Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

THE REDESIGN OF STATISTICS CANADA'S BUSINESS REGISTER

Paul Hunsberger, Yanick Beaucage and Stuart Pursey[1]

ABSTRACT

The structure of Statistics Canada's Business Register (BR) was developed in the mid-1980s. Its major role is to provide a frame that covers more than 2.3 million active Canadian businesses, for more than 90 recurring business surveys. In past decades, there have been enormous changes within and outside Statistics Canada. The globalization and evolution of fiscal policies of the world economy has imposed upon Statistics Canada a revaluation of the conceptual framework of the BR. Changes in informatics technology provide opportunities to rework the informatics infrastructure of the register. The goal of the BR Redesign Project is to simplify, optimize, and harmonize its processes and methods. This paper provides an overview of the BR Redesign with emphasis on the issues that affect the methodology of business surveys. Major methodological issues are the use of a business' Operating Structure in developing a sample design for business surveys; procedures to maintain and update the data of businesses; methods to determine the birth and death of businesses; and the development of a quality assurance strategy.

KEY WORDS: Business Register, Redesign, Frame

1. INTRODUCTION

Statistics Canada's Business Register (BR) was developed in the mid-1980s. A detailed discussion of its infrastructure and approaches can be found in Colledge (1987) and Cuthill (1990). The major role of the BR is to provide a frame of more than 2.3 million active Canadian businesses, for more than 90 recurring business surveys.

The purpose of this paper is to describe the Business Register Redesign, a project designed to completely revise and update the infrastructure and approaches of the current BR. Further details can be found in Bérard *et al*. (2005), Gagné (2004), Pursey *et al*. (2005) and Rancourt *et al*. (2005). Section 2 of this paper provides an overview of the current BR and describes the reasons for a redesign. Section 3 discusses that way in which sampling elements can be derived from the BR. Section 4 describes the use of administrative data in updating the BR. Section 5 addresses the development of a quality assurance strategy.

2. OVERVIEW OF THE CANADIAN BUSINESS REGISTER

The foundation and concepts of the Business Register—the whole economic statistics program at Statistics Canada—are driven by the needs of the Canadian System of National Accounts. The Business Register began to take shape in the early 1980s when efforts were made to establish a central frame that could be used by most if not all business surveys. Gradually over the years, more surveys have been using the BR—it is now the backbone of the business statistics programs of Statistics Canada.

Since the 1980s, there have been changes within and outside Statistics Canada. Fiscal policies, globalization, and the structure of businesses have changed. The conceptual basis behind the BR has become out of date. The revolution in informatics technology has made the informatics infrastructure of the BR difficult and expensive to maintain. New and useful administrative data sources have emerged. The functions of the BR have expanded over the years. There are more tasks and processes to maintain and update the BR. Most importantly, with all these changes, user needs have evolved. The BR needs a rethinking—a redesign.

[1] Yanick Beaucage, Paul Hunsberger and Stuart Pursey, Statistics Canada, Business Survey Methods Division, R.H. Coats Building, 11th floor, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6

2.1 The Business Register

The Business Register is a list of active Canadian businesses. The BR is the central frame for almost all of Statistics Canada's business surveys—and most especially the surveys that provide data for the System of National Accounts. The important BR variables for survey and sample design are the fields that contain data about geography, industry, employment, revenue and contact information.

There are about 2.29 million active businesses listed on the BR and 2.27 million of them are "simple" businesses. For these businesses, a single entity represents best their Legal Structure, Operating Structure, production unit, observation unit, sampling unit, and data collection unit. Although representing 99.1% of the businesses in Canada, these businesses represent only 40% of economic activity. The other businesses are "complex". There are only about 21,000 of them but they represent 60% of economic activity. There are two useful views of a business: the Operating Structure and the Statistical Structure (see Section 3). The Operating Structure is the business' view of itself—how it operates and structures itself. This structure is derived through a comprehensive interview with representatives of the business. This "profile" of the business is important to Statistics Canada because it identifies the types of production units within the complex business, the availability of types of data, and many measures required in survey design (size, geography, and industry).

