique. It is generally used for surveys on pricing and for some agricultural surveys aimed at estimating types and areas of crops (Statistics Canada, 2010).

An example of a survey based on Direct Observation is the "Survey on Prices of Consumption" carried out by Istat- Italian National Institute of Statistics. The collection of prices is performed in two different modes:

- a territorial collection for the most part of goods and services conducted by local offices;
- a centralised data collection performed by the central office about goods and services which have uniform prices at a national level.

The territorial collection is run through the direct observation of prices: interviewers use tablets where an ad-hoc software, developed in Istat, is installed. Data transmission to the central server is in real time through the use of the 3rd generation mobile technology.

The centralised data collection extracts information on databases that are available on the web or from specialised websites (for example, prices of train or air tickets).

2.2 Self-administered techniques

This section describes CAWI and mail surveys that, being both self-administered techniques, share several aspects to be taken into account when choosing the data collection mode. These common features are discussed right below, while peculiarities of each technique are treated in the two dedicated sub-sections.

When these techniques are used, respondents have, in general, more time to provide their answers and therefore these two modes can be adopted in case of long interviews or when respondents need to retrieve the information to answer the questionnaire or in case more respondents are needed to answer the same questionnaire.

At the same time respondents have to be guided and helped in the questionnaire compilation, as they cannot rely on the help of any trained interviewers (Couper, 2001) and, of course, they are not trained on how to answer. To this aim the questionnaire and its layout together with the instructions for questionnaire compilation are of an extreme importance:

- the questionnaire and its layout have to be designed with criteria different in the two techniques, but following two common rules: 1) they have to provide respondents with all the information they need without being chaotic or confusing, 2) they have to arouse and keep always high the respondents’ interest (Istat, 1989).
- instructions have to cover two main information areas, one on content and one on technical aspects:
    - in terms of content, instructions have to make clear and understandable the meaning and the aim of each question. Besides, if necessary, they have to clarify which professional figure(s) has to answer the various questionnaire sections;
    - about the technical aspects, instructions have to inform on how to fill each question and on how to navigate among them. In other words, they have to make respondents immediately understand how and which questions have to be answered.

An important aid for respondents is represented by an “information point” they can easily contact to get any information they need. This is represented, in general, by a toll free line or an e-mail address managed by field staff that has been trained on how to answer all possible requests that can be about the content of the surveys, the technical aspects of the questionnaire or about organisational aspects of the survey.

Another issue to be managed for both techniques is the reminder strategy. As described in “Data Collection – Design of Data Collection Part 2: Contact Strategies”, it is important to plan in advance “when” and “how many” reminders should be sent in order to control the unit non-response rate without increasing the response burden with too many or not well addressed reminders. The reminder strategy has to be planned together with the reporting of unit non-response – coding of the reasons why a unit cannot be enumerated - in order to make reminders more effective and to take the correct actions in due time (substituting the unit, searching for the new address, telephone contact to the unit, etc).

Finally, a common element to be managed is the partial non-response typical of self-administered techniques: to keep it under control, it is necessary to properly design the questionnaire and the compilation instructions and to well organise the follow-up phase, that can be conducted by means of telephone interviews aimed at obtaining answers for those questions with missing values.

All these elements are common to CAWI and Mail surveys but they are implemented in different ways since the way these surveys are run is different.

2.2.1 CAWI – Computer Assisted Web Interviews

The use of web surveys is increasing in general and in particular for business surveys because among enterprises the use of software and hardware equipment is higher than for households/individuals. In fact, web surveys require, at respondent side, the presence of a computer equipped with internet services and obviously that respondents are acquainted with them. On the NSI side, they require a secure web server accessible from internet where to create web pages containing the questionnaire and the entire data collection IT infrastructure. Another important factor for its increasing use, is that the WWW offers the "lowest cost survey environment" especially for ongoing surveys: minor cost of data transmission, no postal charges and less cost for telephone fees (Clayton et al., 2000).

Web surveys are based on Computerised Self-Administered Questionnaires (CSAQ) that can be answered on-line or off-line:

- in the first case, respondents log on to a secure website and enter their data, or can upload dataset of survey data as explained in sections on EDI and XBRL;
- in the off-line case, respondents download, on their pc, an executable file or a "flat" file (excel, pdf, csv, etc.) containing the questionnaire or the structure (record layout) of microdata. Data are then sent back to the NSI via a secure e-mail system or by a fax-server and are automatically stored on the survey database.

On-line compilation assures timely data and a greater control on data collection from the survey manager. This implies a higher data quality but at the same time a higher response burden. Off-line compilation facilitate respondents' co-operation especially in case more respondents are needed to answer the questionnaire, but there is a lower control on how questionnaires are filled in. To enhance response rate, both alternatives can be available. Besides, the paper questionnaire should always be downloadable: in this way respondents can print it and use the paper format as an aid to answer on the web.

To obtain a good response rate, it is necessary to consider many factors for the management and organisation of web surveys. The main ones are described below:

- **Cover letter**: a cover or pre-noticed letter about the starting date of the survey has to be sent to the sample units that were selected during the sample design phase. The letter should contain information about the survey, like its content, its aim and its deadline, together with the web address of the survey, the id-name and (temporary) password that respondents will use to register on the survey web site (see also "Data Collection – Design of Data Collection Part 2: Contact Strategies"). The letter can be sent also to the e-mail addresses of respondents, if an updated list of them exists. For short-term statistics the cover letter can be sent repeatedly at each survey round by e-mail or fax. The question about "to whom" address the letter (to a specific person or to a designated role) is deeply described in the module “Data Collection – Design of Data Collection Part 2: Contact Strategies”. Here it would be sufficient to say that for business web surveys it is more crucial than for the other techniques to know in advance who the respondent unit is. This is because, apart from the need to retrieve information that can be located in different business departments or the presence of technical questions only known by experts, it can happen that respondents need authorisations before submitting the questionnaire. This last thing is especially true for web surveys because the questionnaire compilation requires a person able to use the pc that might not correspond to the target person. Therefore, it would be advisable to let respondents know in advance the content of the interview, by giving them the opportunity to download a paper questionnaire from the survey website, or by sending it by e-mail or as a last chance by post (environmental issues).

