Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Towards a Register-centric Statistical System: Recent Developments at Statistics Canada

Philippe Gagné, Jean Pignal, Tanvir Quadir, and Christian Wolfe [1]

Abstract

The Statistical Registers Integration Project (SRIP) aims to build and maintain a Statistical Register Infrastructure (SRI) that comprises four interconnected base registers (population, building, business and activity) built entirely through the integration of administrative data from various sources that will be used for statistical purposes. The goals of the SRI are to maximize the use of actionable information from data already collected, improve the timeliness of official statistics, and reduce response burden, all while protecting the privacy of Canadians. A brief account of the ongoing development of a de-identified population register and a building register, the redesign of the existing business register, and a conceptual framework of an activity register is provided. This paper also outlines the Canadian System of Integrated Statistical Registers (CSISR), which assembles the underlying base registers and lays the foundation for the SRI.

Key Words and Phrases: Statistical register; Administrative data; De-identification; Statistical Registers Infrastructure.

1. Introduction

As with a number of developed nations, the increased use of administrative data is a central component of many of the modernization strategies for Statistics Canada[2]. Against that backdrop, the Statistical Registers Integration Project (SRIP) aims to maximise information from administrative data that has been repurposed for statistical use. This, in turn, reduces response burden and the cost of survey collection, and enables users to access timely, relevant and high quality statistical information in a central, integrated, de-identified data holding. Hence, Statistics Canada is modernizing the infrastructure that will interconnect a series of base statistical registers (population, building, business and activity) which will be used as a coherent platform by all of the agency's statistical programs. This includes the development of a Statistical Population Register (SPR), the transformation of the existing residential Address Register (AR) into a more comprehensive Statistical Building Register (SBgR), and the addition of new components to its Statistical Business Register (SBR). In addition, as part of the Social Statistics modernization initiatives, a proof-of-concept version of a Statistical Activity Register (SAR), the Life-course Analytical File (LAF) is currently under development.

2. Base Statistical Registers

2.1 Statistical Business Register

Statistics Canada's SBR was first constructed in the 1980's, and has been continuously improved over time. In compliance with international guidelines[3], the SBR collects, compiles and maintains a full repository of businesses

[1] Jean Pignal, Data Integration Infrastructure Division, Statistics Canada, Canada, K1A 0T6 (jean.pignal@canada.ca); Philippe Gagné, Data Integration Infrastructure Division, Statistics Canada, Canada, K1A 0T6 (philippe.gagne5@canada.ca); Christian Wolfe, Data Integration Infrastructure Division, Statistics Canada, Canada, K1A 0T6 (christian.wolfe@canada.ca); Tanvir Quadir, Special Surveys Division, Statistics Canada, Canada, K1A 0T6 (tanvir.quadir@canada.ca).

[2] Other countries currently investigating the increased use of administrative data for statistical purposes include, Australia, New Zealand, the United Kingdom and the United States of America.

[3] UNECE. (2015). Guidelines on Statistical Business Registers. Retrieved from https://www.unece.org/fileadmin/DAM/stats/publications/2015/ECE_CES_39_WEB.pdf

(legal and statistical units) in Canada, to provide a complete, unduplicated, and up-to-date frame for economic programs. It also plays a central role in the modernization of the agency's approach to producing economic statistics.

The SBR was first developed as a frame for survey sampling, and is currently used by more than 200 economic surveys. Its main role is to support survey collection activities by managing and updating business and respondent information, and monitoring and controlling respondents' response burden. The SBR's target coverage includes all businesses and institutions, over 6.6 million active entities, engaged in producing goods or services in Canada, including agriculture businesses/farms.

As with the proposed SBgR and SPR, the SBR is evergreen as it is continuously updated with multiple sources of information and mostly based on administrative data. A very large portion of the SBR is maintained solely with the usage of weekly, monthly, quarterly and annual administrative data files from the Canada Revenue Agency. Ongoing business profiling activities are essential to building comprehensive business structures. Finally, survey feedback complements the administrative data and profiling activities.

