A Canadian perspective on the use of administrative information for statistical purposes

Giles, Philip
Statistics Canada
100 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6, Canada
E-mail: Philip.giles@statcan.gc.ca

1. History and current practice

As is the case with most countries, the origins of the national statistics system in Canada are rooted in the use of administrative records. Apart from the Census of Population and the Census of Agriculture, early statistical information in Canada was produced solely from administrative information, such as Vital Statistics (births and deaths), Public Accounts and limited information on business activity. As this information was collected on paper only, the processing and development of statistical information required significant human resources, thus limiting the production capabilities.

While these uses of administrative information continued, with the development of formal scientific sampling techniques in the 20th century, direct collection of information for solely statistical purposes expanded. Surveys of both households and businesses were introduced, most of which collected information on a regular basis, usually monthly or annually.

During the 1960s, organizations that collected administrative information began to automate their processes. Electronic administrative information permitted significantly greater application for statistical purposes, and Statistics Canada began to exploit this potential.

Response rates for direct survey collection began to decline in the late 1980s and difficulties in collecting information continue today. Statistical organizations have identified various reasons for this, and they have incorporated measures to counteract declining response rates. Reductions in funds available for direct collection have also occurred, leading statistical organizations to consider more cost-effective approaches to meet information demands. Although these different approaches were not entirely new, the extent of their usage was unprecedented.

Among these changes were the use of administrative information, and the linkage of multiple information sources, including the linkage of administrative information and information collected directly through a survey or census. These approaches raise legal, ethical and privacy issues. The objective of this paper is to describe Statistics Canada's approach to the issues, and to indicate its plans for future work in this area.

2. Current uses of administrative information at Statistics Canada

There exist very few statistical programs at Statistics Canada that do not currently use administrative information to some extent. In many cases, the program would not exist were it not for administrative data. In other cases, the quality of the statistical information has improved as a result of the use of administrative information.

Administrative data are used in many aspects of the statistical production process:

- Replacement of direct data collection
- Use as survey frames or to supplement frames
- Edit and imputation
- Estimation - benchmarking, calibration, calendarisation
- Data quality evaluation, including data confrontation
- Enhancing survey data thorough record linkages

One particular source that has greatly impacted Statistics Canada's statistical programs is the federal department of revenue – the Canada Revenue Agency (CRA) – from which we obtain regular information related to the income tax system (for both individuals and businesses), the excise tax system, and information related to payroll deductions.

Statistics Canada's Business Register serves an essential purpose to the production and analysis of economic statistics in Canada. This "list" of all businesses in the country has been significantly improved over time through expanded use of administrative information, and improved quality of this information.

To reduce response burden and to improve the quality of the economic statistical information, tax information is used extensively in place of direct data collection, particularly for small and medium sized businesses. Such businesses are large in number, but account for a relatively small proportion of total production. Statistics Canada uses administrative information to the extent possible for businesses. However, a very large proportion of our information needs (such as all detailed business input and output) are not available from administrative records. Our approach is to collect directly from larger businesses, and to impute the information for smaller businesses (using both administrative records and the directly-collected information). This approach has been demonstrated to reduce costs, reduce response burden, and has a negligible impact on data quality, so long as the sub-population that is subject to imputation accounts for a relatively small share of the economic sector of interest.

On the social-statistics side, linkage of information submitted on income tax returns has opened many doors for analysis. The use of administrative tax information is particularly important because of its broad analytical usefulness as well as the fact that this information is difficult to collect through a survey.

Arguably, the most important trend in the social domain during the last 25 years has been the vast expansion of social surveys. What users wanted to know were answers to complex questions involving the ability to relate events longitudinally, and to measure outcomes, not just outputs. For example, traditional statistics reported student enrolment numbers in post-secondary institutions. They could not report on the characteristics of those who did and did not attend post-secondary institutions, those who completed and dropped out, and the long-term levels of "success" achieved by the various sub-populations. Typically, outcomes cannot be measured using administrative records only.

There are five general areas where we have been making expanded use of administrative records in the social area in novel ways:

- Longitudinal personal income tax data manipulated to resemble longitudinal family income surveys: Annually, the universe of individual tax filers is grouped into families. Based on other administrative files, children and other dependents can be added. This permits the derivation of family income, which facilitates the development of a longitudinal file.
- Linking other administrative records with income tax records to trace income outcomes over time for different target groups. For example, several studies have examined the economic impact of recent immigrants through the merging of immigration records with income tax files.