- **Login procedure**: for a well organised web survey the login procedure and the correlated issues of data security and privacy have to be managed. It is very important to implement an easy login procedure for respondents who, at the same time, have to feel sure about the respect of confidentiality of the information they send via web. Access procedures must be set-up in accordance with the national privacy laws. One way of managing them is to provide each respondent (using a paper letter or an e-mail) with a user-id and a temporary password for the first access and then allow the change of the password with a personal one that has to respect the common standards on passwords. For all subsequent logins for the same survey only the personal password can be used and, as it is known only by the respondent, it guarantees the respect of data confidentiality.
- **Questionnaire and its layout**: as web surveys are based on self-administered questionnaires, it needs to pay a great attention to the way the questionnaire appears on the video screen of the respondent’s pc. In fact, like the other CAI techniques, the layout has to be designed in order to avoid a screen too dense of information to make the user able to easily find what he needs to answer the questionnaire. In addition, for web surveys, it is important to implement an easy navigation of the questionnaire and to avoid vertical or horizontal scrolling that makes navigation more difficult and therefore burdensome (O’Neil, 2008). Like the other CAI techniques, it has to contain automatic skipping rules and customisation of questions’ wording. The questionnaire has to be designed in order to be easy to understand and to complete (Couper, 2001), it must keep respondents’ attention always at high level to make them able and willing to provide the optimal answers and has to make respondents sure about the confidentialities of their answers. Online help is extremely important: it should appear under an icon easily recognisable by respondents and has to contain information on words, concepts, questions’ aim and also instructions on how to fill in the questionnaire.

  The pc support allows for the implementation of any type of questions including the open-ended ones for which an assisted coding function, easy to use, can be provided. Anyway, to avoid response burden, the use of open ended questions should be limited to variables easy to code (i.e., the place the establishment is located) or to variable respondents are used to answer (i.e., NACE[1] sector).

---

[1] NACE, is the nomenclature of economic activities in the European Union.

- **Editing during data collection**: this feature provides the same advantages described for CATI and CAPI. Anyway for web surveys, it is more important than for the other modes to establish good balance between edits and quality of data: since there are no trained interviewers for the questionnaire compilation (Couper, 2001), a too high number of error messages during the questionnaire administration can increase response burden, lowering the quality of the answers, and can induce respondents to skip questions or to stop their cooperation before the very end of the questionnaire. In general, for web surveys it is recommended to use edits during data collection only for crucial or important survey variables (that are defined during the questionnaire design phase). Furthermore, for these variables, edits can be blocking edits, meaning that the compilation cannot proceed until the error is solved. For the other not fundamental variables, it is instead advisable to implement warnings, also called soft edits, that have the advantage of making respondents aware of possible inconsistencies in the information they have provided and to leave them the choice of solving the edit failures or not: this "freedom" reduces respondent burden.

  A peculiarity of web surveys is the possibility of managing edits on server-side and/or on client-side:

  - usually only the edits on server-side are chosen: this means that each time the respondent presses the "submit button" (that could be placed at the end of each page/section or at the end of the questionnaire) data are stored on the database located on the central server. The database contains also all the checking rules and, in case of inconsistencies among the submitted data, errors messages appear to the respondent that is asked to solve the errors. This way of managing edits is the right one for short and not complex web questionnaires, like those generally used in business surveys. It is not suitable for long and complex questionnaires since it would imply too many edit messages at the end of the questionnaire or a too high LAN traffic in case data submission is done at the end of each page (Capparucci et al., 2009);
  - implementation of edits on client-side is strongly suggested for long and complex questionnaires. This solution has also the advantage of solving edits as soon as they happen and therefore to store only consistent data on the central server. The drawback is that edits on client-side need that software able to manage them (in general *Javascript*) is active on respondent' pc. If this is not the case[2], the presence on checks on server-side will solve the problem, meaning that checks on server side must always be implemented in a web application[3].

- **Hardware platforms, software systems and browsers**: a typical feature of web surveys is that respondents use PCs with different hardware components, different platforms and different software systems (Couper, 2001). The consequence is that the questionnaire might not work

---

[2] Respondents might be provided with a link to download and then install the needed software, but, in general, it is not a good practice to ask respondents to do this because it is burdening and because they are not so prone in installing new software on their pc.

[3] The use of Ajax for asynchronous communication between client and server can be seen as a compromise between the two alternatives. Ajax is an acronym for Asynchronous JavaScript and XML. It is a group of interrelated web development techniques used on the client-side to create asynchronous web applications.

properly with some of them or it can be visualised in different ways. A typical example is the use of different browsers that might visualise a simple single-choice question in a completely different way. To avoid all of this a further effort is required to the NSI when designing and implementing the electronic questionnaire because it can be necessary to use ad hoc source code for the development of the electronic questionnaire. This extra work is quite important to reduce respondent burden and to control the non-response rate, because no extra efforts than questionnaire compilation have to be asked to respondents: the web application has to function properly without requiring the installation of any extra software components on the respondents' PCs or the use of specific web browsers.

As mentioned for the other CAI techniques, the presence of hardware and software technology can represent a drawback in a financial sense of the word (O'Neil, 2008), because it is costly to take all the above mentioned actions aimed at reducing the respondents' fatigue. Anyway, if the application is used for periodic surveys, the NSI can easily write off the initial cost.