The Legal and Operating Structures of a complex business are represented by a set of entities (production units) arranged in a multi-level hierarchical pattern that illustrates the reporting relationships and data availability of the production units. There are five types of production units in the Operating Structure. The Business Entity (BE) represents all of the business and all data. The Investment Centre (IC) contains data on revenue, expenses and investment. The Profit Centre (PC) contains data on revenue and expenses—in effect "value-added" can be calculated. The Profit Centre resembles what is usually referred to as an establishment. The Cost Centre (CC) contains data on expenses only. The Sales / Revenue Centre (RC) contains data on revenue only. The diagram below illustrates the Operating Structure of a particular complex business. It happens to have three levels separated into two branches, with several types of production units.

Figure 1a: A complex business

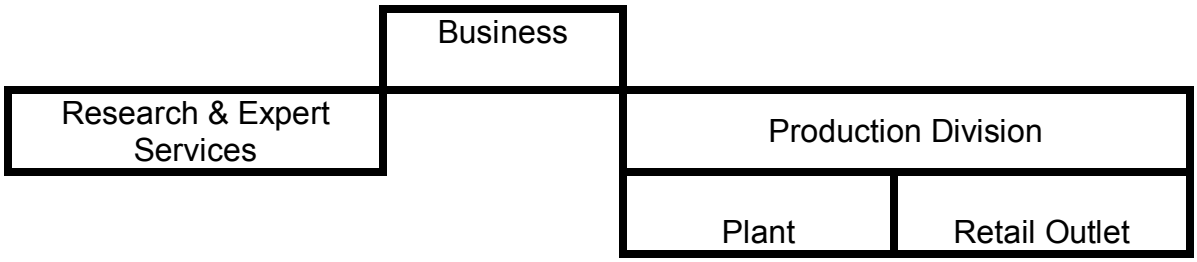
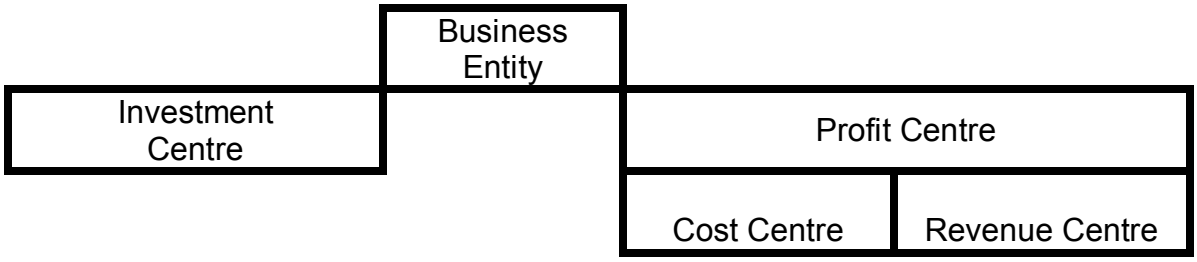


Figure 1b: Operating Structure of a complex business



3. SAMPLING ELEMENTS

3.1 Current Approach

The Operating Structure of a business is used as the foundation for deriving the second view of the business: the Statistical Structure. This view is hierarchal with four levels: Enterprise, Company, Establishment, and Location. Each level represents a type of data availability: Table 1 shows the relationship of data availability, compared to the Operating Structure. The purpose of the Statistical Structure is to provide a structure that facilitates the identification and selection of sampling elements. For example, the frame for an establishment based business survey is derived from the third level of the Statistical Structure (the Establishments).

