United Nations ECE/CES/2011/30

Economic and Social Council

Distr.: General
17 March 2011

English only

Economic Commission for Europe

Conference of European Statisticians

Fifty-ninth plenary session
Geneva, 14-16 June 2011
Item 3 of the provisional agenda
Organization of data collection and sharing, and the management challenges for the implementation of Statistical Data and Metadata eXchange

Organization of data collection and sharing for national statistical organizations

Note by Statistics Canada

Summary

Statistics Canada's organizational model and approach to data collection and sharing has evolved over time in response to changing information requirements, technological advances and a commitment to finding the most efficient and effective ways of fulfilling our mandate. Our agency supports statistical programs in four primary domains: household, agriculture, business and institutional. These statistical programs have evolved around a range of surveys supplemented by information gathered from administrative sources. This supporting paper touches on four important aspects of the organization of data collection and sharing at Statistics Canada: the evolution of our organizational model for statistical collection approaches to multi-mode data collection, the use of paradata for designing and managing surveys and models for sharing survey and administrative data.

I. Organizational models for statistical collection

1. Statistics Canada operates on a centralized model with planning and management functions spearheaded from our head office. Collection is managed by a central branch within the agency including line responsibility for collection operations conducted through three regional offices. The head office Collection Planning and Management group work directly with survey clients and collection partners to conduct collection feasibility assessments, develop survey collection specifications, collection budgets, create and test electronic collection applications, create survey procedures manuals and training materials for interviewers, and monitor survey progress in terms of cost, schedule and quality. Statistics Canada's regional offices manage our national survey interviewing force which currently stands at approximately 800 experienced field interviewers and 1200 telephone interviewers working from five telephone interviewing call centres located in western, central and eastern regions of Canada. A core interviewing capacity is maintained with the flexibility to expand and contract to meet the demands of variable survey workloads. The agency conducts a mix of core funded ongoing surveys and ad hoc cost recovery surveys.

A. Evolution of collection modes

2. In the past, paper questionnaires were the primary collection mode. They were administered by Statistics Canada representatives who visited sampled units to collect the required information. In time, a mail-out mail-back methodology was adopted, where self-administered paper questionnaires were completed by respondents. As technology advanced, Statistics Canada adopted Computer Assisted Interviewing (CAI) making use of the Blaise survey application development software to build electronic questionnaires for use in interviewer administered field and telephone interviewing. When compared to paper based survey questionnaires, CAI allowed for built in edits, savings in mailing costs, efficiencies in questionnaire registration, keying and coding, and improved data quality. The advent of CAI also allowed Statistics Canada to open a number of call centres across Canada to conduct Computer Assisted Telephone Interviewing (CATI). This has proven to be a very cost effective approach for interviewer administered surveying. CATI has yielding significant savings in collection costs over in-person field visits, both in terms of time per case and travel costs to visit sampled units. CATI also allows for more contact attempts per case than would be practical in a field administered survey. Computer Assisted Personal Interviewing (CAPI) remains a key component of Statistics Canada's collection strategy as often times the content, length or nature of surveys require an in person visit from an experienced interviewer. Our national interviewing force is at the core of Statistics Canada's strength as Canada's National Statistical Agency.

3. Statistics Canada also conducts numerous multi-mode surveys. Business surveys often have an initial mail-out component (with or without telephone pre-contact) followed by a CATI component to conduct failed edit follow-up and/or non-response follow-up. For the last few years the agency has also provided an electronic questionnaire option for several business surveys.

4. There are also several household surveys with multi-mode collection strategies. Survey samples are either split between CATI and CAPI or in some cases, a sequential multi-mode approach has been adopted.

5. On the horizon is a move towards increased use of internet based electronic questionnaires as the primary mode of collection in a sequential multi-mode collection strategy.