- **Response time of the web application**: another aspect to be taken into account for the management of web surveys is the response time that represents here the period of time that a respondent (pc client) has to wait to get answers to his queries to the server. For example, the time to be waited after the submission of an answer and the administration of the following question. The entire web application must be implemented in such a way that response time is reduced at minimum levels, in order to avoid respondents abandoning the interview before completing it. Crash tests to establish how many contemporary accesses to the web site are possible with no delay in response time are therefore highly recommended before the beginning of the survey.
- **Partial submission**: it is important to give respondents the possibility of partial questionnaire submissions to let them answer the questionnaire in different moments of the day when they have time. This is particularly true for those surveys that need an information retrieval and/or different respondents to answer the various questionnaire sections.
- **Monitoring system**: in order to keep the data collection phase under control a set of real time indicators has to be implemented. Statisticians can build their own monitoring system that, anyway, should at least report the following indicators:
    - the number of completed interviews;
    - the number of partially completed interviews;
    - the number of refusals;
    - the number of those units that have made the registration but have not answered any questions;
    - the number of those units that did not register themselves.
- **Reminders**: as already said, the reminder strategy is fundamental for self-administered interviews to control the unit non response rate. In web surveys, reminders are generally done according to the results of the monitoring system and through different means of communication like the e-mail address, fax, or telephone.
- **Finalise data collection**: at the end of data collection data are ready for processing and for the editing and imputation phase that cannot rely on already checked data (if compared to CATI and CAPI), due to a limited use of editing during data collection. Anyway, the consistency of final

data can be more easily reached if the questionnaire has been designed following a metadata-driven approach or (Iverson, 2009) any techniques for relational database design, like the "Entity-Relationship scheme (E/R)" (Chen, 1976). In these ways data collected by filling the questionnaire are immediately stored into the relational database, underneath the survey, according to its designed structure that contains data tables, data links and data constraints.

The web site that hosts the survey plays a fundamental role for the success of the survey in terms of response rate, since it represents the "contact point" between businesses and the NSI. The web site should host not only the questionnaire but all the other data collection instruments aimed at supporting respondents in the self-administration. These are:

- instructions on how to access the web;
- instructions on how to answer the questionnaire,
- information on survey contents and aims,
- contacts for any questions or problems,
- list of FAQs,
- contact information of each enterprise (address, telephone number, e-mail address, etc.), that respondents can update after their registration.

All these elements have to be easily accessible from the web site that should be compliant with the following general requirements (Balestrino et al., 2006)[4]:

- to present a homogeneous and stable image of the Institute on the outside;
- to guarantee sender and receiver of each other identity;
- to guarantee the confidential nature of data and the comprehensive environmental security during the collection process;
- to minimise the impact on the operational environment of the external user;
- to replay to the user about the operation he carried out with a confirmation message;
- to favour the monitoring activity about data collection;
- to favour the internal management of the operations related to the data collection;
- to contain costs.

The future of web business surveys is represented by the "Business Statistical Portal", a new way of organising and managing the data collection process for business surveys and already active in some European countries. This model allows to abandon the usual stovepipe model used for the production of business statistics which is "survey centred", adopting a model which is "enterprises centred" and based on integrated production processes that will make NSIs able to organise more efficiently data collection, data processing and data estimation processes.

---

[4] In Istat – Italian National Institute of Statistics – a web site dedicated to web surveys, named Indata (https://indata.istat.it), has been implemented since the '90s with the aim of presenting a unique front-end for respondents to any surveys.

The Business Statistical Portal will strongly reduce the response burden and therefore cost. This is thanks to the integration of administrative data and data provided by enterprises through the use of simple procedures that will allow asking only once information common to all surveys the enterprise is involved in.

A Business Statistical Portal should be compliant with the following requirements:

- it should allow for the sharing of data and metadata on the basis of a common data modelling;
- it should provide a centralised governance for the data collection processes respecting the businesses' needs;
- it should manage a back-office activity that allows to monitor the production of business statistics;
- it should allow the re-use of data that are already available in the statistical system or among the various public administrations, promoting also new protocols for data exchange;
- it should use IT instruments that make simpler and less expensive the exchange of information.

2.2.2 Mail surveys

Data collection for business surveys can be run by paper questionnaires which are sent to the target units by post and sent back to the NSI still by post or by fax. Due to low response rate and the environmental impact caused by the use of a great amount of paper, the adoption of this technique for business surveys is decreasing in favour of the use or the combined use (mixed mode) of those assisted by computer especially the web based ones. Anyway this mode has still a great importance to collect information for various reasons (explained in its advantages listed below) among which the low cost plays the major role.

It has advantages and disadvantages (see also "Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method") that are typical of self-administered techniques based on paper questionnaires and that influences the way data collection is set-up:

Advantages:

- as already said, it requires a low budget effort;
- it is quite useful when respondents need time to answer because: *i)* the interview is long or *ii)* it is necessary to retrieve information before answering the questionnaire or *iii)* the questionnaire contains questions with very technical concepts only known by experts or *iv)* because the questionnaire has to be filled in by different persons inside the same enterprise;
- due to its low cost, samples can be larger than those used with other techniques (keeping budget constant);
- the questionnaire can contain difficult questions (calculation, ordering, etc.);

Disadvantages:

- it has low response rate that requires to plan a reminder strategy (as described in the following) and therefore a longer data collection period;
- it has a high risk of partial non-response or of incomplete questionnaires that requires the design and setting of a more accurate editing and imputation phase;

- data are not soon available since there is a need to plan the data-entry phase to finalise the data collection (as described in the following).

In setting-up a data collection with mail surveys different elements should be taken into account as described in the following list.

- **Sending material to sampled units**: sampled units have to receive by post all the material necessary to participate to the survey (see also "Data Collection – Design of Data Collection Part 2: Contact Strategies"). Monthly or quarterly deliveries of all the material are generally planned for short terms statics. In these cases e-mail or fax are used instead of a mail-out system.