As to the attributes, the SBR contains direct identifiers such as legal and operating names and addresses of economic units. All entities on the SBR are assigned a unique business statistical identifier (BR_SN) to avoid duplication and allow for the de-identification of entities. The register also contains statistical variables such as the "birth" (date of incorporation or official recognition as an economic operator) and cessation dates, the economic unit activity status, size variables (for example, sales or number of employees) and classification variables such as NAICS and CCIUS[4], among others. All values are time referenced. Communication variables are also present, including the contact, mailing address, landline and cell phone numbers, as well as email and website addresses of the entities.

With the increased availability and use of administrative data files, the SBR is becoming a central hub for business data linkages, a direct input to the analysis of business populations and for the dissemination of official business statistics. The SBR is leveraging the development of the register-centric statistical system by adding the relationship with units in other base statistical registers – for example the location of the SBR entities will soon be recorded using the statistical building and building unit identifiers derived from the SBgR. Furthermore, in order to better support analytical requirements, the SBR is creating a new longitudinal module. It will include longitudinal identifiers, predecessor/successor tables with detailed information, as well as historical updates of the past data to reflect the most recent information available. The new features to the SBR with linkages to the other base statistical registers will provide the infrastructure with a more rapid response to the evolving needs for detailed longitudinal business data and facilitate the production of business statistics and entrepreneurship indicators, along with new and innovative statistics.

2.2 Statistical Building Register

The address register (AR), initially built in 1986, has been used as the basis for our Census collections and as a frame for household and social surveys. But the register's role appears not to be as pivotal as the SBR. For example, the AR does not record the units sampled to monitor respondent burden, and it has little or no direct interaction regarding collection activities and survey process. Hence, as part of the transformation initiatives, the AR will be transformed into a more exhaustive statistical building register (SBgR) that covers both residential and non-residential buildings, where the entity is the building and its units. *Building* refers to a roofed independent free-standing permanent structure usually enclosed within external walls or dividing walls that extend from the foundations to the roof and comprises one or more rooms or other space, while *building unit* is a part of the building, either residential or non-residential, that must have its own entrance through an outer door or an interior door in a shared hallway. A unique and stable statistical identifier will be assigned to each building (Bg_SN) and building unit (BU_SN), which will be used to link to base registers, any other statistical registers or administrative files in the agency and to safeguard privacy.

[4] NAICS is the North American Industry Classification System and CCIUS is the Canadian Classification of Institutional Units and Sectors.

The SBgR will be updated on an evergreen fashion and interlinked with the SPR and the SBR. The SBgR will be supported by a Statistical Geospatial Infrastructure (SGI[5]) to improve geocoding of administrative files and optimize linkage of administrative geolocation data (e.g. addresses) to the SBgR and indirectly to the SPR and the SBR. Some of the administrative files that are currently being analyzed for inclusion into the SBgR are Canada Post Corporation files (already used by the AR), provincial 911 emergency files, property assessment and land register files and smart meter files from provincial hydroelectric companies.

The attributes of the SBgR fall into the following four categories:

- Geography (Basic Block, Block Face, GPS coordinates, dissemination and collection geographies)
- Contact information (civic address, mailing address, phone number) to be used for collection and linkage activities
- Status (inactive, active, temporary, placeholder)
- Quality indicators (validity, residential occupancy, geographic precision, etc.)