B. Recent initiatives to enhance Statistics Canada's data collection model

6. Statistics Canada is currently implementing a major corporate business architecture (CBA) re-engineering initiative with collection operations at the heart of many of these changes. The goal of CBA is to be more efficient, responsive and adaptive as an organization and in doing so reducing duplication through centralization and optimization of our business practices. The key initiatives for the collection organization within CBA include: maximizing our regional interviewing capacity by moving business survey collection (that had been conducted out of our head office divisions) out to our regional call centers, modernizing and streamlining collection systems (including shared systems between ongoing survey and the census organization), developing an improved electronic questionnaire option and deploying it as the primary collection mode for most surveys. The key drivers behind these changes are ensuring the relevance of the information we collect and disseminate, improving the quality of the information we collection, improving our organization efficiency and flexibility so as to remain adaptive to change.

C. Challenges

7. E-questionnaire technology has to be adapted for use in complex social surveys.

8. As take up rates for the electronic questionnaire mode increase, to what extent should we adapt different collection approaches for other modes?

9. With the advent of new collection systems and wireless technology for interviewing can we replace expensive call centre infrastructure and have virtual call centres using home based interviewing? How do we adapt our management approach to effectively operate in a virtual mode?

10. Can we evolve to the point where all electronic collection applications are driven by one integrated platform (i.e. one electronic questionnaire used for CATI, CAPI and Internet based collection)?

11. Can collection systems be adapted to allow for real-time collection and transmission of results?

II. Multi-mode data collection

A. Collection challenges – survey environment

12. Maintaining survey response (especially amidst budgetary pressures) has become a preoccupation among collection/survey managers. Over the past few years there has been a gradual downward trend in many household survey response rates. This is a direct result of increased difficulty in making contact with respondents and gaining their cooperation. Although the drop in response has not been dramatic enough to raise major data quality flags, if the trend continues, this may be inevitable. Of equal importance is the fact that for many surveys the amount of collection effort required to try and maintain response rates has increased resulting in added pressure on collection budgets. Statistics Canada is currently examining several initiatives aimed at increasing awareness of the importance of respondent participation in our surveys

13. Another ongoing challenge for Statistics Canada is response burden, particularly as it relates to business and agricultural surveys. Response burden is particularly heavy for large companies/operators due to the importance of their information to the overall survey estimates. These entities tend to be chosen in the sample files for numerous surveys.

Statistics Canada has programs in place to work with large significant respondents to help organize and prioritize survey requirement. In addition, we also have an ombudsman for response burden to help address respondents concerns.

B. Response options

14. One method to help increase the likelihood of response for all surveys and responding units is to have a robust and pertinent choice of response mode options. Statistics Canada surveys make use of a wide array of response options including paper, telephone interviewing (CATI), personal interviewing (CAPI), secure electronic file transfer, and a move towards enhanced use internet based electronic questionnaires. The modes used for any particular survey depend on the specific requirements of the survey program, the respondent population being targeted, as well as time and cost constraints.

15. As mentioned previously, Statistics Canada, under its Corporate Business Architecture Initiative, will introduce electronic questionnaire as the primary response mode for many of its surveys. This will provide a response option that is convenient for respondents to use, more time effective, and will provide a high quality, cost effective data collection option. This will also allow the agency to use electronic questionnaire as a pillar in a sequential multi-mode approach to data collection.

C. The electronic questionnaire vision

16. Under Corporate Business Architecture, over the next 5 years, Statistics Canada plans to create an E-questionnaire option for most business surveys as well as a number of key household surveys (including the Labour Force Survey). The goal is to develop a generic, electronic questionnaire development system to meet the needs of most surveys. This will encompass the principle of reusability through the development of standardized content blocks, generic specification templates and standardized functionality. Statistics Canada is already working with other National Statistics Offices, to share experiences and best practices to help ensure an efficient and effective roll-out of this new mode.

D. Challenges

17. Research is required to determine the ‘mode effect’ impacts associated with Electronic Questionnaire. This is especially important for surveys whose current primary mode of collection is interviewer assisted (CAPI or CATI).