  In general a unique envelope is sent, containing the following material (Istat, 1989):

  - a cover letter that, apart from describing the content and aim of the survey (similarly to any other techniques) has to explain to respondents *i)* what they are asked to do, *ii)* the importance of their cooperation, *iii)* what they can do in case of any doubts, *iv)* which telephone number or e-mail address they can contact, *v)* how confidentiality is guaranteed and *vi)* acknowledgments for their collaboration;
  - instructions for questionnaire compilation, that explain the content of each question, the meaning of concepts, how to fill in questions and how to read navigation instructions. To avoid a too long questionnaire, instructions are in general written on separate paper sheets. It is advisable to re-write instructions (in a shorter format) on the questionnaire itself, next to the question they refer to. In fact, it has been tested (Istat, 1989) that respondents tend not to read instructions if they are written in documents different from questionnaire;
  - the questionnaire that, if possible, should be customised with pre-printed information like for instance enterprise master data (name, address, ect.);
  - a pre-paid returned envelope to send the questionnaire back to the NSI with no additional cost for respondents.

- **Questionnaire**: the questionnaire design and its layout (see also "Questionnaire Design – Main Module"), in other words, the way questions and instructions are organised and graphically represented, are extremely important for mail surveys even more than for web surveys, because in mail surveys the questionnaire is static and not dynamic (Couper, 2001) and therefore it is not possible to create different versions that are customised according to the interview flows. The first page of the questionnaire should contain a short presentation of the survey and must indicate a code (a bar-code or an alphanumeric sequence of characters) that represents the univocal key assigned to each respondent. This key is repeated in all pages to help finding and joining separated pages of the questionnaire and it is fundamental to link survey data to each respondent during the data registration phase. Besides in case an enterprise has more local units involved in the survey, it has to be created in such a way to link all the questionnaires.

  Different fonts and colours should be used for texts according to their functions, in order to make respondents immediately understand if they are reading a question or an instruction or the section header. For questionnaire background a light colour should be chosen (Fanning, 2005) in order to create a contrast with texts that can be better read. Obviously it is important not to use too many colours that might increase respondents' fatigue.

Questions should be organised according to a logic flow and, in general, this means to group into a section those questions referring to the same theme (see "Questionnaire Design – Main Module"). If possible a questionnaire page should correspond to a questionnaire section and, in any case, the beginning and ending of each section has to be made clear by using lines, boxes or other graphical elements.

Response items must be placed on the same page of the question they refer to, because if written on the next page there is the risk that respondent may miss reading them, creating a potentiality for measurement errors on data. Besides, to allow data-entry with specific software, they have to be numerated and a box for checking the answer has to be placed next to them.

Skipping instructions have to be graphically represented through symbols or short instructions (i.e., >>, →, goto, etc.) placed close to the filter questions in order to facilitate respondents in filling the correct branches of the questionnaire (see "Questionnaire Design – Main Module").

Any types of question format can be used, although open ended questions should be used rarely. There are three main reasons for this recommendation: *i)* hand-written material is difficult to be registered both from the data-entry operator and the OCR (Optical Character Recognition) software; *ii)* the content of the answer could be meaningful or generic or ambiguous since there is no interviewer probing to get a meaningful answer and *iii)* a coding phase is necessary at the end of data collection that becomes more costly in terms of time and resources.

- **Reminder strategy**: reminders are necessary to increase the response rate and should start when questionnaires arrivals at NSI start decreasing regularly. This does not apply to short term statistics for which the first reminder is generally sent before the end of the data collection period (see "Data Collection – Design of Data Collection Part 2: Contact Strategies"). Reminders can be done by telephone or by post. As it may happen that some questionnaires are missed or do not reach respondents, it would be advisable to send, during the first or the second reminder, another questionnaire (Istat, 1989) paying attention, at the end of data collection, to the presence of duplicated questionnaires. The structure of a reminder (by letter or telephone) should be the following:

  - a kind but determined invitation to answer the questionnaire;

  - to re-state the importance of respondent's cooperation;

  - apologies for those who already answered or did not receive the questionnaire.

- **Organising the data-entry phase**: the information collected through paper questionnaires has to be gathered and stored in an electronic format. This is done through data-entry with specific software or OCR systems that are described in the following two sub-sections.

- **Finalising data collection**: at the end of the data entry stage, the NSI has at its disposal a set of raw data as supplied by respondents, that is used as input for the editing and imputation phase where all types of inconsistencies are treated an solved. Besides, the raw data set enables statisticians to carry out systematic error analysis, which might be interesting for testing the clearness of questionnaires. Furthermore, by saving the original data the value added of editing operations can be determined. Thirdly, during subsequent stages of the processing, discussion might arise as to the correctness of certain edits. This holds in particular when consistency checks with data from other surveys reveal differences between edited data (Willeboordse, 1998)**.**

2.2.2.1 Data-entry with specific software

Data-entry can be done by means of an electronic questionnaire which is developed using specific software (like Blaise or CSPro)[5]. In case of mixed mode, the same program used for the other(s) CAI technique(s) can be used for data entry, thus saving implementation time and costs. The electronic questionnaire has to be developed in such a way to make the typists' job easy: this means that the electronic questionnaire layout should be quite similar to the paper one and no blocking edits have to be implemented. This is because typists are not trained on how to solve inconsistencies and the only thing they can do is to compare the entered data with the data on the form. In general, only soft consistency edits are implemented to reduce typing errors while blocking edits are about the questionnaire key number to avoid duplication of the same questionnaire. Anyway if the amount of questionnaires is limited the revision phase can coincide with the data-entry one. This requires the organisation of a training session for typist about how to solve inconsistencies.

2.2.2.2 Data-entry with Optical Character Recognition (OCR)

Another way to electronically store paper questionnaires is OCR that is particularly suited to large data collections. It allows simple edit checks, like valid values and value ranges. Readability is the crucial factor and shortcoming of this method: statisticians should take into account that the method does not enable systematic controls on the readability of the data reported, that numbers are more easily readable than plain text and that hand-written material is more difficult to be recognised than typed data. Although modern OCR packages use dictionaries and quite sophisticated software for texts recognition, the main caution to be considered is that OCR requires very accurate questionnaire layout and printing standards to ensure that the answers can be read by the sensors correctly.

