Statistical Journal of the IAOS 38 (2022) 231–243
DOI 10.3233/SJI-210923
IOS Press

The Canadian experience of building a privacy-responsible integrated statistical register infrastructure

Julie Trépanier
Statistics Canada, 170, Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6, Canada
Tel.: +1 613 371 5777; E-mail: julie.trepanier@statcan.gc.ca

Abstract. Statistics Canada has maintained statistical business and address registers for decades. Its Statistical Business Register is continuously being modernized to adapt to needs for more and more timely business and institutional statistics at lower levels of geography. The statistical address register is being replaced in 2022 by a Statistical Building Register that expands coverage to the non-residential building units and includes more attributes. Since 2016, Statistics Canada had also been investigating options to add a population component to this integrated system of registers. The organization settled in 2021 on a privacy-responsible population linkage infrastructure that is designed with privacy in mind from the onset. This paper presents how the privacy landscape has evolved and shaped the statistical register infrastructure in Canada. It also describes the Secure Infrastructure for Data Integration that will be elaborated to produce reference population files to support the production of statistical information for Canadians, and how it is entrenched in privacy principles.

Keywords: Administrative data, privacy, statistical registers

1. Introduction

Registers built and maintained for administrative purposes have existed in the world for centuries, well before the digital age. Their use in the production of official statistics (not to be confused with the statistical use of administrative records in general) dates back to the late 1960s or early 1970s, at least in the Nordic countries [1]. Publications from the United Nations Economic Commission for Europe [1,2] and Wallgren and Wallgren [3] are useful references for establishing definitions of administrative data vs. registers, administrative vs. statistical registers, etc. To avoid any confusion while reading this paper, the most useful definitions reviewed and compiled by the author in 2018 with international colleagues from the International Census Forum have been appended to this paper.

Early signs of statistical registers at Statistics Canada date back from 1972. The first statistical register, the Statistical Business Register (SBR), has evolved since the 1980s to become an exhaustive central repository of baseline information on businesses and institutions operating in Canada, with more than 200 survey programs in the Agency utilizing it [4]. The SBR is not only serving as a sampling frame for business and institutional surveys, it is also pivotal to collection activities by providing business contact information and managing response burden centrally.

The SBR Program is constantly evolving and being modernized. In the recent years, the SBR has grown from a strictly industry-based frame to a frame that provides some information on key activities of businesses (e.g., agricultural activities, digital economy). A Longitudinal Business Database created from the SBR will also reach completion in 2022 to better reflect real business closures, openings and changes in industry classification to support the production of business demographics. On the modernization path for the next three years are the improvement of the accuracy and timeliness of geographical and industrial classification of business and institution locations by using more lean and automated business profiling processes and alter-

native data sources, as well as a review of its security framework to reflect the expanded role and increasing number of users of the SBR across the Agency.

Another statistical register that was initiated in the 1990s, first as a coverage improvement tool for the Census of population, is the Address Register (AR) [5]. Over the years, the AR has become the central frame of residential dwellings (both private and collective dwellings) for the Census of population and more than 20 social survey programs at Statistics Canada. It contains civic addresses and related geographic coding, but also some basic dwelling-level socio-economic information to support survey design, as well as mailing addresses and telephone numbers to support data collection.

In 2022, as many Census 2021 activities are wrapping up, the AR will be replaced by the Statistical Building Register (SBgR). The SBgR uses more administrative data sources than the AR (e.g., addresses from Canada Post's non-residential files, 911 emergency files, electricity distribution files, property assessment files). It extends its coverage to both residential and non-residential buildings and building units. It contains additional attributes that the AR does not have such as alternative address formats that exist in Canada and GPS coordinates. To achieve that, Statistics Canada had to modernize its linkage methods and processes of integrating locational data and put in place the Register Matching Engine.

By 2023, the SBgR will also provide a more complete suite of frame services similar to the SBR by going beyond the provision of survey frames and aiming at a full integration with Statistics Canada's collection systems to exchange contact information and survey feedback and monitor response burden on dwellings. As part of the modernization of the SBR, the building and business registers will be better connected by 2024, ensuring in the future that a business or institutional operating location on the SBR is associated to a building unit on the SBgR. This should greatly enhance the capacity of the Agency to produce disaggregated business and institutional statistics.