18. Statistics Canada is also examining the known limitations of electronic questionnaire as it relates to other modes currently in use. For example, limitation in the use of electronic edits, and gaps in the availability of collection paradata for this mode (i.e. in comparison to the robust Blaise audit trail data available from surveys currently in CAI modes).

19. In order for electronic questionnaires to become established as a cost effective method of administering surveys, certain take up rates need to be achieved. What type of ‘push strategies’ would be best adopted to achieve this end?

20. Research is also needed to determine which respondents are using the electronic questionnaire mode. If these are typically respondents who were already highly cooperative under current collection modes, then the average cost per unit for remaining respondents requiring collection through other modes will rise.

21. Coverage – electronic questionnaire will initially target areas of high internet connectivity within Canada. As there are still parts of the country where either broad band

coverage or internet take-up rates are low, research will be required to examine quality and cost impacts on survey outputs.

22. One benefit of the electronic questionnaire mode is the ability to more efficiently transfer questionnaires with pre-filled data to respondents. Security and confidentially concerns surrounding this plan need to be reviewed.

III. Using survey paradata in designing smart surveys and operational approaches

A. Paradata

23. With the advent of computer-assisted data collection, the compilation of detailed information about the data collection process became feasible. Such information is referred to as paradata or process data, though we prefer to reserve the latter term for the broader concept of information about the survey process more generally. At Statistics Canada, we produce paradata for both computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI). The data fall into three major categories: transaction histories, audit trails and information from the interviewer pay system.

24. Statistics Canada uses the Blaise system extensively to conduct its surveys. One output of Blaise is the Blaise Transaction History (BTH) file. BTH files contain detailed information about the call attempts made for each case (a case is a sampled unit such as a dwelling and its associated telephone number). For each contact attempt, the BTH file contains a variety of information including the start and end times of the attempt and the outcome code. For some CAPI surveys, transaction information comes from the Case Management (Caseman) system. The data is similar to, but more limited than, the information in BTH files. A second type of paradata available under computer-assisted interviewing is audit trail data, which provide a "blow by blow" record of each interviewer-case interaction (e.g., the question-by-question interaction during an interview). The information generated includes start and end times for each question, which edits were triggered, and events such back-tracking to a previous question. Such data is very useful for finding "problem" questions. The data generated is very voluminous since a record is generated for each event. The third major source of paradata is the pay systems for interviewers. Interviewers keep track of their activities (travel, interview, training, etc.) and use task codes to record these and claim hours worked, kilometres travelled, parking fees, etc. Unlike transaction and audit trail data, these are not automatically generated, therefore there can be "response errors" present (e.g., underreporting of small tasks).

25. To be useful, the raw paradata need to be processed. What give the paradata their value are the variables, such as call duration, that can be derived from the raw data and the subsequent analysis of these variables. By linking pay data with transaction data using interviewer ID, project code and date, a variety of interesting cost and productivity analyses are possible. Because of the large volume of data that is generated, particularly for audit trail data, storage, tracking and archiving of the data becomes a challenge. As part of the Corporate Business Architecture initiative mentioned earlier, Statistics Canada will establish data service centres. One such centre, for survey data and the associated paradata, will have meeting this challenge as one of its goals.

B. The use of paradata for collection operations

26. Paradata are useful principally because they shed light on the data collection process by augmenting anecdotal information with objective measurement. Furthermore the data are useful at various stages of the survey process. Before collection, data from previous surveys are useful for budgeting and planning. For example, they can be used to determine how many interviewers to have on hand at different times of day and how to distribute call attempts for a given case (some in the morning, afternoon, evening). During collection the data are useful in continuous monitoring, in detecting and predicting problems, and so on. We can produce daily or even (nearly) real-time reports on the collection process. This allows *active* management of the collection effort. This can be formalized into a *responsive design* approach to data collection, described below. Finally, after collection, the paradata can be analyzed and used to prepare for future surveys.