At the end of the data collection by OCR, raw data are submitted to a program that checks which records have had problems in recognition of texts. Those records that fail this check are then submitted to the video screen correction and correct texts are inserted manually on the bases on what reported in the paper questionnaires. After this phase, the set of raw data is ready for the editing and imputation phase.

2.3 Data collection using EDI and XBRL

2.3.1 EDI: Electronic Data Interchange

EDI represents the "*Electronic exchange of data usually in forms that are compatible so that software or a combination of individuals and software can put the data in a compatible form at the receiving end if necessary*". (SDMX, 2009).

---

[5] Blaise is a computer-assisted interviewing (CAI) system and survey processing tool developed by Statistics Netherlands. CSPro - Census and Survey Processing System - is a public domain software package for entering, editing, tabulating, and disseminating census and survey data.

EDI offers businesses the opportunity to retrieve information electronically from their internal systems and to forward that information to trade partners/suppliers/customers/government through a communications network (from Context of SDMX, 2009).

The use of EDI requires standardisations from both technological and conceptual sides (Willeboordse, 1998):

- since businesses develop their own information system on the basis on their needs, it is necessary to map the concepts and then to standardise them accordingly to the statistical use. Conceptual dissimilarities may concern: *i)* the naming and coding of data items, *ii)* the level of aggregation of data items - a statistical item may be composed of different accounting items -, *iii)* the existence of data items - a statistical concept may have no accounting counterparts-. This standardisation has to end up with a standardised set of metadata to be used in any kind of business surveys;
- data should also be organised in a standard technical form in order to be readable by the NSI.

Therefore, implementation of EDI for data collection comprises the design of an electronic *translation* facility (Willeboordse, 1998) in order to bridge the technological and conceptual gap between the worlds of respondents and of NSI. As a consequence, enterprises have to use software for the translation and therefore they will have some start-up costs to adapt their information system. This cost also includes an overlapping testing period where both the old data collection method and the new EDI system are used. Besides these costs depends on the nature and complexity of edification projects that varies among surveys and depends on (Willeboordse, 1998):

- the distance between business accounting systems and information needs, with respect to the technological and conceptual dissimilarities as mentioned above;
- the degree of standardisation of business accounting practices.

The greater the distance and the lowest the degree of standardisation the higher the initial costs that anyway can have a counterpart in the reduction of respondent burden that can also be reduced if the NSI supplies standard software packages for free to the respondents.

This means that the use of EDI as a mean of data collection has an impact on the entire statistical process. It reduces respondent burden because it avoids the compilation of questionnaires and because it requires the harmonisation of similar of equal questions asked in different surveys. It improves the timeliness of data since it reduces the time that elapses between data collection and data processing, it can improve statistical integration since same data can be used for different statistical figures (Hans R. Stol).

Examples of the use of EDI are UN/EDIFACT and GESMES.

UN/EDIFACT - United Nations / Electronic Data Interchange For Administration, Commerce and Transport (http://www.unece.org/trade/untdid/welcome.html) is the international EDI standard developed under the United Nations. It comprises a set of internationally agreed standards, directories, and guidelines for the electronic interchange of structured data, between independent computerised information systems. In particular the EDIFACT standard provides:

- a set of syntax rules to structure data;
- an interactive exchange protocol (I-EDI);

- standard messages which allow multi-country and multi-industry exchange.

Recommended within the framework of the United Nations, the rules are approved and published by UNECE in the UNTDID (United Nations Trade Data Interchange Directory) and are maintained under agreed procedures. EDIFACT has been adopted by the International Organization for Standardization (ISO) as the ISO standard ISO 9735.

GESMES - Generic Statistical Message
(http://www.sdmx.org/docs/1_0/SDMX%201_0%20SECTION_04_SDMX-EDI.pdf)

It was developed by a group of European statistical organisations working within the international UN/EDIFACT standards body. GESMES was accepted as UN/EDIFACT Status 1 messages in 1995 and was first published in the UN/D95A directory. The statistical office of the European Union, EUROSTAT, who has lead the development of statistical UN/EDIFACT messages, is implementing GESMES into the data flows between it and the Member States of the EEA (European Economic Area) and promoting the use of the messages by other international organisations and by other sectors.

GESMES has all the features required to exchange multi-dimensional arrays and time series data, including metadata (such as attributes and footnotes). The advantage of using GESMES, in preference to a proprietary data format, is that it is an internationally agreed standard which is both open and fully functional. It is not tied to the format and constraints of one particular application. In particular GESMES supports the exchange of: metadata, multi-dimensional arrays, time series, administrative data.

An application of GESMES is GESMES/TS - GEneric Statistical MESsage for Time Series – (http://stats.oecd.org/glossary/detail.asp?ID=5874) which is a data model and message format (a GESMES profile) allowing the exchange of statistical time series, related attributes and structural definitions using a standardised format. The initial name of GESMES/TS was GESMES/CB (GEneric Statistical MESsage for Central Banks), but has been changed in order to reflect its wider application. The model and format are maintained under the auspices of the SDMX initiative. In this context, GESMES/TS is known as SDMX-EDI. In the same context it must be mentioned SDMX-ML which is the XML syntax used by the European Central Bank and the national central banks in the web dissemination of statistics.

At present, the use of the web and all instruments correlated to it and based on the XML standard have allowed the implementation of the XBRL described in the following section.