2.3 Statistical Population Register

The Statistical Population Register (SPR) is an evergreen unduplicated and highly protected database of all people found on a variety of administrative files starting in 2016 and moving forward. Although administrative files have long been used at the agency, the idea of a longitudinal statistical population register evolves from the use of administrative data in the Census of Population Program to replace the collection of income information. The fact that the coverage of the Canadian population from the integration of existing administrative files has been shown to be quite extensive has encouraged the Census to undertake a research project known as the Canadian Statistical Demographic Database (CSDD) to measure our ability to replicate a headcount of Canadians by putting the right people in the right place solely through the use of administrative data. In all, five versions of the CSDD have been created since 2013, two representing the 2011 Census and 3 representing the 2016 Census. The final iteration of this database will be used to initialize the SPR with new entrant files continuously bringing the register up to date by adding new records and by adding new attribute values to existing records. The contributing administrative files for the CSDD (and the future SPR) include a variety of tax files from the Canada Revenue Agency (CRA), births and deaths from vital statistics, temporary and permanent residents files from Immigration, Refugees and Citizenship Canada (IRCC), the Indian Register from Crown-Indigenous Relations and Northern Affairs Canada (CIRNAC) and the Social Insurance Number (SIN) register from Employment and Social Development Canada (ESDC). In addition to the aforementioned CSDD files, it is envisioned that, once the SPR has reached maturity, all incoming administrative files containing individuals could be filtered through the SPR. Some, such as driver's license and cell phone service files, could add records and attributes to the SPR, while others may simply be de-identified and assigned a unique Individual Statistical Number (ISN).

The goal of the ISN is to alleviate privacy concerns by removing personal identifiers (names, administrative numbers such as the SIN, etc.) from all incoming administrative files and substituting these with a permanent random number that would allow the agency to continue to use the information for statistical purposes by allowing record linkages, while masking the identity of individuals on the files. The ultimate goal is to maintain public trust while Statistics Canada acquires administrative data sources that can create more timely and relevant statistics.

The demographic attributes included on the SPR are quite basic: sex, immigration status, Indigenous status, date of birth, date of death, income (for stratification purposes), marital status, immigration date and emigration date. Contact information includes name, address (replaced by the BU_SN from the SBgR), cell phone number and email address and language of correspondence (for potential future contact). These are primarily divided into static attributes, such as date of birth, that do not usually change and dynamic attributes such as last name, that may have variability over time. The SPR maintains the stack of values for each attribute that originate from disparate sources and has developed

[5] , the SGI is a subset of the Spatial Data Infrastructure (SDI) for these key concepts: Basic Blocks (BB), Streets (ST), Blockfaces (BF) and Census Subdivisions (CS).More details regarding the SDI are indicated in the Reference section.

business rules for the most-likely value of each attribute to be used by statistical programs. To this end, the SPR has implemented the following quality indicators:

- ISN status/Person validity: It is an indicator that shows the degree of confidence we have that a person is indeed a unique statistical entity in the SPR. The ISN can be designated as Temporary ("T"), Permanent ("P"), or Expired ("X") depending on the signals received from the underlying administrative files.
- Valid link/Validity indicator: This validity indicator is a binary (0, 1) or (T, F) value at the identifier level. This will be used to manage false-positive matches when information is erroneously assigned to the wrong record (ISN).
- Attribute quality: Calculated or assigned value based on the trust we have in the source for a specific attribute.
- In and out of scope indicator: This value is calculated at extraction to denote if, given the extraction parameters (date ranges, desired age groups, proximity of last recorded signal, indication of emigration, date of death, etc.) the record is in-scope or out-of-scope.
- Entropy: This is a measure of discrepancy or disparateness among all sources recorded for a record's attributes. The closer the value of the entropy is to zero, the more agreement is deemed between the sources as to the value of the attribute. The weighted entropy value is dynamically calculated when an extraction from the SPR is performed. For dynamic attributes, the number of occurrences of a value as well as its temporal proximity is also considered.

