A GENERALIZED METHODOLOGY FOR ECONOMIC SURVEYS

Michael Colledge and Robert Lussier, Statistics Canada

ABSTRACT

In 1985, Statistics Canada embarked on a project to redesign its major surveys of economic production and financial statistics. The objective of the project is for all surveys to use a common central frame and a generalized methodology. This should facilitate the integration of statistics and reduce the cost of future survey designs, redesigns and upgrades. The paper provides an overview of the methodology. It focusses on policies and strategies rather than techniques.

1. INTRODUCTION

1.1 Background

The economic statistics program at Statistics Canada includes surveys of financial, industrial, commodity, employment, capital expenditure and taxation statistics collected on a monthly, quarterly, annual and occasional basis. There are roughly 300 surveys (depending upon the precise definition of a survey), of which about 125 are subannual and the balance annual or occasional.

In 1985, Statistics Canada initiated the Business Survey Redesign Project, the main objective of which is to standardize and integrate all the economic program systems and ouput data. The objective is to be met through the mandatory use of a common central frame and of a generalized survey methodology. This should not only facilitate the integration of statistics and rationalize operations, but also reduce the cost of future survey designs, redesigns and upgrades as specific methodologies and tailormade systems are replaced by generalized ones.

It should be mentioned at this point that the term "business" which appears within the project title is used within Statistics Canada to include units of economic production engaged not only in the trade and commercial service industries but also in manufacturing, construction, transportation, etc. Indeed, the term is often extended to include the economic activities of professionals. Thus "business," "organization" and "economic entity" are used synonymously in this paper to refer to any form of economic production unit.

The project constitutes a major effort to build quality into the economic program. The benefits which it is hoped will materialize include:

(a) standardization of concepts, definitions, classification schemes and survey procedures;
(b) more comprehensive frame coverage, more precise delineation of large economic entities and their reporting arrangements, and more reliable classification;
(c) increased use of administrative data, reduced response burden and improved respondent relations;
(d) reduction in overall frame maintenance, mailout and data collection costs;
(e) enhanced facilities for integrating data and increased scope for audit;
(f) development of generic systems and procedures for a wide range of survey functions including automated/computer-assisted industrial coding, sample size determination, allocation and selection, edit imputation, etc.

The original project strategy was formulated three years ago (Cain et al.,1984). It has been summarized by Colledge and Lussier (1985) and Colledge (1987).

1.2 Contents of Paper

The paper deals with the generalized survey methodology being developed during the course of the project. It covers all the functions of an economic survey in chronological order, i.e., in the order in which they usually take place in production. The description focusses on strategy rather than technical details.

Section 2 summarizes the basic design for the acquisition, processing and use of frame data. This design provides the foundation upon which all future survey development will take place. Section 3 describes the set of generic functions which are applicable, in principle, to all surveys. Sections 4 and 5 deal with special features of these functions for subannual and for annual surveys, respectively. The concluding remarks in section 6 are followed by abbreviations and references.

2. FRAME DATA DESIGN

Economic surveys will be based upon standard statistical entities and classification schemes. The current practice of multiple exceptions will not be continued. However, the target population and data requirements for a given survey can not be specified without consideration of the availability and ease of extraction of the data from respondents' accounting records. It is a complicated procedure to determine the set of statistical entities within a large organization for which the data are to be collected, and to set up reporting arrangements by which the data can actually be acquired. To assist in this process, an information model which recognizes the complexity of large economic organizations has been developed. It incorporates five distinct types of entity:

(a) legal - for example, incorporations under federal or provincial charter;
(b) administrative - for example, payroll deduction account holders, income tax filers;
(c) operating - for example, divisions, profit centres, plants, etc., corresponding to the way in which the business organizes itself and keeps its operating accounts; the legal, administrative and operating units jointly define the view the business has of itself;
(d) statistical - the target entities for statistical measurement purposes, i.e., the statistical agency's view of the business;
(e) reporting - providing the linkage between the statistical target units and the business world economic operating entities.

Experience indicates that a single statistical entity is inappropriate for collection of the full range of economic data from a large organization. Different classes of data are available at different levels within the organization. To deal with this problem yet maximize the potential for subsequent integration of the various data items, a four level hierarchy of statistical entities has been defined. Each level corresponds to the capacity to report data, as follows:

(a) statistical enterprise - the highest level of statistical entity, associated with an autonomous organization, capable of reporting all forms of

economic data about itself;

(b) statistical company - a subdivision of the statistical enterprise, corresponding to an entity capable of providing an unconsolidated financial report, e.g.,a division of an enterprise;

(c) statistical establishment - a subdivision of the statistical company, roughly corresponding to a profit centre, able to report production statistics and the components of value added;

(d) statistical location - a subdivision of the statistical establishment, corresponding to an individual plant, warehouse, retail outlet, etc., capable of reporting revenue and, possibly, employment.

Thus, for example, an organization may be viewed for measurement purposes as comprising one statistical enterprise, two statistical companies, 4 statistical establishments and 6 statistical locations. The complete set of entities at each level in the hierarchy will, in principle, cover the entire business universe, and will be matched to the capacity to report a particular class of data.

Delineation of statistical entities will thus depend upon the way in which organizations structure their operations and keep their records. The process of creating and maintaining these "structures," i.e., lists of entities and their relationships, is termed "profiling" at Statistics Canada. It is a costly and time consuming operation for large organizations and it can involve considerable respondent burden. To reduce the resources required and the burden on respondents, maximum use will be made of administrative data.

2.1 Administrative Data Sources

Three data files from Revenue Canada, namely the employer payroll deduction (PD) and the personal (T1) and corporate (T2) income tax files will be the principal administrative sources. However, as the files do not have a common identification scheme, their complete integration into a single frame would require a massive expenditure on record linkage and cannot be contemplated. The problem will be circumvented by having a new "Central Frame Data Base" (CFDB) with two components: an "integrated portion" (IP) and a "non-integrated portion" (NIP).

The IP will contain a unique and unduplicated list of the statistical entities covering every large and complex business. The creation and maintenance of this list will require complete linkage and reconciliation of all administrative and other sources. The statistical entities will be generated automatically from the operating structure. They will be fully classified, linked to the corresponding legal, operating and administrative entities, and tracked through time.