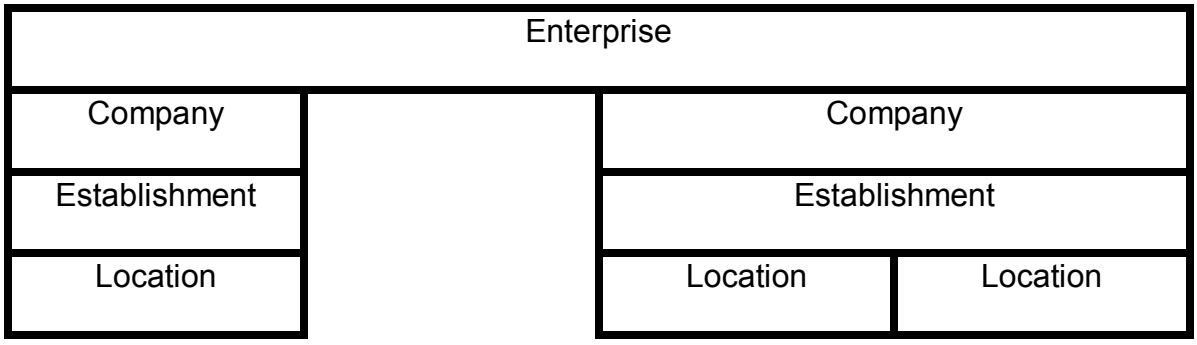
Table 1: Operating and Statistical Structures – Types of Units

Operating Structure	Statistical Structure	Availability of Data
Business Entity (BE)	Level 1: Enterprise	All Data
Investment Centre (IC)	Level 2: Company	Revenue, Expenses and Investment
Profit Centre (PC)	Level 3: Establishment	Revenue and Expenses
Cost Centre (CC)	Level 4: Location	Expenses
Sales/Revenue Centre (RC)		Revenue

STATGEN is the algorithm used to generate, from the Operating Structure, the Statistical Structure. For the 2.27 million "simple" businesses STATGEN maps the single production unit of the business into each of the four levels. For the 21,000 complex businesses (two or more levels in the Operating Structure) the algorithm expands or contracts the Operating Structure to get four levels. Generally – although there are complexities when the Operating Structure is not of four levels (or more exactly, each branch of the Operating Structure is not of four levels) – Investment Centres become Companies, Profit Centres become Establishments, and Revenue Centres and Cost Centres become Locations.

Figure 2 shows the construction of the Statistical Structure of the example shown in Figures 1a and 1b. In this example the left side branch must expand from 2 levels to 4 levels and the right branch must expand from 3 levels to 4 levels. In more complicated businesses with many levels in the Operating Structure, the levels must be collapsed to four levels. But of the 21,000 complex businesses only 2,000 of them have more than two levels in their Operating Structure.

Figure 2: Statistical Structure of a complex business (following Figures 1a and 1b)



3.2 Difficulties with the Statistical Structure

The major advantage of the Operating Structure is that it provides a realistic, accurate, and almost untouched picture of the structure of a business—the structure of the business as viewed by the business itself. If we base our sample

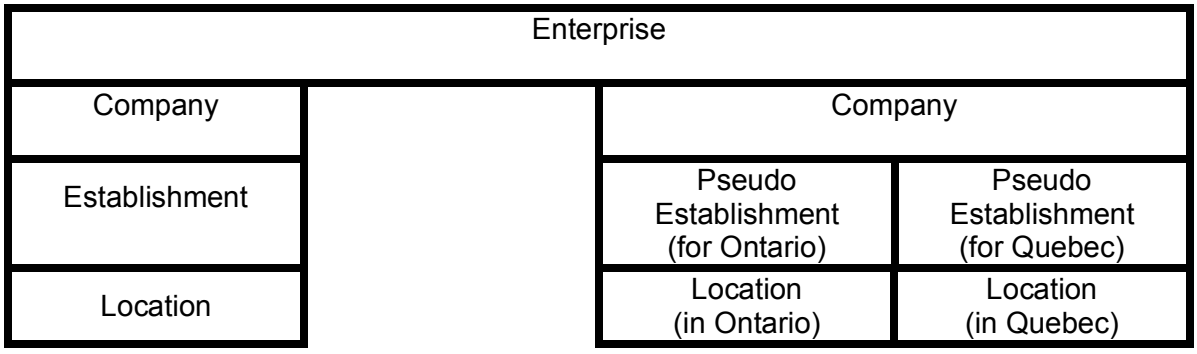
designs and data collection arrangements directly on the Operating Structure, rather than the highly structured Statistical Structure, we expect gains in collected data quality and a reduction in response burden.