C. Responsive design at Statistics Canada

27. Responsive design (RD) is an adaptive approach to survey data collection that uses information available prior to and during collection to adjust the strategy for the remaining cases. In responsive design, the usual goals of maximizing the response rate and controlling costs are augmented by other considerations: quality, productivity, the response propensity of in-progress cases, the mode of collection and competition from other surveys for resources. The RD strategy for CATI surveys tested at Statistics Canada breaks down the data collection process into four phases: planning, initial collection, RD-1 and RD-2. The *planning phase* occurs before data collection starts. The main activities of the planning phase include the analysis of previous data collection cycles to identify improvement opportunities, frame and sample assessment and validation, the development of active management tools and reports and the establishment of staffing plans. The second phase, *initial collection*, begins with the start of collection and ends when the first responsive design phase, is initiated. During this phase, new features can be introduced in the collection process. The third phase, *RD-1*, categorizes and prioritizes in-progress cases using information available prior to the beginning of collection and paradata accumulated during collection. The objective is to improve overall response rates. The last phase, *RD-2*, aims to reduce the variance of response rates among the domains of interest to improve the representativity of the sample by targeting cases that belong to domains with lower response rates.

28. This approach to responsive design was tested on two Statistics Canada surveys. The impact of responsive design on the collection process observed in these tests was generally small but almost uniformly positive. In an era of declining response rates, the fact that the two surveys increased their response rate is encouraging, especially because less effort was required to achieve comparable response rates under RD. The observed improvement in representativity for one of the surveys is also reassuring. For more details, see the papers by Laflamme and Karaganis (2010) and Tabuchi et al. (2009).

29. The implementation of responsive design led to the development of new management tools and procedures that will be useful for future surveys. Implementation for other surveys and for more than one survey at a time will require generalization of the tools and programs, good documentation and well-trained, dedicated staff.

D. Challenges

30. The greater role of electronic questionnaires has implications for paradata as well. There will be paradata from this mode, but it will be of a different type, posing new challenges for the survey manager.

31. The growing importance of multimode surveys has already been noted. The use of different modes implies different types of paradata will be available and these need to be combined in a sensible manner for proper analysis and use.

32. Keeping track of all the paradata (storage, archiving) is a challenge—suggesting the need for metadata on paradata! On the positive side, having a long history of paradata (going back years) allows us to conduct studies even years later.

33. Reconciling paradata from different sources for the same survey is not always as straightforward as one would hope (e.g., time data from BTH files are system-generated whereas the corresponding pay system data are interviewer-coded).

34. The approach to responsive design developed and tested by Statistics Canada is not the only option. Other approaches should be studied and tested. In addition, approaches to responsive design under multiple modes, including electronic questionnaires, need to be developed.

35. User-friendly software tools are needed both to aggregate and analyse paradata and to actively manage the data collection process. The tools developed for responsive design at Statistics Canada are only a starting point.

36. Paradata are useful for methodological research. For example, Statistics Canada is developing a simulation model of certain aspects of the collection process which uses paradata as inputs. Another example is the development of a new indicator of interviewer productivity. Paradata can also be used to improve non-response adjustment. We expect that these are just the first of many potential uses of paradata.

IV. The approach to sharing data

37. The sharing of data refers to the provision of detailed microdata to persons or organizations that are not employees of Statistics Canada. Data collected as part of a household survey or through acquiring administrative data from organization present distinct challenges for data sharing. In both cases, the practices of the sharing data are embedded in the Statistics Act. The Statistics Act awards the necessary powers to Statistics Canada in order to collect the data towards institutions and identifies the obligations of Statistics Canada in regards of preserving the confidentiality of the data.

38. In past years, Statistics Canada developed an organisational model comprised of different access modes that recognize the various needs of users (researchers, political analysts, students and the general public).