The NIP will contain two sets of statistical entities giving two completely independent views of the small business universe not covered by the IP. The first set, termed "income tax based" will be derived from the T1 and T2 income tax files. The second, "PD-based" set will be obtained from the payroll deduction source, which provides a more timely flow of updating information. No systematic attempt will be made to relate the records in these two sets. They will provide alternative frames. Within each set, the hierarchy of statistical entities will be kept simple: all four types of statistical entity will be presumed to coincide with one another and with the corresponding administrative entity. There will also be size boundaries below which entities will be considered out of scope for survey purposes.

The reason for dividing entities into integrated, non-integrated and out-of-scope categories is to focus resources on the integrated portion entities which account for a large proportion of national economic production, and to adopt simplified, less costly procedures for the remainder. In particular, processing the administrative files and maintenance of the small entities will be automated as far as possible. This seems to be a plausible strategy as the economy is dominated by a relatively small number of large businesses.

The CFDB will be updated to reflect, as far as possible, all relevant changes of structure and of classification which occur in the business world. This is a complex process. Updating information is available from a variety of different sources. There is a bewildering number of alternative ways in which the sets of legal, administrative, operating, and statistical entities and their relationships can be updated, and there are some pitfalls to avoid. For example, changes of legal structure such as mergers, amalgamations, takeovers, creations of subsidiaries, etc., do not necessarily imply any changes in the corresponding operating or statistical structures. Thus, if the sets of statistical entities were to be updated automatically as a result of information from administrative or legal sources, there might be a speciously high incidence of apparent "births" and "deaths" of statistical entities, and an attendant risk of incomplete or duplicate coverage.

To handle this situation Armstrong et al.(1986) have defined a comprehensive set of fifty or so "standard events" which can take place in the business world. Any indication of a change will be considered as a "signal" in response to which the corresponding entities will be investigated, and reprofiled if necessary, to deduce which, if any, of the fifty standard events have occurred. CFDB updating will always be in terms of these events. Precise definitions of births, deaths and changes to statistical entities will thus be embedded in the rules for definition of standard events and for statistical entity generation. There will be quite large numbers of signals, obtained from essentially three types of source: survey feedback; administrative processes; and CFDB routine reprofiling. These signals will be placed on the CFDB "workbench" where they will be sorted according to complexity, batched into work units, and automatically allocated to CFDB maintenance staff when they request future assignments. In keeping with the general IP/NIP approach, the treatment accorded to entities in the NIP will be much simpler than for the IP and will be automated as far as possible.

More comprehensive details of the procedures for profiling and for the initialization and subsequent updating of the CFDB are contained in Clark and Lussier (1987) and other project working papers.

3. GENERIC SURVEY FUNCTIONS

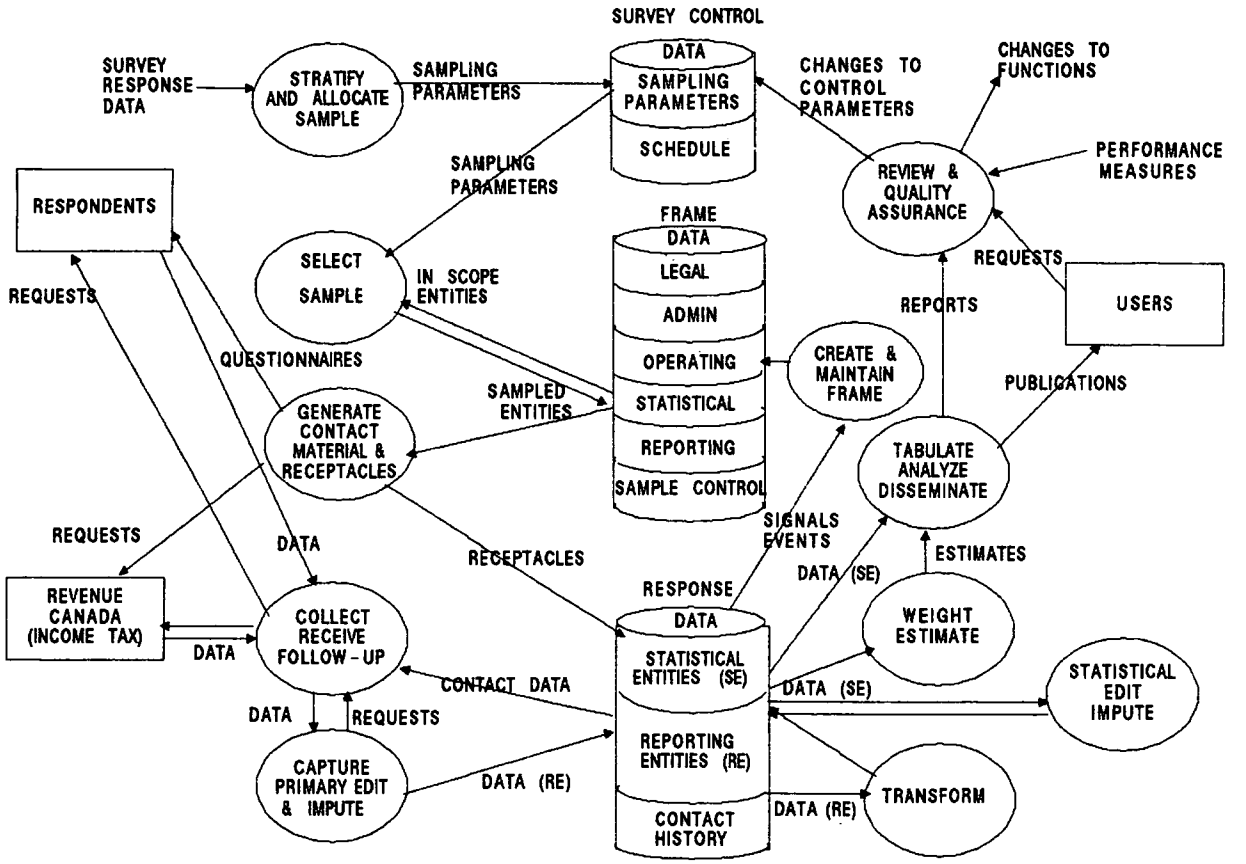
3.1 Introduction

This section deals with the basic elements of economic survey methodology. It focusses on concepts and procedures which have been and are being developed during the course of the Business Survey Redesign Project and which represent, to a lesser or greater extent, departures from current practice.