The main advantage of the Statistical Structure is that it provides a straightforward way of selecting units for an enterprise, company, establishment, or location based survey. STATGEN arranges the various production units of the Operating Structure into their place on the Statistical Structure, following the data availability guidelines shown in Table 1. There are three main disadvantages to the Statistical Structure. These disadvantages have emerged over the years as we have gained experience in the use of the Statistical Structure and as our priorities to reduce response burden and to improve data quality have increased.

The first disadvantage is that the Statistical Structure is maintained within our informatics infrastructure. This is a huge cost, especially when one realizes that the impact of STATGEN is only for the 21,000 complex businesses, of which 19,000 have only two levels.

The second disadvantage is based on the requirement from the System of National Accounts that estimates be provided by province. But often a business does not operate by province, nor keep its accounting books by province. Consider Figure 1b. Although not stated explicitly in the diagram, the Revenue Centre and the Cost Centre are from the same province leading to Figure 2 where a single establishment is above the two locations. Suppose that the Revenue Centre and the Cost Centre are from different provinces. Therefore the Profit Centre has activities in two different provinces. The solution, from the view of an establishment based survey, is to create two pseudo establishments—neither is real—and Figure 3 shows the outcome.

Figure 3: Statistical Structure of a complex business (from Figure 1): the pseudo establishment case



Pseudo establishments cause problems in data collection because some survey respondents are aggravated and puzzled in trying to provide revenue and expense data for a production unit that does not exist. Still, one way or another data is collected or imputed but potentially the data quality is poor due to nonresponse or due to the difficulty in knowing exactly what the collected data refer to. This is called a "provincial split" pseudo establishment but there are two other types. When a location supports an establishment and both are on the same level in the Operating Structure the location is rolled (combined) inside the establishment and called a "roll up" pseudo establishment. When a location supports more than one establishment and all are on the same level in the Operating Structure, the location becomes an establishment in its own right (and it is called an ancillary pseudo establishment).

The third is really a generalization of the second. The Statistical Structure modifies each business into a four level structure, regardless of the number of levels it actually has. For the simple businesses (one level) this is not really a problem (beyond the cost of data processing and storage) because the result is simply a repeat of the single production unit: one for each of the four levels. For almost all of the complex businesses, there is a real distortion of the Operating Structure. The 19,000 two-level businesses are turned into four levels. Those with many levels must be squeezed into four. Two very different Operating Structures may have identical Statistical Structures.

It is often the case that a business, upon viewing its Statistical Structure, finds itself unrecognizable. The difficulty with the Statistical Structure is that it makes data collection arrangements more difficult and more confusing, potentially leading to needless response burden and poor data quality. With the current experience in today's economic context, we realize that the Statistical Structure, in effect, is a distortion of the reality of a business that at

best has no impact on data collection arrangements, response burden, and data quality and at worst leads to difficulties with these issues.

3.3 Using the Operating Structure directly to create sampling elements

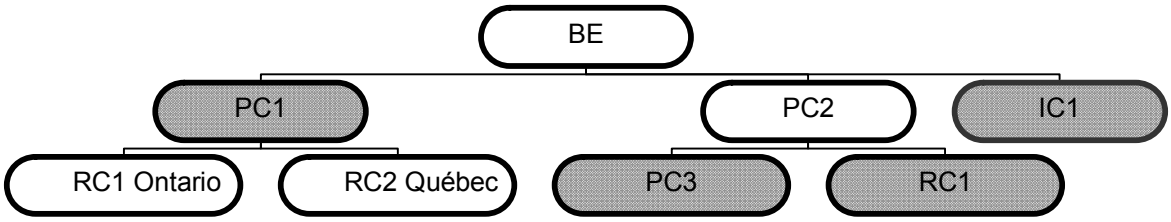
Figure 1b shows the Operating Structure of a business after its profile has been created—this structure is identical to that in Figure 1a. In an establishment based survey the objective is to survey Profit Centres (to measure Value Added = Revenue – Expenses) and to cover completely without duplication the Business Entity. Thus the sampling element is the Profit Centre. There are two difficulties that are easily overcome.