2.3.2 XBRL: eXtensible Business Reporting Language

While different EDI solutions may be very efficient in some cases, as shown in the previous section, there is also a strive for more generalised technical structures and formats that may aid statistical offices and other collectors of business information connect with businesses in an even more efficient way. This could involve sharing of data between authorities or the possibility for businesses to re-use the data in their administrative systems for many purposes. XBRL, short for eXtensible Business Reporting Language, may very well become this standard format. The XBRL format, developed and maintained by a consortium of regulators, accountants and software builders, can offer a link between the data kept in book keeping systems and the data terms of regulators, such as national statistical and tax offices. XBRL offers the same advantages as other EDI solutions; a possibility of reduced respondent burden and data collection costs, especially after over time since the first implementation may invoke some costs. The main difference between XBRL and other more specific EDI solutions is that XBRL is an open format that is intended to be used for many different purposes from the business side; exchanging data within the enterprise (or enterprise group), exchanging data with accountants, sending data to government authorities and also sending data to any interested party such as banks, analysts et cetera.

What XBRL is has been described by Roos (Roos, 2008). XBRL is an XML-based computer language specifically developed for the exchange of business facts between computer systems. Business facts are defined as administrated events that are of economic interest to the company or other related organisations. The XBRL-standard provides a precise, predictable structure for describing and expressing those business facts in a way that can be used and processed by computer systems.

One advantage of XBRL compared to other file formats is that it is an open standard based on the globally well-known language XML. The idea of XBRL is rather simple. Instead of treating information as a block of text, like on a website or in a written document, XBRL tags the individual information in a document with the necessary information. This makes each piece of information readable and possible to interpret electronically.

In order for the systems to understand each other, an agreement on terms and definitions is needed. This is defined in a taxonomy. An XBRL taxonomy defines variables and the relations that may exist between those variables. A taxonomy may also refer to variables defined in other taxonomies. The taxonomy is developed by for example a data collector to describe which information is required. If this is done, a software provider or businesses themselves can link the data in the administrative systems to this taxonomy, and provide the requested information automatically as soon as the link is set up. Taxonomies can be created globally or locally, and relate to any kind of concept. On an international level, there are for example taxonomies created based on the accounting standards US GAAP and IFRS (for more information, see www.xbrl.org). It should be noted though that even though XBRL is an open format, the immaterial rights to the XBRL format are owned by XBRL International Inc. Therefore, when developing a taxonomy it is important to follow the specifications and guidelines given by the international consortium, and also to follow how others use the standard to ensure that the taxonomy created is in line with other ongoing initiatives. According to Bohlin et al. (2009), there are three fundamental design principles for the creation of taxonomies:

- Immaterial rights: The development and maintenance of taxonomies must follow the Intellectual Property Policy of XBRL International. They can be found at www.xbrl.org
- Technical guidelines: The taxonomy must follow XBRL 2.1 and should follow the Financial Reporting Taxonomy Architecture (FRTA) of XBRL International as much as possible. The FRTA can also be found at www.xbrl.org
- International "best practice": The taxonomy must strive to follow similar design as other established taxonomies to ensure comparability and interoperability.

There are a number of factors influencing how effective introducing XBRL in the collection of statistical information can be:

- If XBRL is used for other means of sharing information or not. If other government agencies also use XBRL, it is easier to set up something. For example, if XBRL is used to send annual reports or tax reports, there is already an experience of using taxonomies and mapping business information to them. Moreover, if some data can be used to fulfil data provision to several authorities, there might be a better business case to get software providers and others interested.

- How the taxonomy is set up. If the taxonomy is set up according to the requirements above that is a good start, but it must also be usable and understandable to businesses. The terms used must be clear and unambiguous.
- Mapping the taxonomy to the business systems. The possibility to map a taxonomy to business systems may vary a lot depending on the situation in different countries. Some countries have mandatory ore very well-spread standardised accounting systems, meaning that it is possible to create more general mappings (for example, by software providers or even the statistical office itself) that can be used by many enterprises. For example, Statistics Finland has created a taxonomy that relates to hotels (Konttinen, 2012). In other countries, more or less every single enterprise would have to make their own mapping. For enterprises to be willing to make such a work, there must be potential gains over time, for example the possibility to re-use the mapping many times. For large enterprises, this might be certain since they are almost always included in the statistical samples, while for smaller enterprises they might very well rotate out of the sample after only a short while. Such enterprises might be less willing to make a mapping exercise.
- Mapping the taxonomy to the metadata systems at the Statistical office. Using definitions in a taxonomy in the data collection is an undertaking that must also be upheld. Changes to variables and definitions linked to XBRL taxonomies that are used by several parties cannot be done without informing the others, or indeed the businesses using these for reporting data. The links between the metadata systems and the taxonomy must be upheld and maintained, but it can also help in giving standardised definitions that are generally agreed upon, meaning the statistical office does not need to prepare its own definitions.
- Legislative issues. In a few countries, using XBRL has become mandatory for some reporting, in a few cases also statistical reporting (mostly for financial enterprises at the time of writing). Such a support of course makes using the XBRL solution for statistical collection much easier.
- Building technical solutions. After the mapping, there needs to be a technical solution to transfer the data from businesses to the statistical office. There are three possible cases to cover:
    - Complete coverage (everything requested in the statistical requirements is covered by the mapping and there is no need for adjustments).
    - Adjustments needed (for example, everything is covered but several sources at the business are involved and need to be combined, or some values have to be recalculated, e.g., adjusted from an accounting period to a calendar year).
    - Incomplete coverage (not everything in the statistical requirements is covered).

  In the first case, it would be enough to build some mechanism for transferring the file from the business to the statistical office. It is probable that business would request an encrypted possibility. In the other two cases, there is also a need to build some tool to do the adjustments and add the missing data. This requires a more sophisticated technical solution. It can be built by an outside provider or by the statistical office itself, depending on what is deemed most suitable. Regarding this point, there are large similarities between systems for XBRL data and other EDI solutions.