The methods described are generic in the sense that they have been reviewed and accepted as standard for application to all surveys under the project umbrella. The methods are grouped into eleven generic "functions," described with reference to the data flows and storage files appearing in Figure 1. The principal storage elements are:

FIGURE 1:--GENERIC FUNCTIONS, DATA FLOWS AND STORAGE

(SIMPLIFIED, AND EXCLUDING FRAME CREATION AND MAINTENANCE DETAILS)



(a) the survey control data base, containing sampling parameters and schedules specific to each survey;
(b) the CFDB, comprising legal, administrative, operating and statistical entities, as previously noted, which are shared by all surveys, and reporting entities and sample control files which are survey specific;
(c) the response data base, containing data from reporting entities and for statistical entities, for each survey together with contact history data to be shared by all surveys.

The use of these files and, in particular, the distinction which is made between reporting entity (RE) and statistical entity (SE) data, is elaborated in the following sections.

3.2 Frame Creation and Maintenance

Every economic survey will, in principle, draw its frame for each survey "occasion" from the CFDB. The frame is the appropriate set of statistical entities and associated classification data, determined in accordance with the survey coverage and sampling requirements. For certain surveys the set of entities used for sampling purposes will be at a different level in the statistical hierarchy than the set which comprises the statistical target. For example, the target population for the wholesale trade survey will be the set of statistical establishments classified to the wholesale industry, whereas the sampled population will be the set of statistical companies with one or more wholesale establishments. The reason for this particular complexity is a preference, based on experience, to collect data for an entire organization rather for just a part of it. This has been interpreted, in terms of statistical entities, as a requirement to obtain data covering a complete statistical company not just individual component establishments.

To promote good coverage and efficient sample selection, it will be standard practice for new entities ("births"), inactive entities ("deaths"), and changes in classification of entities to be recorded on the CFDB and utilized in the formation of survey frames as soon as the data become available. Such information will be obtained from administrative research, profiling operations, and the surveys themselves. Survey feedback is recognized in being a particularly rich source of deaths and, in certain circumstances, of classification changes. This has lead to the following considerations and decisions.

Any data obtained during the course of survey processing which is of significance for frame updating will be fed back to the CFDB, via the contact history file. This will be the only mechanism by which the survey process will update the survey frame. Thus no deviation between the frame as derived from the CFDB and as used by the survey will be allowed to develop.

In forming survey frames after CFDB updating, the incorporation of births poses no particular problem, but purging deaths and introducing classification changes is not as straightforward. Most sampling procedures

require a systematic overlap of the entities selected on successive survey occasions. This can mean, for example, that entities should remain in the sample even after they have been reclassified to strata with different sampling rates. Furthermore, the statistical dependency between updating based on survey feedback and the selection of future samples implies a potential for bias in future estimates as would happen, for example, if deaths detected by the survey process were simply deleted from the frame.

As the starting point for addressing these problems, i.e., for controlling sample overlap and producing unbiassed estimates, the sampling history of each entity for each survey will be recorded in the survey control file. This file will be the device by which the frame data items associated with statistical entities will be extended to include the sampling information, specific to each survey, required for weighting and selection of future samples. This is further elaborated in Section 3.4, and in the context of subannual surveys (Section 5).

3.3 Stratification and Allocation

For most surveys stratification will be by geography, by industry and by size. The first two items will be used to meet publication objectives and the third is required for efficiency reasons. The highly skewed nature of distribution of economic entities by size dictates the need for a "take-all" stratum of large units to be selected with certainty. The methods being considered to determine the boundary between the take-all and remaining "take-some" strata are given in Hidiroglou (1986) and Lavallée and Hidiroglou (1987). The first of these methods provides a boundary on the assumption that a given level of sampling precision has to be satisfied by a sample of minimum size drawn by simple random sampling from two strata, one being the take-all. The second method is an extension which allows for any number of take-some strata.

The basic procedure for determining sample size and allocation will be to specify, for a key data item, the coefficients of variation required nationally, and provincially, and for various groups of industries, then to use Hidiroglou's method, or a "power" allocation as summarized by Bankier (1986).

In contrast to most of the other functions which will be carried out on every survey occasion, stratification, sample size determination and allocation will be performed from time to time, on an ad hoc basis, as needed. The corresponding strata definitions and sampling parameters will be stored in the survey control data base.

3.4 Sample Selection

As far as possible this function will be completely automated. On a preset date recorded in the survey control data base, sample selection for the survey occasion will take place. The set of statistical entities which are in scope will be accessed from the CFDB together with the complete survey sampling history contained in the sample control file. The historical data will include the classification values of all entities as of the dates they were first subjected to sampling. It will cover all entities ever subject to sampling, even those subsequently found to be dead but knowledge of which is still required to enable the production of unbiased estimates.

A simple random sample will be selected in each stratum in accordance with the survey sampling parameters, and the sample control file will be correspondingly updated. Occasion to occasion sample overlap will be controlled so as to balance the competing requirements for sample replacement (to reduce individual respondent burden and to facilitate efficient estimates of "level") and for sample continuity (to produce efficient estimates of "change"). These ideas are further discussed in Sections 4 and 5.

3.5 Generation of Contact Material and Survey Receptacles

Reporting entity data embody the reporting arrangements made with respondents to collect data for statistical entities in the sample. When a statistical entity is added to the sample, default reporting arrangements will be generated automatically from the operating entity data in the CFDB. This will provide the required information for the first contact. There is provision at subsequent stages of processing to update default reporting entities with the information collected during the first contact.

For surveys with a mail out/mail back data collection procedure, contact material will be automatically generated on a preset date recorded in the survey control data base. The material will be mailed to the selected reporting entities. The completed questionnaires will be returned to the local regional office or to the head office. Also, regardless of the particular data collection methodology, contact data for control purposes possibly including "organizational portraits," will be automatically generated and sent to the regional offices and/or the head office, as appropriate. These data will be used in initiating telephone interviews if a telephone procedure is used, in logging returned questionnaires, and for follow-up and frame data editing.