2.4 Statistical Activity Register

The notion of a Statistical Activity Register (SAR) appears in UNECE (2013)[6] model of Nordic Register-based Statistical Systems and Wallgren & Wallgren (2014). The idea of a statistical register called Social Data Register (SDR), which is deemed as a potential contributor for Statistical Activity Register, was first conceived at the agency's strategic plan of Social Statistics Program modernization initiatives (MacNabb, 2017). The pilot version of the SDR, called the Life-course Analytical File (LAF), is currently underway, which is a register of person-level activity information of Canadian population in the domains of Work, Education and Human Capital, Economic Well-being, Family, Crime and Victimization, Health and Well-being, Immigration and Indigenous Status. Cooke (2011) portrayed the integration of a life-course perspective within the Canadian social data system in a strategy paper produced for Employment and Social Development Canada (ESDC) and Statistics Canada. This seminal reference lays the foundation for the LAF that will serve the following:

- To potentially contribute to the creation of a statistical activity register, which will form the fourth base statistical register in the Canadian System of Integrated Statistical Register (CSISR) along with the statistical population, building and business registers;
- To facilitate the integration of underlying sources with all due diligence regarding security and confidentiality of the data by building a de-identified statistical register that contains information in the aforementioned domains, which are not maintained by other base registers (population, building, and business);
- To foster interdisciplinary statistical research by providing an 'analysis-ready' file containing key cross-sector variables and indicators required by most social statistics programs;
- To record and maintain the historical activities of the Canadian population that will enable analyzing life-course perspective, developmental trajectories, and event histories;
- To contribute to improved survey support (frame service, content replacement);
- The coverage of the LAF is from 2005 to 2015. The LAF is using the Social Data Linkage Environment (SDLE) to link records across the domains and record pertinent activities. The list of the administrative sources used for building the LAF includes by domain and topic: Work: Tax and employment insurance benefit files; Education and Human Capital: Post-secondary institution and apprentice Records; Economic Well-being: Tax file (detailed economic performance), Family: Tax file (family-level information); Crime and Victimization: Criminal court data, Health and Well-being: Vital statistics, cancer register and

[6] UNECE (2013). Using Administrative Data in Statistical Registers. Retrieved from https://statswiki.unece.org/display/adso/Using+Administrative+Data+in+Statistical+Registers

hospitalization data; and Immigration and Indigenous Status: Permanent and non-permanent resident files and Indian Register.

The coverage of the LAF is from 2005 to 2015. The LAF is using the Social Data Linkage Environment (SDLE) to securely assign activities to anonymized individuals across domains. The domains will be treated separately and will not be combined into a database. A broad category of the attributes of LAF is given below, where some of the attributes may overlap with SPR which will not be maintained by LAF in the long run to avoid inconsistencies:

- Work: Employment status, employment income, employment insurance benefits
- Education and Human Capital: Enrolment status, program/certification type and duration, institution, educational attainment
- Family: Family composition, (census) family income, household type, household income (if household indicator is available from address register)
- Health and Well-being: Hospitalization, use of health services, chronic condition, birth and death information
- Economic Well-being: Detailed economic performance (employment income – wages, self-employment income, investment income, private retirement income, other market income, government transfers)
- Crime and Victimization: Types and dates of offenses and charges, incarceration, recidivism
- Immigration and Indigenous Status: Characteristics of immigrants, aboriginal indicator

3. Statistical Registers Infrastructure

The mandate for Statistical Registers Infrastructure (SRI) is to create a framework of integrated base statistical registers to be used as a coherent foundation for statistical programs and to deliver processes for register linkage to support frame reconciliation, integration between registers, geocoding and de-identification of administrative files.