The first concerns pseudo establishments. Can we accomplish the goal of representing the provincial dimension of the data without creating artificial sampling elements? As described in Section 3.2 there are three types of pseudo establishments. The solution for provincial split pseudo establishments is to flag the Profit Centre as having multi-province activity and attach to it a vector of data that indicates the size of this activity in each province. This is the data required to create a sample design by province. (An alternate is to link the Profit Centre to the RCs and CCs below it.) For both roll up or ancillary pseudo establishments the RCs and CCs become sampling elements in their own right. This is somewhat unfortunate because RCs and CCs cannot provide value added data—but that is the reality of the business! With a realistic picture we can more easily develop data collection arrangements with the respondents that take into account production units that are not Profit Centres.

The second is that in a business with many levels, as we travel down each branch, we may encounter more than one Profit Centre. The rule is to select the "lowest PC" in a branch as the sampling element. This preserves greatest amount of the industry and geography information.

Figure 4 shows the sampling elements (shaded) for an establishment based survey using the lowest PC rule. The sampling element of the left branch is PC1 (a vector of data is needed to show provincial activity). The sampling elements of the middle branch are PC3 and RC1 (it must be recorded as a Revenue Centre). The sampling element of the right branch is IC1. The Business Entity is completely covered without duplication. Data collection arrangements can be made with respondents and value added can be calculated. There will be challenges in arranging data collection for the middle branch, but this difficulty is not hidden through the creation of an artificial establishment.

Figure 4: The sampling elements (shaded) of an establishment based survey for a business.



The timeline for the redesign of the sampling universe includes the creation and analysis of test sampling frames for major Statistics Canada business surveys throughout 2006 and 2007 before the new sampling element creation process is finalized and integrated in 2008.

4. USE OF ADMINISTRATIVE DATA

4.1 Measure of Business Size

Currently a model is used to estimate the revenue of a simple business. For employer businesses, the model is based on a monthly administrative data file called PAYDAC that provides the remittances made to the Canada Revenue Agency (CRA) of a business. This number is used to estimate an annual revenue and annual number of employees

for a business. For non-employer businesses, revenue is based on an administrative data file from CRA, the Canadian Goods and Services Tax (GST). It provides monthly, quarterly, or semi-annual data for a business. Sometimes neither PAYDAC nor GST is available and if the business is incorporated then its annual revenue (T2) from CRA is used in the model.

In the BR Redesign we want to avoid the use of models and instead provide measures of size directly with no modification beyond the usual need for editing, imputation, and outlier detection. The GST will be used to develop an annual measure of revenue size for both the employers and non-employers. As well from CRA, we will provide, as separate variables the annual T1 revenue (for unincorporated businesses) and the annual T2 revenue (for incorporated businesses). There is a relatively new file available from CRA: the monthly Payroll Deduction Accounts (PD-7). It provides directly the number of employees and it will be used to derive an annual measure of the number of employees of a business.

For complex businesses, profiling will continue to be the data source for BR measures of size data.