A. The context of sharing administrative data with Statistics Canada

39. Replacing direct data collection with the use of administrative data reduces the response burden and generates savings. Over the past ten years, productive efforts have been made to expand the use of administrative data for statistical purposes, especially the use of tax data for producing business statistics. Data from administrative sources presents the distinct challenge of being a “census” of events which by its very nature increases the risk of individual records being identifiable. Statistics Canada has taken two approaches to

providing access to administrative data. The first has been to obtain the permission from the data supplier to share the data with third parties. Of course, the application of this approach is limited to a select group of trusted partners who will preserve the confidentiality of the data. The second approach is to provide access to the data under strict conditions in accordance with the Statistics Act. Analysts and researchers granted such access are sworn in as "deemed employees" of Statistics Canada and are subject to the legal obligations under the Statistics Act of ensuring the confidentiality of the data.

B. The context of access to data held by Statistics Canada

40. Over the years, Statistics Canada has developed a model consisting of different access modes designed to accommodate the diversity of users' needs (researchers, policy analysts, students and the general public), as well as the challenges unique to each data type (survey and administrative data, economic and social data, microdata and aggregate data).

41. Currently most social surveys produce microdata files that have been reviewed to ensure no confidential data are accessible. These files are made widely available through university research facilities and other methods of dissemination. The master files for many social surveys, which contain confidential data, are also available under strict conditions in Statistics Canada Research Data centers at over 20 locations affiliated with Canadian universities across the country. All researchers accessing microdata in these centers must be sworn in as deemed employees of Statistics Canada subject to the legal obligations of ensuring the confidentiality of the data.

42. Recently we have successfully piloted a remote access approach to master file microdata. Although the functionality of the remote access is somewhat limited it holds the promise of dramatically improving access to a wide range of survey and administrative data. This new application will automate data processing and verify the results to ensure that no confidential information is disclosed.

43. Future improvements will include enhancing the model to handle more complex requests as we expand access to data to strategic decision-makers and academic researchers.

44. In terms of protecting the confidentiality of respondents, the challenges created by microdata from business surveys are different from those associated with social data. By their size, industry type or other distinct characteristics respondents to business surveys are considered to be easily identifiable despite any efforts to mask or suppress characteristics. This has traditionally been considered an insurmountable problem for providing more access to these data.

45. Statistics Canada has begun consulting with businesses and trade associations to elicit their advice on procedures for accessing the databases. This consultation guides the development and implementation of a strategic framework for access and also allows us to maintain collaborative ties with businesses. Statistics Canada pays particular heed to its collaborative relationships with business, and to the fact that other businesses have an interest in accessing the databases of the former in order to glean trade secrets. Statistics Canada must consider the impact that a complaint from a business could have on the entire Canadian statistical system.

46. In light of rapid developments in communications technology, as well as heightened sensitivity to the issue of protecting personal data that is associated with this technology, Statistics Canada is under increasing pressure to justify its activities in the context of protecting personal and business data. The challenge is to satisfy the demand for access to data while controlling costs and protecting the confidentiality of respondents.

47. Recognition of the diverse needs of users creates some challenges. The more access modes there are, the greater the risk of residual disclosure. Diversifying and broadening access to microdata must thus take into account the existence of each access mode and its interaction with other modes. Some users also fear that one option may disappear as a result of another being delivered.

48. The development of all these access options places additional burdens on Statistics Canada's human and financial resources, but also on its ability to rapidly disseminate the collected data, given that timeliness is an important component of data quality. Strategies to mitigate this impact must be proposed.

V. References

Laflamme, F. and Karaganis, M. (2010), Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada, presented at the European Conference on Quality in Official Statistics (Q2010), Helsinki, Finland.

Tabuchi, T., Laflamme, F., Phillips, O., Karaganis, M. and Villeneuve, A. (2009), Responsive Design for the Survey of Labour and Income Dynamics, Proceedings of Statistics Canada Symposium 2009, *Longitudinal Surveys: from Design to Analysis.*