Considering the limitations outlined above, it is clear that it is still too early to recommend a generalised common solution for implementing XBRL in the statistical collection all across the European Union. It is however clear that XBRL is a standard format which is used more and more extensively for many purposes, and statistical offices need to consider the possibilities to exploit in the statistical collection in the extent possible given the country specific situation. It can be foreseen that the use of XBRL will continue to grow, also in the statistical area. If a country is contemplating implementing an EDI solution (see above), making use of XBRL should also be considered. Using XBRL has a large potential in reducing response burden as well as data collection costs, but there is still a long way to go before it is a widespread possibility for businesses to provide data through XBRL on a broad scale.

2.4 Use of administrative data

Although administrative data do not represent themselves a data collection technique, they have to be mentioned in this topic since their use is going to change the way NSIs organise the data collection phase of the survey process.

This section will briefly describe the advantages and disadvantages in using administrative data and when and how the can be used in the statistical survey process. Full and detail information on this subject can be found in the module “Data Collection – Collection and Use of Secondary Data”.

Administrative data are “*the set of units and data derived from an administrative source*” and an administrative source is “*a data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations*” (SDMX, 2009).

The ESSNet on the “*Use of Administrative Data for Business Statistics*” (Admin Data ESSNet 2011) https://essnet.admindata.eu), which is part of the European MEETS program, is aimed at developing recommended practices on the use of these type of data in business surveys. It also reports information on projects, carried out by NSIs, to improve or increase the use of administrative data.

The use of administrative data has the great advantage of reducing data collection costs as well as respondent burden and, sometimes, to improve the timeliness of data delivery because surveys can use already existing data (Statistics Canada, 2010). But, as these data are collected for administrative purposes and not for statistical ones, their use in surveys has to be done under certain bounds/limitations. In fact, administrative data are collected by public organisations to administer or to control or to tax or to regulate the activities of enterprises or individuals. This approach is different from that followed by NSIs that collect data to study and analyse individuals or enterprises. Besides, administrative data may differ from statistical data because public organisations and NSIs may adopt different definitions of units, different definition of variables and different classifications (Calzaroni, 2010).

Very important is the issue of the quality of administrative. The definition of quality is quite complex and therefore not yet commonly shared among NSIs (Casciano et al., 2011). Apart from definition, the problem about quality lays on the fact that NSIs do not control the data collection process which is set up by public organisations that use their own control procedures that can be based on different and less stringent criteria than those used by NSIs (Statistics Canada, 2010). Besides, the quality level could be lower for those variables which are not fundamental for the administrative study but are important for statistical purposes. An overview of projects and approaches for assessing the quality of a secondary source can be found in the theme module "Data Collection – Collection and Use of Secondary Data".

The existing practices in the use of administrative data, among the NSIs, for producing business statistics, are reported in the table below. The use of administrative data combined with survey data, is divided in the four domains studied by the Admin Essnet: Business Register, Short-term statistics (STS), Structural Business Statistics (SBS) and Prodcom statistics.

The use of administrative data – from Admin Data ESSNet

Countries of the EU & EFTA by combination of direct sources used for producing business statistics and business statistics domain (end of 2010) – Table 4

DOMAINS		*COMBINATIONS*				*NON-RESPONSE*	*TOTAL*
		Admin/ register data only	*Admin/ register & survey data*	*Survey data only*	*Not specified*		
BUSINESS REGISTER		12	16	-	-	2	30
STS	Turnover	2	15	12	-	1	30
	New orders	-	10	19	1	1	31
	Production prices/costs	-	14	16	-	1	31
	Building permits	14	2	13	1	1	31
	Employment	3	16	10	-	1	30
SBS	Annexes I-IV	1	23	4	-	2	30
	Annex V	13	11	3	-	3	30
	Annex VI	16	8	2	1	3	30
	Annex VII	13	7	4	3	3	30
	Annex VIII	-	13	14	-	3	30
	Annex IX	20	7	1	-	2	30
PRODCOM		-	10	13	1	2	26

Elaboration from Admin Data ESSnet WP1, *Deliverable 1.2/2010. Database "Overview of Existing Practices in the Uses of Administrative Data for Producing Business Statistics in the EU and EFTA"* (2011).

Briefly, this table shows (Admin Data ESSNet 2011- Deliverable 1.1, pages 20-28), that: for the Business Register the majority of countries update it also by means of regular surveys; for Short-term and Structural statistics the exclusive use of administrative data is not common, but that administrative data do exist although they cannot be used as the unique source of information for reasons of quality, comparability, timeliness, etc.; from Prodcom, due to the nature of its statistics, the use of administrative data instead of direct surveys is limited.

Anyway (Admin Data ESSNet 2011- Deliverable 1.1, page 13), the statistical use of administrative data is increasing because it is recognised and sustained by national statistical laws and because the cooperation among NSIs and public bodies is improving as well as the organisation for their collection and transmission. Obviously, the situation in Europe is not homogeneous with countries that represent the optimum and others quite far from it. Examples of the optimum are represented by France and Scandinavian countries: in the first case the NSI directly manages the business register, called SIRENE, which is used for both administrative and statistical purposes; in the second case there is a very good cooperation between the NSIs and the public organisations that hold the administrative data to set up and define strategies for collecting and using these data.

Examples of use of administrative data inside the statistical survey process are:

1. Direct processing or analysis: when administrative data can replace survey data.
2. Indirect processing: when ad hoc statistical surveys are run to cover lack of information of the register or to update it.
3. Indirect estimation: when administrative data are used as input in some estimation models.
4. Survey frames: when administrative data are used as survey frames or to update them.
5. Matching with statistical archives: this could be horizontal – different archives to obtain data for the same unit - and/or vertical – different archives to obtain information for different types of units. This also known as Hybrid Data Collection (HDC) that, hence, represents a collection process based on heterogeneous administrative archives, that can change also over time (Calzaroni, 2010).