At the same time as contact material is generated two sets of "receptacles" to receive the data will be created in the response data base. The first of these sets, the reporting entity (RE) receptacles, will be in one to one relationship with the reporting entities from which the data are to be acquired. The second set, the statistical entity (SE) receptacles, will refer to the target entities for which data are to be derived by transformation of the reporting entity data (as elaborated later). The motivation behind this explicit identification of the entities from and for which data are to be collected is to ensure that the survey response data base remains precisely in step with the CFDB frame. Data for which there are no receptacles will not be used directly in forming survey estimates.

3.6 Collection, Receipt and Follow-Up

Generally, annual surveys will use a mail-out/mail-back procedure whereas sub-annual data will be obtained by telephone interview conducted from the regional offices. Consideration is being given to the use of CATI. When the regional offices or head office receive data from respondents, either by mailback or by telephone, the corresponding identification information will be logged immediately on a computer data base. In this context, the use of optical bars on questionnaires is being investigated.

All contacts made by each survey will be recorded on the contact history data base. This facility is intended to make possible the fast exchange of responses, respondent status and collection problems between surveys and the frame profiling operations. The data base will be consulted prior to any contact of a respondent. By this means, unnecessary duplicate contacts will be avoided, and, if more than one contact in a short space of time is essential, then the staff member making the contact will be able to explain the situation. It is believed that this sort of procedure will

enhance the relationship of the agency with its respondents.

Follow-up will occur not only for non-response but also to resolve some edit failures. The policy will be to contact all non-respondents, not just a sample. The procedure used will be a telephone interview from a regional office or from the head office, depending on the survey and on the reason for follow-up.

3.7 Data Capture, Primary Editing and Imputation

Capture of the reporting entity data and the primary editing will be done in the same operation. "Primary edits" are defined as those which check the validity of data without resort to inter-entity comparisons and which may require reference to the original source documents to resolve edit failures. They include within fields edits (e.g. checks on numerics, alphabetics, and maximum values), between field edits (e.g. checks on accounting relationships), and between occasion edits. Once the data for an entity have passed through this process, the source document will be archived and probably not used again.

Requests for follow-up to resolve certain types of edit failures will be generated and handled as previously noted. Captured, edited reporting entity data will be stored in the corresponding receptacles on the RE response data base. In certain circumstances, however, there may be no appropriate response receptacle. This may occur if a change takes place in the structure of a business but is not reflected on the CFDB at the time the receptacles are generated. The change may well be detected, though not fully identified, by the response on a survey questionnaire, or by the subsequent follow-up of an edit failure. The survey staff will feed back all the pertinent information as a "signal" to the CFDB staff. As and when investigation of the signal is complete, and the corresponding business world events have been identified, the CFDB will be updated and appropriate new receptacles may then be generated. Prior to this point in time, the associated survey data will be in temporary storage and will play no part in subsequent processing, except possibly to impute values for the original receptacles.

To illustrate these points, consider the Monthly Retail Survey which collects data on the number of locations, and the sales, within prescribed geographic by industrial strata. Suppose a business in the sample opens a new location and this is not recorded on the CFDB at the time the response receptacles are created. Information about the new location will be fed back to the CFDB. If, however, the location is in a domain for which no RE receptacle exists then the data will be ignored, or at most used for imputation, until the frame information has been processed and a new receptacle created via the CFDB.

3.8 Transformation

The data for the reporting entities will be transformed into data for the target statistical entities according to prescribed rules defined at the time the reporting entities are created. For example, it is possible that a business which has two operations in different provinces cannot break down its sales data by province on a monthly basis, yet such a breakdown is required for provincial statistics. In this case, there will be two statistical establishments but only one reporting entity. Aggregated data will be collected from the reporting entity and then allocated internally, at the time of transformation, to the corresponding statistical entities.

Transformation is a vital step in the process of data integration. Items of information collected in accordance with survey-specific reporting arrangements are converted into data for the sets of standard statistical entities. By this means, data from a variety of surveys can more easily be brought together.

3.9 Statistical Editing and Imputation

Statistical edits involve comparisons between the characteristics for each entity and the estimated distributions of these characteristics. They include, for example, edits based on percentiles. These edits will be performed upon statistical entities rather than upon reporting entities because the former should, in principle, be more homogeneous, being defined on the basis of standard concepts rather than individual reporting arrangements.

The application of statistical edits will be automated. Edit failures will be listed for manual inspection in an order determined by the magnitude of the deviation from average behaviour and the potential impact upon the estimates. It is entirely possible that, due to time and resource constraints, not all the listed items will be investigated. Data items failing statistical edits will continue through subsequent stages of processing unless corrected by manual intervention.

Review of statistical edit failures will reveal not only incorrect data but also legitimate values which will greatly influence survey estimates. Amongst these values will be a certain number that are believed to be unrepresentative of the domain to which they belong. They will be defined as "outliers" and action will be taken to reduce their impact upon the estimates by adjusting the sampling weight, or reported data, so that the product of the two becomes more appropriate. If the sampling weight is reduced, the weights of other entities in the same sampling stratum will have to be correspondingly augmented in order to retain the same stratum population count. Certain entities flagged as statistical edit failures, whether defined as outliers or not, may be labelled as inappropriate for use in imputation. Such exclusion is motivated by the need to preserve the average distributions of the various characteristics being measured.

The final step performed by this function will be to impute for missing and incorrect data items. According to the circumstances, various methods will be used including hot deck, ratio, regression and use of stratum means. In most cases knowledge of the distributions of the characteristic to be imputed will be required, just as for statistical editing. This is the reason why, as far as possible, imputation will be performed on statistical rather than reporting entities. The only imputation which must be performed at reporting entity level is that required to enable transformation of the data.

3.10 Weighting and Estimation

A weight equal to the inverse of the original probability of selection will be assigned initially to each entity. In general it will, subsequently, be adjusted for achieved sample size so that the weights will sum to known population totals. Usually a single weight will be used for all data items for a given entity, i.e., data items will not be separately ratio adjusted to different totals.

Domain estimation will be used to produce estimates. The domains will be defined by the most current classification values available from the CFDB for the statistical entities and the given survey reference period. Thus, these domains may differ from the original sampling strata because of subsequent changes in size, industry or location of entities or

because the stratification for sampling was done at a coarser level than is required for estimation. The original stratification may even have been based on statistical entities at a different level in the hierarchy, as previously noted in section 3.2.