4.2 Births

The BR loads, on average, 18,000 business births per month. Currently the BR loads births every three months for non-employers and every month for employers. The process used for births sometimes causes "avalanches" of births. For example, a business must be coded to its full 6-digit NAICS (North American Industry Classification System) before being birthed, often leading to a backlog of businesses waiting for coding. When these backlogs become too large, more people are assigned to the coding task, creating peaks. Once or twice a year there is a large influx of births due to outside administrative sources. The difficulty with irregular birth processes is that potentially there are sudden peaks in the estimates from business surveys. Although it is difficult to manage the various sources of administrative data so that births are timely, we need to explore ways to get around this difficulty and if not, find ways to lessen the impact of untimely births. For example, one possibility is to accept industry coding at 3-digit or 4-digit NAICS, rather than the longer time required to code at 6-digit NAICS. This would most likely go a long way to allowing monthly updates of births. Careful negotiation with partners, with the goal of modifying processes within and outside Statistics Canada, may avoid the "avalanches" at less frequent intervals.

4.3 Deaths

The current rules to determine that a business is dead are conservative. It is easier to deal with businesses that are dead during estimation (using domain estimation) than it is to deal with businesses thought to be dead, but which are alive (leading to under coverage). But it is costly to maintain dead businesses on the BR, sample designs are not as efficient as they could be, and data collection funds are not spent effectively. We need to find a balance between protecting against the bias due to under coverage yet keeping sample designs and data collection efficient. There are several administrative data sources that are accurate for determining, with near certainty, the death of a business. These files will be incorporated into the BR Redesign and a process created to "death" these businesses twice a year. Survey feedback also provides a source for identifying dead units, but it is restricted to businesses selected in a sample. To avoid bias in estimation, the source indicating that the business is dead (i.e. the survey or the administrative data file) should be kept and used to update survey frames only when it is independent of the survey.

There is another type of business death, a business may be in the process of closing its operations and be found dead at the time of data collection. As well, a business may be inactive (while administrative data files indicate that it is alive) and spring back to activity. In the BR Redesign we will explore the development of methods to manage businesses with inconclusive death signals. Here are three examples of possible approaches. A survey process may decide to use a decision rule; based on the inconclusive signals to exclude a business from a survey frame, yet be kept alive on the BR. The rule would consider both the risk of over-coverage and the risk of under-coverage. A survey may use a relatively inexpensive pre-contact before data collection to confirm the business' dead / alive status. Other possible uses of the inconclusive signals are in the imputation strategy and in estimating the quality of the alive / dead status on the BR.

5. QUALITY ASSURANCE STRATEGY

The Business Register Redesign provides an opportunity to implement a completely new and updated quality assurance strategy. This strategy should focus on the six dimensions of quality; accuracy, relevance, timeliness, accessibility, interpretability and coherence. Currently, quality reports from the BR operations are neither centralized nor standardized. They are mostly ad-hoc reports and special operations. Updates to the BR are done every month using administrative sources. Feedback from these updates needs to be clearly reported in such a way that we can evaluate the process over time. A number of data items should be part of this report including the number of updates to key variables and the magnitude of these updates, the number of new entrants (births) and new records considered dead. In order to assess the quality of the NAICS classification coding for new entrants, an informal quality control process is currently in place. For new coders, there is 100% quality control done until they reach a 5% rate of miscoding, and afterwards, only spot checks are done.

Another measure in place to assess the quality of the coding and the coverage of the BR is the Quality Assurance Survey (QAS) which is done from time to time on the BR. This survey was implemented most recently for reference year 2002 and should be held again in 2006. The QAS provides a report card assessing the proportion of correctly coded NAICS on the BR. When units are miscoded, it could be due to one of two potential reasons: the unit was incorrectly coded, or the unit has recently changed code. Proportions can be calculated for each type of miscoding leading to measures that could be used to evaluate the quality and timeliness of the NAICS. The QAS also measures the rate of dead units that were deemed live on the BR. This proves to be useful when trying to predict the effective sampling size.

5.1 Plans for the New Business Register

We would like to standardize procedures and ensure that all steps are subject to reports assessing what is being changed or updated in terms of counts and magnitude. This should be reported systematically and be kept in a database for longitudinal analysis. Any updates to the size measures (revenue, number of employees, etc.) should also be analyzed and extreme cases should be looked at carefully, especially when the updates come from an administrative source which could contain uncorrected values. Again there should be regular reports assessing changes in measure of size which could be analyzed over time.