The use of administrative data for statistical purposes should be done after a strict evaluation of many aspects like their quality, their coverage, the concepts and definitions they use. In their decision process, statisticians should consider and evaluate a set of factors whose composition depends on the type of source used. A not exhaustive list of the main factors is described in the following:

- **Response burden**: evaluate whether the use of administrative data really reduces the response burden (questionnaires are not administered or shorter versions of questionnaires can be used);
- **Cost**: evaluate whether the use of administrative data can eliminate some of the steps of the data collection process thus reducing cost;
- **Coverage**: evaluate whether the population the administrative source refers to is defined with the same criteria of the survey population;
- **Concepts and definitions**: evaluate whether the concepts, the definitions of units and variables as well as classifications are coherent and suitable for the survey needs;
- **Quality**: evaluate the control process used by the public administration and whether its criteria fit with those used by the NSI;
- **Timeliness**: evaluate whether the availability of administrative data fits with survey deadlines;

- **Consistency over time (stability)**: evaluate whether data can change over time because of new administrative laws or rules or because of political changes;
- **Physical integration**: evaluate whether data are available in a convenient format in order to be easily matched with the statistical ones (if they are aggregated or not, which standardisation criteria have been used, etc.).
- **Legal issues**: be sure about the fact that their use is not limited by any privacy constrains.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Admin Data ESSNet (2011), Work Package 1 – Deliverable 1.1 "Main findings of the information Collection on the Use of Admin Data for Business Statistics in Eu and EFTA Countries" – June, 2011 (htts://essnet.admindata.eu).

Balestrino, R., Macchia, S., and Murgia, M. (2006), Data capturing strategies used in Istat to improve quality. UNECE – Work session on statistical data editing, Bonn, 25-27 September 2006.

Blanke, K., Brancato, G., Hoffmeyer-Zlotnik, J. H. P., Koerner, T., Lima, P., Macchia, S., Murgia, M., Nimmergut, A., Paulino, R., Signore, M., and Simeoni, G. (2006), *Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System.* http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/Handbook_questionnaire_development_2006.pdf.

Bohlin, M., Holmgren, T., Persson, A., Rydell, L., and Thorling, P. (2009), *Guide till svenska taxonomier för årsredovisning och revisionsberättelse*.

Budano, G. (2008), *Design and implementation of the CAPI IT system for the Labour Force Survey* (only in Italian). Istat – Metodi e norme, n.36 2008.

Calzaroni, M. (2010), *The use of administrative sources for Statistical Registers*. Naples, October 2010.

Capparucci, L., Degortes, M., Landriscina, M., and Murgia, M. (2009), Comparative analysis among open source and commercial software for the development of electronic questionnaires for statistical surveys. NTTS 2009 – Bruxelles, February 2009.

Casciano, C., De Giorgi, V., Luzi, O., Oropallo, F., Seri, G., and Siesto, G. (2011), Combining administrative and survey data: potential benefits and impact on editing and imputation for a structural business survey. UNECE - Work Session on Statistical Data Editing (Ljubljana, Slovenia, 9-11 May 2011).

Chen, P. (1976), The Entity Relationship Model: Towards a Unified View of Data. *ACM Transaction on Database System* **1**, 9–36.

Clayton, R. L., Searson, M. A., and Manning, C. D. (2000), *Electronic data collection in selected BLS establishment programs* (Clayton_R@BLS.gov).

Couper, M. P. (2001), *Web Surveys: The Questionnaire Design Challenge*. Survey Research Center, University of Michigan (mcouper@umich.edu).

Fanning, E. (2005), Formatting a paper-based survey questionnaire: best practices. In: *Practical Assessment, Research & Evaluation*, Volume 10 m.12, August 2005.

Iverson, J. (2009), Metadata-driven Survey Design. *IASSIST Quarterly*, Spring – Summer 2009.

Istat (1989), *Manual of surveys techniques*. Notes and reports, 1989, volume 3.

Konttinen, J.-P. (2012), Rationalising data collection: automated data collection from enterprises. UNECE Seminar on New Frontiers for Statistical Data Collection, 31 Oct-2 Nov 2012.

Murgia, M. and Simeoni, G. (2005), Improving the Quality of the Assisted Coding of Occupation in CATI Surveys through Control Charts. SIS Conference - Classification and Data Analysis - Cladag 2005 (Parma, June 6-8, 2005).

OECD Glossary of Statistical Terms, http://stats.oecd.org/glossary/.

O'Neil, G. E. (2008), *Developments in Electronic Survey Design for Establishment Surveys*. United States Census Bureau.

Parent, G. and Jamieson, R. (1999), *The use of CAI for the collection of business surveys in Statistics Canada*.

Roos, M. (2008), The Dutch Taxonomy Project and structural regulatory business reporting: impact for Statistics Netherlands. 94th DGINS Conference 25–26 September 2008, Vilnius, Lithuania.

SDMX (2009), *Metadata Common Vocabulary*.

Statistics Canada (1998), *Survey Methods and Practices*. Catalogue no. 12-587-X.

Willeboordse, A. (ed.) (1998), *Handbook on the design and implementation of business surveys*.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Main Module
2. Questionnaire Design – Electronic Questionnaire Design
3. Questionnaire Design – Editing During Data Collection
4. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method
5. Data Collection – Design of Data Collection Part 2: Contact Strategies
6. Data Collection – CATI Allocation
7. Data Collection – Collection and Use of Secondary Data
8. Coding – Main Module

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1.

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

Data Collection-T-Techniques and Tools

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-02-2012	first draft	M. Murgia	ISTAT-Italy
0.2	08-05-2012	second draft	M. Murgia	ISTAT-Italy
0.3.1	05-09-2012	third version	M. Murgia	ISTAT-Italy
0.4	08-03-2013	glossary review	M. Murgia	ISTAT-Italy
0.5	19-11-2013	fifth version	M. Murgia	ISTAT-Italy
0.6	03-12-2013	EB revision	M. Murgia	ISTAT-Italy
0.6.1	05-12-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:50