It is worth stressing that changes in classification will be reflected immediately in the estimates, not accumulated over time and then introduced en bloc. It is recognized that, due to time lag in the detection procedures, the time at which a particular change is introduced may not necessarily reflect the time at which it actually occurred in the business world. However, it has been judged preferable to accept the slight time lag bias than to explain the discontinuities in the estimates which would occur after accumulated classification changes have been made. It is believed that, at the aggregate level, time lag bias will not substantially affect the trend.

One of the ideas that will be pursued in the future is estimation of components of the change between two occasions. The components may be defined as:

(a) changes of classification of entities;
(b) changes of structure within entities, for example the creation of a new establishment;
(c) changes of structure affecting more than one entity, such as a take-over;
(d) true "births," i.e.,new entities not resulting from (a), (b) or (c) above;
(e) true "deaths," i.e., disappearance of entities not resulting from (a), (b) or (c) above;
(f) changes of data values for continuing entities, i.e.,changes not resulting from modification to classification or structure.

The magnitudes of each component will be determined, for important variables, and in terms of the number of units involved.

3.11 Tabulation, Analysis and Dissemination

Survey products will continue to be disseminated as they are at present, in the form of regular and occasional printed publications (catalogues, bulletins, newsheets), and a publicly accessible data base (CANSIM). In accordance with agency policy (Statistics Canada, 1986), descriptions of survey objectives, concepts and methodology, and measures and comments on reliability will be made available. The results of data analyses may be disseminated regularly, for example, as publication highlights, or on an occasional basis in special publications or at conferences, seminars, etc. Special requests for estimates or micro data will continue to be serviced. All dissemination will be subject to confidentiality constraints.

3.12 Review and Quality Assurance

The major role of senior survey staff will change from supervision of survey operations to review of the output, and quality assurance. Survey staff will be responsible for analyzing the estimates and the micro-data and for producing the text which will accompany the figures in the publications. They will also examine the performance measures routinely generated as part of every survey function. Their observations will lead to improvements in future repetitions of the survey, for example, changes in sample size, modifications to the statistical edits, or revisions to data collection procedures to improve unacceptable response rates.

4. DISTINCTIVE FEATURES OF SUB-ANNUAL SURVEYS

4.1 Introduction

The generic methods described in the previous section will apply to all types of surveys covered by the project. This section refers to special features of subannual surveys.

Sub-annual economic surveys at Statistics Canada can be classified into four groups as follows:

(a) Industry-specific surveys of production. Examples are the monthly surveys of: manufacturing shipments, inventories and orders; retail trade; wholesale trade; building permits; and restaurants, caterers and taverns. Not all industries are covered by these surveys and the data content varies considerably from one to another.
(b) Surveys of finance. These are two quarterly surveys, of industrial corporations and of financial institutions, which together embrace almost the whole industrial spectrum.
(c) The monthly survey of employment, payroll and hours, which also has broad industrial coverage.
(d) Miscellaneous other sub-annual surveys.

In general terms, the objectives of sub-annual surveys are to provide timely measurements of economic trends nationally and for various geographic and/or industrial breakdowns. Timeliness, not detail, is the essential criterion in survey design. More specifically, the level of detail, both as regards data content and geographic/industrial breakdown, should be much less than for annual surveys. The reasons for this are the difficulty or impossibility of acquiring detailed structural data sub-annually, the response burden, costs, and processing time.

4.2 Frames

Frames for sub-annual surveys will be drawn from the sets of statistical entities in the CFDB IP and PD based NIP. Only in exceptional cases will alternative frame sources be used.

For most surveys, the target statistical entity will be the statistical establishment and the sampling unit will be the statistical company. Exceptions will be the monthly retail trade survey for which the target unit will be the statistical location, and the quarterly financial survey for which the target and sampled units will be the statistical enterprise.

For each survey, the frame for the first occasion will be created from scratch by identifying the statistical entities which are in scope. This set of entities will be recorded on the sample control file with their classification values and with sampling data produced by the ensuing sample selection process. For subsequent occasions, the frame generation process will start with the set of in-scope statistical entities and classification data, which will have been updated since the previous occasion. These data will be compared with information recorded on the sample control file for the previous occasion. Births will be identified and automatically incorporated in the frame. However, deaths will not be simply purged, nor entities moved to new sampling strata in accordance with changes in classification, because of the potential bias due to survey feedback.

Five approaches to the treatment of deaths have been considered in some detail, as follows.

(a) Delete all dead entities from the frame and the sample, regardless of the source of information, and ignore the bias on the grounds that it is insignificant. This is not a practical proposition as studies have shown that as many as 20% of the entities on the NIP PD-based frame may be dead, and the survey process will certainly detect those in sample.
(b) Retain all dead entities on the frame, suitably

flagged so as not to be subject to data collection if they fall into the sample. This is unconditionally unbiassed, but very inefficient.

(c) Delete from the survey frame only those entities signalled as dead by a procedure independent of the survey process. This is somewhat difficult to implement in practice as the original source of information can be obscured by subsequent profiling operations, or by feedback from other surveys. It is also inefficient, but an improvement on (b).

(d) Delete all dead entities from the frame but keep running estimates, based on the sample, of the numbers of deaths within each stratum, and adjust the sampling weights as needed to provide (nearly) unbiassed estimates. This option was originally developed in the context of the survey of employment (Schiopu-Kratina and Srinath, 1986).

(e) A variant of the fourth option, but instead of adjusting the weight, certain numbers of dead units are retained on the sample control file and in the sample so that normal estimation procedures will produce unbiassed estimates.

No final decision regarding the choice of approach has yet been made, though it is likely to be one of (c), (d) or (e).

As regards changes in classification, an extension of approach (e) for deaths could be applied to allow updating of classification values based on survey feedback, without significant bias. The underlying principle would be to ensure that updating of misclassified entities took place at the same rate for those entities in the survey sample as those outside it. However the procedure would be more complicated than for the treatment of deaths, as it would involve maintenance of a matrix of misclassified sampled entities to represent misclassification by stratum on the frame. Hence a simpler procedure, analogons to approach (b) for deaths, will be adopted. The original classification of each statistical entity, as of the date it was first subjected to sampling, will be retained on the sample control file and used for weighting. Consideration is being given to procedures for bringing these data into line with the current classification values, possibly in association with a sample redraw, but the details have not been finalized.