- Linking individual administrative records with survey data to exploit the relative strengths of both. The linking of provincial health records with our ongoing household health survey has greatly expanded the analytical potential of the household health database.
- Using administrative records to facilitate survey operations. In some cases, such as for longitudinal surveys, it is necessary to trace people. Address and telephone information is obtained through administrative records, such as telephone billing files and driver's license files. To facilitate census operations and survey operations and to reduce collection costs, Statistics Canada maintains an Address Register of private dwellings.
- Reducing reporting burden by using administrative information in lieu of direct survey collection. The primary current use is with personal income information.

Apart from the use of tax data, Statistics Canada's use of administrative data for social statistics remains in areas where they represent the primary source of information: health, justice (police and court files) and education (institutional and student-level information). Although the quality of the statistical information has improved over time as a result of representation by Statistics Canada to the collectors of the administrative information, there are currently only infrequent uses that could not have been possible, say 25 or 30 years ago.

Other examples of other current administrative information holdings are:

- Import and export declarations
- Transportation administrative files (airports, marine and rail travel)
- Information from various agricultural marketing boards
- Employment Insurance claimants
- Building permits

In general terms, the use of administrative information reduces response burden and data collection costs. In many cases, the data quality is higher for administrative information as compared to directly-collected.

3. Legal, ethical, privacy issues

Legal authority for National Statistics Office to obtain administrative information

Through the federal *Statistics Act*, Statistics Canada has the legal authority to obtain and use administrative data files held by any organization, public or private, in Canada. This information assists the Agency in meeting its mandate. As not all organizations accept that Statistics Canada has the legal authority to compel them to provide copies of administrative files, Statistics Canada has chosen a non-confrontational approach: it negotiates with other organizations on mutually-acceptable terms for Statistics Canada to obtain and use administrative information.

Statistical use is secondary to primary purpose

By definition, the statistical use of administrative information is secondary to the primary purpose for which it is collected. Usually these primary purposes are for program administration related to individuals (eligibility for government transfers or for benefits, registration such as for a permit to drive) or records created for business purposes such as police records and student records kept by educational institutions. In many cases, these organizations are limited in their ability to collect information; generally, the information must be needed for the business purposes of the organization.

Legislation affecting statistical information

Statistics Canada strives to make representation with respect to legislation on all issues related to its operations. Principally, this review attempts to ensure the preservation of the confidentiality of the information. For example, our assurance of confidentiality is extremely important when we collect information. We would not want legislation that would, even inadvertently, compromise the confidentiality of our statistical information holdings by authorizing another organization to have access to all government records.

As another high priority, review of draft legislation attempts to ensure continuity of access or increased legal authority for access to administrative files.

Control over one's own information

This is the basic definition of information privacy. Although legally permissible, the use of administrative information for statistical purposes does increase the level of privacy invasion. Canada's federal *Privacy Act* clearly distinguishes between the use of personal information for administrative and statistical purposes; generally, the constraints are less stringent for statistical uses. In many situations, it is impractical to seek consent. The privacy invasion is mitigated by the confidentiality requirements of the *Statistics Act* in two ways. First, the information itself must be protected by the NSO so that it is not accessible to those who do not have the legal authority. Also, all published statistical information must be in a form that does not divulge, even indirectly, information related to an individual person, business or organization.

Record Linkage

Record linkage (also called statistical matching) is the matching of individual micro-records for the purpose of producing a composite record. Record linkage is a valuable statistical tool with many benefits. It is more privacy invasive than the use of administrative files individually as it puts information together in a new manner of which individuals are generally unaware. This is mitigated by the confidentiality requirements of the *Statistics Act* and Statistics Canada's own privacy and security policies. Statistics Canada has a strong governance model for record linkage that ensures that there is a strong business justification for the record linkage. Although consent for linkage is usually sought only when there is a legal or policy requirement, some transparency is achieved through proactive disclosure on the Agency's web site, as descriptions of all approved record linkage projects are posted.

It is important to understand linkage for statistical purposes. Although the objective is to match the same individual across the various files, a certain level of error will generally not impact the output of any statistical analysis using the linked file. On the other hand, such a file could not easily be used for administrative purposes, since an incorrect decision would be made for individuals who have been incorrectly linked.