In addition, we want to have more formal quality control put in place for crucial BR operations such as NAICS coding and profiling. The NAICS coding needs to be accurate and one way of ensuring its continuous accuracy is through quality control. This is valid both for the manual coding (assessing coder's work) and automated coding (assessing automatic coding quality). We want to put in place software that BR operation managers will be able to modify depending on the experience of the coder or any external constraints. For automated coding, this will be extremely important to verify new automatic coding put in place as well as sectors where the coding might generate confusion. For example, automatically coding "food bank" into the banking industry should not happen and verification should be put in place in order to prevent this kind of error.

In order to do manual coding there are a few tools available at Statistics Canada but this is still seen as a subjective exercise. There is however one new tool developed in collaboration with the Canada Revenue Agency. It uses the description of the activity to prompt a set of questions that will help selecting the appropriate NAICS code. This tool is currently available for roughly 80% of NAICS codes and it is planned that it will be refined and expanded to cover more NAICS industries. It should be used by all coders in order to standardize procedure and yield consistent results.

At Statistics Canada, there is an exercise called profiling (described in section 2.1) which needs to be done for large companies. The profiling exercise is conducted at regular intervals and is performed to update the Operating Structure of the business. Depending on the importance of the business, profiling may be done more regularly. One important aspect of profiling is determining the accountability of each portion of the business. This can be done in as many ways as there are profilers. Under the new BR we want to give training and find ways of standardizing it. Also we might want to consider including some quality control process in the profiling exercise in order to assess its quality. This will probably be more difficult to implement than a regular quality control process as profiling is very long and complex.

Improving the quality of the Business Register through standardization, training, addition of quality control mechanisms, timely updates or profiling will reflect on all surveys using the BR. Reporting that quality on a regular basis and following changes to the Business Register through time will help keep BR managers and users informed. This information will be useful to make decisions regarding the BR maintenance or improvement; it can also be useful for new surveys that are contemplating the BR as their frame. The redesign of the BR provides an excellent opportunity to add, formalize and report data quality for the years to come.

6. FUTURE CHALLENGES

Changing the sampling element for establishment surveys may have an impact at different levels of the survey process from sampling to estimation. We need to assess changes that will occur based on our new definition of the sampling elements, especially with regards to industry and provincial estimates. This major change may also affect survey operations and programs. Each survey will have to individually evaluate the impact of this new approach on their system.

While conducting the BR Redesign, one goal is to have more timely updates and optimize our use of administrative data. But are we making the maximum possible use of these inexpensive data sources? Now that these sources have been in place for a certain period of time, we might use them for additional operations like structure changes or NAICS verification. This is one of the possible ways of improving the quality of the new BR. We also outlined a number of other possibilities to report or improve the data quality, but are they enough? The only certainty is that the more we know our frame, the more confident we can be in our estimates.

REFERENCES

Bérard, H., Gagné, P., Rancourt, E. and Pursey, S. (2005), "La Refonte Du Registre des Entreprises de Statistique Canada", Proceedings of the Colloque francophone sur les sondages, Québec, Canada

Colledge, M. J. (1987), "The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada", Proceedings of the Third Annual Research Conference, Bureau of the Census, 550-576.

Cuthill, I. (1990, revised 1997), "The Statistics Canada Business Register", internal document, Informatics Branch, Statistics Canada

Gagné, P. (2004), "Projet de refonte du registre des entreprises", internal document, Statistics Canada.

Pursey, S., Beaucage, Y. and Hunsberger, P. (2005), "The Redesign of the Statistics Canada's Business Register" 2005 Proceedings of the American Statistical Association (to appear)

Rancourt, E., Bérard, H. and Pursey, S. (2005), "Re-Thinking Statistics Canada's Business Register"; Federal Committee on Statistical Methodology Research Conference, Arlington,VA. On CD-ROM.