Once the frame for a survey occasion has been identified and a sample selected, the frame for that occasion will be frozen, i.e.,no units added, deleted or reclassified. Subsequent updating information will be used in defining the frame for the next occasion. The rationale behind this decision is that the minor improvements in the frame which could be expected by allowing last minute updates would be far outweighed by the associated statistical and operational complexity.

4.3 Sample Selection and Rotation

A sample will be selected from scratch on the first survey occasion, and updated for each subsequent occasion. The intention of updating will be not only to include new statistical entities and remove dead ones but also to rotate the sample so as to reduce the burden on individual respondents wherever this is possible, i.e., in the take-some strata.

The basic criteria which have been established for rotation are that once an entity has rotated into sample it should be there for "x" survey occasions; and that once an entity has rotated out of sample it should not be rotated back in for at least "y" survey occasions. Typically, the parameters x and y will have values of 12 for monthly surveys, implying a one-year period in the sample, followed by at least one year out.

Three approaches have been investigated for rotation; rotation groups; sampling intervals; and panels. In all three cases, the rotation procedure is applied to each sampling stratum separately.

The rotation group method consists of randomly selecting the entities to rotate into sample on the next survey occasion from a set which are "waiting for selection," i.e., not currently nor recently in the sample. The entities are assigned to a particular sample "rotation group." At the end of a fixed period of time, e.g.,one year, the entities in the group will be rotated out of the sample and placed in a "not eligible for selection" category. At the end of the appropriate period out of sample, the entities will again be waiting for selection, and they will have lost their original rotation group label. This is the method currently used for the survey of employment, payroll and hours. Schiopu-Kratina and Srinath (1986) provide details.

The sampling intervals method consists of first assigning equi-spaced "sample selection numbers" at random to the statistical entities. Those entities which fall within a particular interval are then selected for the first sample. Rotation is performed subsequently by shifting the position of the interval on each survey occasion.

The panel method consists of allocating the entire set of entities to equi-sized panels prior to selection of the first sample. The number of panels within each stratum is chosen in accordance with the number of units available and required rotation rate. The panels are randomly ordered, and the first x panels are chosen to be in the first sample. ("x" is the number of survey occasions for which entities are to remain in sample.) On each subsequent occasion, births are added to the panels, then one panel is rotated out and another in. The detailed methodology is presented in Hidiroglou and Srinath (1987).

These methods are similar in their advantages and their drawbacks. For example, panels and rotation groups both tend to become uneven over time and, equivalently, the equispaced numbers no longer evenly spaced. The panel method will probably be adopted.

4.4 Generation of Survey Response Receptacles

In accordance with the decision to freeze the frame for a given occasion at the time of sample selection, the survey response receptacles will be correspondingly frozen. Data for entities for which there are no appropriate receptacles will not be used, except for imputation.

4.5 Data Collection by Accounting Period

Sub-annual surveys will produce estimates for each calendar month, or for each calendar quarter (beginning January, April, July and October). However, respondents will be allowed to report data for accounting periods of their choice to suit their particular accounting practices. Thus, for example, a respondent may report for a four week period to a monthly survey, or for quarters, beginning February, May, August, November, to a quarterly survey. Allowing this flexibility will increase the respondent's capacity to report and reduce the response burden, but at the expense of having to adjust the reported data to the required reference period.

4.6 Weighting and Estimation

The possibilities of using administrative data and annual survey data to improve subannual survey estimates or to reduce sample sizes, are being

investigated. In particular, payroll deduction data are being studied in connection with the survey of employment, and it will be standard practice to benchmark subannual estimates to annual figures whenever more reliable annual data exist. In the latter context the various possibilities have been outlined by Laniel (1987).

As noted in Section 4.2, there may be an adjustment of the sampling weights to compensate for the removal of deaths from the survey frame and sample.

5. DISTINCTIVE FEATURES OF ANNUAL SURVEYS

5.1 Introduction

This section refers to special features of the annual surveys covered by the project. Subsections 5.1 to 5.6 cover sampling, data collection and processing for statistical entities of all sizes whereas the final two subsections deal with features peculiar to small entities in the NIP.

Annual surveys at Statistics Canada can be categorized as follows:

(a) Industry specific–surveys of economic production. Each of these surveys is concerned with a particular group of industries. Examples are the annual surveys of manufacturing, construction, transportation, retail and wholesale, and selected services. Not all industries are covered.

(b) Surveys of finance, labour income, capital and technology. This group of surveys is concerned with financial, taxation, labour income, capital stocks and expenditures data, balance of payments and external trade. Ownership and control, and technology data are also collected under the Corporations and Labour Unions Reporting Act.

(c) Surveys of small areas and small businesses. (Neither of these programs is as yet fully established).

(d) Miscellaneous other annual surveys, e.g., surveys of energy, of research organizations.

In general terms, the objectives of annual surveys are to provide structural information regarding production, finance, employment, ownership, etc., data, at the finest level of detail for which there is a demand, and which data sources and agency resources can support on a regular annual basis.

The project is focussed on surveys falling within the economic production and the financial categories. A major component of the project strategy is to bring together all the industry specific production surveys and to regard them, conceptually if not operationally, as being a single "survey of economic production."

5.2 Frames

The frames for annual surveys will be drawn from the sets of statistical entities in the IP and tax–based NIP. Only in exceptional circumstances will alternative frames be used. For production surveys, the target statistical entity will be the statistical establishment and the sampling unit will be the statistical company. For financial surveys, the target statistical entity and the sampling unit will be the statistical enterprise. The target population for reference year Y will be defined as all statistical entities with fiscal years ending in the period April 1, Y to March 31, Y+1. (This is further explained in Section 5.4).

It will be possible to update the frame during the course of processing for a given reference year, i.e., the frame will not be frozen at the time of sample selection as it is for sub-annual surveys due to shortage of processing time. New entities will be added and dead entities will be excluded until the point at which the frame must be fixed for weighting and estimation purposes.

Statistical entities in scope but not in sample for annual surveys will be subject to regular re-identification and classification review as part of the CFDB maintenance profiling procedure. Thus it will be possible to incorporate classification feedback from a survey into the frame for subsequent years with negligible risk of introducing bias.