To achieve an appropriate balance between the benefits to its statistical programs and the privacy invasiveness, Statistics Canada sets the following conditions for record linkage activities:

- the purpose of the record linkage activity is statistical and is consistent with the mandate of Statistics Canada as described in the *Statistics Act*;
- the benefits to be derived from such a linkage are clearly in the public interest;
- the results of the record linkage activity will not be used for purposes that can be detrimental to the individuals involved;
- in the case of research projects, the proposed analysis is methodologically sound and the linked dataset and analytical techniques are appropriate for the intended objectives;
- the products of the record linkage activity will be released only in accordance with the confidentiality provisions of the *Statistics Act* and with any applicable requirements of the *Privacy Act*;
- the record linkage activity has demonstrable cost or respondent burden savings over other alternatives, or is the only feasible option;
- the record linkage activity is judged not to jeopardize the future conduct of Statistics Canada's programs;
- the linkage satisfies a rigorous review and approval process, culminating with the approval by the Chief Statistician.

4. Challenges and approaches

Negotiating access

It is increasingly difficult to negotiate access to administrative information held by other organizations. In the past, organizations would usually sign Statistics Canada's standard agreement template. Now, they are increasingly challenging not only the terms of the agreement, but also the premise that Statistics Canada has the legal authority to obtain the information. This prolongs the negotiations, and in some important areas, has so far prevented access.

One important area where access has been difficult and continues to be so is in the area of provincial individual health records, clearly extremely-sensitive information. Currently, along with representatives of all provincial and territorial departments of health, Statistics Canada is engaging in an initiative to address the privacy and security concerns while providing access to compatible national health information. A rigorous set of procedures is being negotiated to address transmission of the information, storage and retention by Statistics Canada, and researcher access.

For many years, Statistics Canada has had a small group of employees dedicated to the development of statistical agreements. Although statisticians by training (rather than lawyers), this group uses its knowledge of the statistical needs and its understanding of the legal requirements of the *Statistics Act* to negotiate agreements with other organizations, involving the Agency's legal advisors as needed.

Influencing the content of administrative files

Often Statistics Canada finds that a few extra pieces of information on an administrative data file would greatly increase the statistical uses. In several situations, we have been able to convince the collecting organization to add content that they do not require, but which is extremely valuable to us. The limitation always is whether the organization has the legal authority to collect information that they do not require.

Use of common identifiers

The quality of Statistics Canada's economic statistics increased significantly with the introduction of the Business Number by the federal government. Intended to simplify the way businesses dealt with government, the use of this identifier became mandatory in 1997 for all transactions with government.

Unlike several other countries, Canada does not require a resident to register to a population register. Several personal identifiers exist, each for different programs. Legislation and public sentiment limit the use of these identifiers. This makes record linkage much more difficult since the matching of records must be accomplished using information that is more difficult to use for matching and which has errors. An approach that Statistics Canada uses when different databases may be linked for different purposes is to adopt an approach of linkable, but not linked files.

An example will illustrate what is meant by linkable files. Assume three data files where information on the same individual may occur on one or more of the three files. The three files have three different identifiers; call them ID1, ID2, ID3. There are other identifiers in common on the various files that are more difficult to link; for example, name, address, date of birth, sex. A linkage activity takes place by which a random identifier, for purpose of this example, called RID, is created for every individual with information on at least one of the files. Then three different id link files are created, each containing only two pieces of information. The first id link file contains the link between RID and ID1 for all individuals on the first file. Similarly, the second file contains the link between RID and ID2, and the third file the link between RID and ID3. These three id link files are retained and protected so that only those responsible for linkage can access them.

The use of these files balances privacy considerations with the efficiency of future projects that require these files to be linked. More specifically, as described above in the section on record linkage, it is necessary to have a strong governance over record linkage projects. Linkage is a violation of privacy that can only be justified if no harm comes to individuals concerned, either individually or collectively, ***and*** if the social benefits are sufficiently important to justify the violation involved. The latter condition can only be assessed on the basis of a case by case examination. Hence each use of the linkable files must be approved by the Chief Statistician.

Census of Population

Statistics Canada conducts the Census of Population every 5 years. Since 1971, there are two census questionnaires: the short form for 80% of households and the long form for 20% of households.[1] Traditionally, census information has been obtained through direct collection from individuals. For the 2006 and 2011 censuses, individuals have been offered the choice with respect to the income information. One could provide the information directly in the traditional manner or request that Statistics Canada use the administrative information already submitted for income tax purposes. In 2006, 82 % of individuals chose the administrative option.