A custodian division at Statistics Canada is responsible for the acquisition and receipt of an administrative data file (ADF), pre-processing it, determining fields with identifiers/quasi-identifiers, preparing metadata, storing the ADF and controlling accesses to it. The ADFs processed in the custodian division will undergo the spatial geocoding if the data file contains non-tabular geographic attributes or is a spatial file. The pre-transformation phase of data includes any extra processing that may be required to transform the data into the formats expected by the Register Matching Engine (RME). The on-boarding of an ADF consists of defining the layout of information in the ADF, associating the variables in the ADF to register fields and quality evaluation of the ADF. The RME provides the SRI with tools to perform linkage of data to/within the statistical register system. The RME will have the following fundamental elements:

- Assigning a unique statistical number to each new statistical unit (person, building, building unit, business or activity);
- De-identifying units by concealing the raw personal identifiers from the register and storing them in a secure, controlled environment;
- Developing business rules to determine with reasonable certainty whether a new record on an administrative source truly involves a new unit;
- Placing individuals in their dwellings. There will be continuous queries from the SPR to the SBgR to obtain the unique identifier. This requires a solid geospatial database that contains information on the road network, address ranges and all standard geographic boarders that are used at Statistics Canada.

4. Benefits and Uses of a Statistical Register System

While the design and construction of the new Statistical Population Register and the Statistical Building Register are being developed first to support and supplement current census collection efforts, there are other potential statistical uses for such registers within our National Statistical Organisation. For instance, the SPR could be used to supplement our existing Demographic Estimates Program. In addition, where quality is sufficient, it could provide, for the first time, a sampling frame of individuals for the Social Statistics Programs giving us the ability to construct more efficient sampling design. This would, complement the dwelling frame provided by the SBgR. Furthermore, it could produce regular baselines for population studies (e.g. snapshots of the entire population at a specific time, or of a population subset (youth, Indigenous Canadians, new immigrants), or for a specified geography).

The term ‘integration’ in the CSISR should be perceived in terms of the interactivity of the base statistical registers within themselves and with non-base statistical registers through statistical identifiers, while the base statistical registers contain very basic information about the statistical units and the non-base or satellite registers contain a number of details about underlying construct/concept of interest (e.g., employment, health, justice, education).

5. Conclusion and Future Directions

The raison d’être of register-centric statistical system is hinged on the ‘Administrative Data First’ model of the modernization strategy of Statistics Canada. The Census of Population is currently exploring means of replacing all, or parts, of the traditional census collection using administrative data. Continued simulation of a combined census (i.e. SRI combined with traditional enumeration) until 2021 will lead to the decision for the format of census in 2026 and beyond. Once the base statistical registers are up and running, it would be imperative to ensure their seamless integration with non-base (i.e. satellite) statistical registers. Guidelines and frameworks need to be developed to produce official statistics from register-centric statistical system or mix of statistical registers and surveys as well as quality indicators to help interpret the results.

References

Bakker, B. F. M. et al. (2014), “The System of Social Statistical Datasets of Statistics Netherlands: An Integral Approach to the Production of Register-based Social Statistics”, *Statistical Journal of the IAOS*, 30, pp. 411-424.

Cooke, M. (2011), “Integrating a Lifecourse Perspective within the Canadian Social Data System”, A study paper prepared for Human Resources and Skills Development Canada (HRSDC) and Statistics Canada, Canada.

Gagné, P., and G. St-Louis (2017), “The Statistical Registers Transformation and Integration Project at a Glance”, a presentation for Advisory Committee, Statistical Registers and Geography Division, Canada: Statistics Canada.

MacNabb, L. (2017), “Social Statistics Transformation Phase 1 Strategic Plan: Social Health and Labour Statistics Field”, an internal report, Social Health and Labour Statistics Field, Canada: Statistics Canada.

Statistics Canada, Definitions. Data sources and methods / Statistical units / Building http://www23.statcan.gc.ca/imdb/p3Var.pl?Function=Unit&Id=450333

Statistics Canada, Definitions. Data sources and methods / Statistical units / Building http://www23.statcan.gc.ca/imdb/p3Var.pl?Function=Unit&Id=450334

Statistics Canada, Illustrated glossary, Spatial Data Infrastructure (SDI) https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/other-autre/infrastructure-eng.htm#n1

Wallgren, A. and Wallgren, B. (2014), *Register-based Statistics: Statistical Methods for Administrative Data*, United Kingdom: Wiley.