5.3 Sample Selection and Control: Rotation

Rotation of annual samples is not such a significant factor as for subannual surveys. For the larger statistical entities there will be little scope for rotation as most of them will fall into the "take-all" category. For the smaller entities data will be obtained primarily from income tax returns (as described shortly), and the requirement for rotation is thereby reduced. Thus, as a general rule, year to year sample overlap will be maximized to facilitate efficient estimation of annual change.

5.4 Data Collection: Concepts and Strategy

Income tax returns will be an important source of financial data for annual surveys. This will alleviate respondent burden and it will reduce agency costs. The collection strategy will be different for statistical entities in the IP than for those in the NIP.

IP entities will be subject to direct, full–scale survey. If there is a one–to–one correspondence between a statistical entity selected for sampling and a tax record then, as an alternative to direct survey, financial data may be acquired from the tax return. If the correspondence is not one to one, tax data will not be used. This may be the case for a large business in which the set of statistical entities, defined on the basis of operating structure, bears no direct relationship to the legal structure and hence to the set of legal entities for which tax returns are submitted.

For NIP entities, financial data will be obtained from tax returns. These data will be supplemented where required by direct survey of "other characteristics" i.e., of non-financial items. Details of the sampling design are given later in subsections 5.6 and 5.7.

For both IP and NIP, entities the procedure for acquiring financial data from tax returns will be along the following lines. The returns for which data are to be obtained will be indicated to Revenue Canada as a set of identification numbers and/or a sampling algorithm. The corresponding returns will be intercepted and copied. The copies will be sent to Statistics Canada where the required financial data will be extracted and captured.

Respondents to direct survey will be asked to report data for their particular fiscal years. These data will be summed, with adjustment to account for the range of fiscal periods, to create data for the reference year. A process of continuous mailout and data collection will be adopted, with timing matched to each respondent's fiscal year end. This arrangement will suit respondents by making it easy for them to report. It will ensure that data are collected on as timely a basis as possible and, to some extent, will spread work load. It will also be in accordance with the way respondents complete their tax returns.

With the above approach, data can, in principle, be published for the population of economic entities having fiscal years ending in any given twelve–month period. However, for consistency and ease of interpretation, a particular period has been nominated as the basis for

"reference year Y" statistics. In making this choice, the principal options considered for the range of fiscal year ends were:
(a) January 1, Y to December 31, Y;
(b) April 1, Y to March 31, Y+1; and
(c) July 1, Y to June 31, Y+1.

The problem with option (a) is that it produces statistics covering a calendar period from January Y-1 to December Y, i.e., is not centred on calendar year Y, even taking into account the preponderance of fiscal periods ending in December. Option (c), though nominally centred on calendar year Y, does not provide much better coverage due to the uneven distribution of fiscal periods; and it implies a six months' delay in production of yearly statistics. Thus option (b), which does give a more balanced coverage than (a) and implies only 3 months delay, was chosen.

5.5 Generation of Response Receptacles

In accordance with the provision for updating the frame for a given year, new response receptacles will be added as required during the course of survey processing, thus providing a more up-to-date structural image of the business universe.

5.6 Estimation

The annual surveys will produce estimates that represent the calendar year. However, as the data will be requested and collected according to the fiscal year of the respondent they will have to be adjusted to calendar year. Where available, subannual survey data will be used as the basis for adjustment, but the details have yet to be finalized.

5.7 NIP Use of Tax Data

A sample of NIP statistical entities will be selected and financial data will be acquired from the corresponding tax returns to meet the collection needs of all annual surveys. A two phase design will be used. An overview of the design is presented in this paragraph; the following paragraphs contain more detailed descriptions of each phase. The first phase sample will be built, and maintained from year to year, by sampling and copying tax returns at Revenue Canada, transporting the copies to Statistics Canada, identifying the corresponding statistical entities and classifying each by industry, geography and size. The set of statistical entities thus created will constitute the first phase "master" sample, from which a second phase sample will be selected. For each member of the second phase sample required data items will be extracted from the financial statements which form part of the tax return.

For first phase sampling, the population of tax returns will be stratified by industry, province and size. The stratification by industry will be fairly coarse (at approximately 2-digit level of the 1980 SIC), corresponding to the degree of precision which can reasonably be expected from the business descriptive items which appear on tax returns. Selection will be based on a unique pseudo-random number generated for each tax return by applying a hashing function to the tax identification number as suggested by Sunter (1986). Tax returns for which the hashed numbers fall within prescribed intervals will be included in the sample. Use of the same hashing function and sampling interval parameters each year will produce a master sample with maximum year-to-year overlap. This will not preclude the possibility of changing the sampling intervals from time to time, to expand or contract the master sample, for example, to improve efficiency or reduce resources. Some changes in the master sample from year to year will inevitably take place due to births, deaths and changes of classification. Procedures for purging returns which become superfluous are being considered. However, as a general guideline, once a tax return has been identified as belonging to the master sample it will remain there for all subsequent years. The original sampling weight associated with a return will be retained except where the classification change moves the return into a stratum with a higher sampling fraction. In this case the return will take on the smaller weight corresponding to the new stratum.

For each tax return selected in the first phase sample, one statistical establishment will be created for each separate, complete set of financial statements attached to the return. Most tax filers submit only one set of financial statements and, in this case, a single statistical establishment will be defined. If, however, a tax filer attaches two or more sets of statements, each of which relates to a separate business, then a separate statistical entity will be created for each business. Tax filers may also be in partnership with one another. No attempt will be made to link the partners. Instead, correction for the duplication on the frame due to partnerships will be made by deflating the sampling weight of an entity so as to reflect the partner's share.

The first phase master sample will be stored in the CFDB tax based NIP. (It will actually constitute the tax based NIP.) It will provide the frame of statistical entities from which the second phase sample of tax returns will be selected with stratification by 3 or 4 digit SIC. The tax identification numbers of this second phase sample will be sent to Revenue Canada so that the corresponding tax returns can be intercepted and copied. In addition, a proportion of "births", i.e., returns entering the master sample for the first time during the current reference year will be added to the second phase sample to ensure it is representative. From each copied tax return the required financial items will be extracted, captured and stored on the tax response data base. They will subsequently be merged with other characteristics data obtained by direct survey for the same entities, weighted, and added to data from IP statistical entities to produce the estimates for each survey.