For the 2016 Census of Population and beyond, the federal government has asked Statistics Canada to evaluate collection options. One option which Statistics Canada will be considering is to make greater use of administrative information. Among these considerations are: the potential sources for comprehensive reliable information, the accessibility, and the level of error associated with the file linkages. A possibly more important aspect of this option will be whether the general public accepts this expanded use of administrative information.

[1] Technically, the 2011 Census consists of only the short form. The long form content was collected through a voluntary survey, with one in three households selected into the sample for the National Household Survey.

Privacy Commissioners

Privacy Commissioners in Canada recognize and support Statistics Canada's use of administrative data. A major contributor to this is the governance model that Statistics Canada has put into place with respect to privacy protection.

Central control of administrative information

Recognizing the importance of the tax data that it receives from the Canada Revenue Agency, Statistics Canada has set up a central organizational unit to be the liaison with CRA. This central unit has the following responsibilities:

- Ensuring compliance with the terms and conditions of the inter-departmental agreement on the provision of administrative tax data to Statistics Canada
- Controls access to the tax data within Statistics Canada
- Acts as the "subject matter" experts with respect to the file content
- As the statistical use is somewhat different from the program purposes of CRA, conducts a "clean up" of all files received to ensure that they are fit for their intended statistical uses

This approach is the model at Statistics Canada for dealing with administrative information. One area is responsible on behalf of the department.

Use of information on the Internet

An emerging data source is information available on the Internet. This is somewhere between direct collection and administrative information. Usually one can only obtain information on a particular individual (usually a business, not a person). However there is no increased response burden on the part of the individual for Statistics Canada to collect the information. While such an approach is available to specific situations only, it is appealing for many reasons.

One example of its use is to collect price information for the purpose of the production of the Consumer Price Index. For certain products, all the information required to obtain a price is available on business web sites, so this is a cost-effective manner to collect information. A limitation, which to date has not posed a significant concern, is that many web sites restrict the use of their information to actual and potential customers. Statistics Canada seeks permission from the businesses to collect information for its statistical purposes, and has been granted this permission.

5. Summary

Not unlike many national statistical organizations, Statistics Canada is increasingly looking towards administrative information to meet its information requirements. Major shifts such as these inevitably lead to new challenges. Statistics Canada is working with a variety of partners to meet these challenges: data users, data providers, and other statistical agencies.

6. Bibliography

Brackstone, G.J. (1988). Statistical Uses of Administrative Data: Issues and Challenges. Proceedings: Symposium 1987, Statistical Uses of Administrative Data: An International Symposium, Statistics Canada, Ottawa.

Brackstone, G.J., and White, P. (2003). Data Stewardship at Statistics Canada. Proceedings of the Section on Survey Research Methods, American Statistical Association, 2002.

Cloutier, M. (2010). A Strategic Vision for the Use of Administrative Data at Statistics Canada. Proceedings: Symposium 2010, Social Statistics: The Interplay among Censuses, Surveys and Administrative Data, Statistics Canada, Ottawa.

St-Louis, G. (2008). The Evolution of Administrative Data Use for the Canadian Business Register (BR). Proceedings: IAOS (International Association for Official Statistics) 2008 Conference, Shanghai, October 14-16, 2008.

Abstract

ISI Dublin 2011: Session IPS34 "Legal, ethical, privacy, etc. issues arising out of increasing use of linked admin files"

Abstract: A Canadian perspective on the use of administrative information for statistical purposes

Author: Philip Giles, Statistics Canada

Statistics Canada has a long history of use of administrative records for statistical purposes. For example, for several decades, vital statistics records have been obtained from the provinces and territories. Other areas where long-standing use has been made of administrative records are agriculture, international trade and transportation.

More recently, since the mid 1990s, Statistics Canada has had access to detailed income tax information (for both persons and businesses). This has expanded the quality and quantity of the statistical information that Statistics Canada produces for its data user community. Just as important is the use of this information to improve the efficiency and quality of our statistical operations. For example, the use of tax data has greatly improved the quality of the Business Register, the basis for all economic statistics.

Administrative information is also used for many other purposes, such as benchmarking to key variables or for data quality evaluation.

The paper will explore current uses of administrative information at Statistics Canada, discuss issues related to the use of administrative information and consider possible future avenues for new uses.