In conclusion, it is worth noting that the processing of NIP tax data will follow the same generic procedures outlined in sections 3 and 5.2-5.6 for annual surveys, with the provision that, in accordance with the general IP/NIP philosophy, operations will be simplified and costs reduced. Thus, administrative sources rather than profiling operations will provide the basic sampling frame; statistical entities per se will be defined only for a master sample and, even here, partnerships will be allowed for by adjustment of the weights rather than by establishing partnership links. The "reporting entities" will be tax returns; the "contact material" will be sampling parameters and tax return identifiers sent to Revenue Canada.

5.8 NIP Other Characteristics Survey

To complement the acquisition of financial data from tax returns for NIP entities, there will be direct, industry-specific surveys of other characteristics. Each such operation will collect data on production, commodities, services, employment, etc., according to the particular group of industrial activities being covered. These operations will be collectively viewed as constituting an "Other Characteristics" (OC) survey, the design and processing of which will be in accordance with the same generic methods already described.

The first phase tax master sample will provide the OC survey frame. The sample selected from this frame will be constrained to be a subsample of the second phase tax sample, so that, for each statistical entity in the OC sample, financial data items will have been extracted from a sampled tax return.

The OC survey will be mailout/mailback. The generation of reporting entities, contact materials, response data base, and the data collection, capture and primary editing will follow the same lines as for collection of annual data by full-scale questionnaire. Edited OC data will be merged with the corresponding sampled tax data, jointly edited, imputed and weighted, then added to data for the IP entities, as previously noted. Details are given in Foy (1987).

6. CONCLUSION

This paper presents an overview of the generic methods which have been developed in the context of the Business Survey Redesign Project, for widespread application to the Statistics Canada program of economic surveys. Methods corresponding to earlier stages in the survey process, i.e., frame creation and maintenance, sample allocation and selection, data collection and capture, including the use of income tax data, etc., have been fairly comprehensively defined. For later stages in the survey process, many details remain to be finalized, for example, the benchmarking and confidentiality procedures, and the revision policy. In addition, a complete quality assurance program has to be developed, also a strategy for coping with the discontinuities in the data series, which changes of methods will inevitably produce.

Ultimately, all economic surveys will be redesigned in accordance with the generalized methodology. The redevelopment schedule takes into account the urgency of upgrading each survey and the associated resource implications. The first surveys currently being redesigned are the monthly and annual surveys of wholesale and retail trade and the quarterly and annual financial surveys.

The implementation of generic methods will take place through the development of generalized systems. Here "generalized system" implies a menu-driven meta-system, capable of generating specific systems to meet the particular needs of individual surveys.

In principle, generic methods and the corresponding systems should first be defined and developed, then utilized subsequently for survey redesign. In practice, due to a shortage of time, parallel development of generalized and of specific systems is taking place. More explicitly, systems to meet the design requirements of specific surveys are being developed as early prototypes of more generalized systems to follow. This is not an ideal approach, as it can compromise the generalized design. It does, however, provide the generalized development with well-defined milestones and practical applications, hence ensuring its timeliness and relevance.

The gains to be expected from the introduction of a generalized approach will accrue not only from the reduction in the costs of future redesign and maintenance work but also from improvements in quality. Each generic method is being very thoroughly discussed, analyzed and tested by all interested parties, including statisticians, systems analysts and survey operations staff. The result will be a bank of standard methods which have been given a universal stamp of approval. This should lead to better quality, and more easily integrated data.

7. ABBREVIATIONS

CFDB - Central Frame Data Base.
IP - Integrated Portion of the CFDB.
NIP - Non-Integrated Portion of the CFDB.
OC - Other Characteristics.
OOS - Out-of-scope for surveys.
PD - Payroll Deduction Account.
SIC - Standard Industrial Classification.
T1 - Personal Income Tax Form.
T2 - Corporate Income Tax Form.

8. REFERENCES

ARMSTRONG, G., MONTY, A., WOODS (1986), "Definitions of Standard Events," Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa.

BANKIER, M.D. (1986), "Power Allocations: Determining Sample Sizes for Sub-National Areas," Social Survey Methods Division working document, Statistics Canada, Ottawa.

CAIN J. et al. (1984), "Infrastructure Development, Objectives, Policy and Strategy," Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa.

CLARK, C., LUSSIER, R. (1987), "The Use of Administrative Data for Initial and Subsequent Profiles of Economic Entities," Proceedings of the Section on Survey Research Methods, 1987, American Statistical Association, Washington (in this volume).

COLLEDGE, M. (1987), "The Business Survey Redesign Project - Implementation of a New Strategy at Statistics Canada," presented at the Bureau of the Census Third Annual Research Conference, Washington.

COLLEDGE, M., LUSSIER, R. (1985), "A Strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys ," Proceedings of the Section on Survey Research Methods, 1985, American Statistical Association, Washington.

FOY, P. (1987), "Development of the OC Capability for the Annual Surveys of Economic Production," July 1987, Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa.

HIDIROGLOU, M.A. (1986), "The Construction of a Self-Representing Stratum of Large Units in Survey Design," The American Statistician, 40, 27-31.

HIDIROGLOU, M.A. and SRINATH, K.P. (1987), "Sample Rotation," Business Survey Methods Division working document, Statistics Canada, Ottawa.

LANIEL, N. (1987), "Benchmarking of Business Surveys: Problem Discussion and Strategy to Derive a Solution," September 1987.

LAVALLÉE, P., AND HIDIROGLOU, M.A. (1987), "On the Stratification of Skewed Populations," Business Survey Methods Division working document, Statistics Canada, Ottawa.

SCHIOPU-KRATINA, I. and SRINATH, K.P. (1986), "The Methodology of the Survey of Employment, Payroll and Hours," Methodology Branch Working Paper No. BSMD-86-010E, Statistics Canada, Ottawa.

STATISTICS CANADA (1986), "Policy on Informing Users of Data Quality and Methodology," March 12, 1986 , Ottawa.

SUNTER A. (1986), "Implicit Longitudinal Sampling from Administrative Files: A Useful Technique," Journal of Official Statistics, Vol. 2, No.2, pp 161-168.