Since 2016, Statistics Canada had been exploring the possibility of adding a Statistical Population Register (SPR) to its suite of statistical registers. In 2019, recognizing the importance of privacy and confidentiality and the sensitivity of Canadians to privacy issues, Statistics Canada took a different approach by designing a privacy-responsible population linkage infrastructure rather than building a more traditional population register.

This paper presents how the privacy landscape has evolved and shaped the statistical register infrastructure. It also describes the Secure Infrastructure for Data Integration now being elaborated and how it is entrenched in the privacy principles [6] that exist in Canada and elsewhere and respects the Necessity and Proportionality Framework that Statistics Canada adopted in 2019.

2. Statistics Canada turning to an "administrative data first" approach

Statistics Canada is responsible under the *Statistics Act* [7] for conducting the Census of Population every five years. The government (by an order in council) prescribes the questions to be asked in the census as per subsection 21(1) of the Act. In the summer of 2010, the government approved 10 questions to constitute the 2011 Census. This so-called "short form" remained mandatory and was distributed to all households. The government asked Statistics Canada to collect the remaining information proposed to be collected in the 2011 Census (via a mandatory long-form questionnaire) through a new voluntary sample survey named the National Household Survey. At the time, the notion of privacy intrusiveness was brought to the forefront of the public conversation, raising questions as to whether Canadians should be obliged to answer certain questions and whether the information collected by the Census Program was relevant.

This decision made by the Canadian government at the time, which is no doubt historical in nature for the Agency, triggered a study to examine census options for the 2016 Census. The study, known as the 2016 Census Strategy Project, reviewed the approaches for population censuses that existed around the world and evaluated their applicability to the Canadian context [8]. Statistics Canada considered three main types of methodology approaches used internationally at the time: (1) the traditional census approach, which collects characteristics from all individuals and housing units at a specific point in time; (2) the census approach employing existing administrative registers, including a population register and a building/dwelling register to produce basic characteristics on all individuals and housing units at a specific reference point in time; and (3) the census approach employing continuous measurement, where part or all of the characteristics are collected from individuals and housing units on a continuous basis.

The project concluded in 2012 that the only viable option for the 2016 Census was the traditional census

approach [8]. A census approach employing existing administrative registers requires both a population register and a universal personal identification number, neither of which existed in Canada, nor were they likely to exist in the short or medium term. The Privacy Commissioner of Canada at the time was consulted and clearly expressed concerns for the creation of an administrative population register and a universal personal identification number in Canada.[1] Finally, there was simply not enough time to implement a census approach employing continuous measurement by 2016 and to consult census stakeholders and policy makers on this important change of census methodology. Moreover, this approach was unlikely to cost less than the traditional census.

Although the traditional census approach was recommended for 2016, the report added that the most promising alternative to reduce respondent burden and/or improve efficiency and quality in the future was one that would rely more heavily on administrative data. This recommendation was made at a time when Statistics Canada was shifting to an "administrative data first" strategy [9], which was later put into a more theoretical framework [10]. The Agency even stated in its Report on Plans and Priorities 2014/2015 [11] its objective to conduct a comprehensive review of the potential for administrative records and other alternative data sources to replace, complement or supplement the Agency's census and survey programs. This led to exploratory work for the Census Program that revealed that it would be possible to accurately enumerate the Canadian population at high levels of geography by combining administrative data, but to do so for smaller geographies, in particular remote and rural areas, would require more accurate and timely address information than what the sources available at the time (taxation, immigration, vital statistics) offered.

These encouraging results led to the creation of the Census Program Transformation Project (CPTP) in 2016 [12]. The CPTP was launched to further research the possibility of creating integrated statistical building and population registers to support a combined census model. The idea was to explore whether it would be possible to obtain new administrative data to build statistical registers of sufficient quality to be used to reduce the proportion of Canadian households being asked to complete a short form questionnaire (3 in 4 households in 2016), while still collecting detailed census information from a long form questionnaire (1 in 4 households in 2016).

The longer term vision was to further integrate the statistical building and population registers with the statistical business register that has been in place and has constantly improved since the 1980s at Statistics Canada [13]. Considerations were also given to the notion of a statistical activity register. The vision was greatly inspired by Wallgren and Wallgren [3], but with the important difference that Statistics Canada's registers would be statistical and not administrative constructs, meaning that the registers were and would continue to be built entirely through the integration of administrative data from various sources and be used for statistical purposes only. The three (or four) integrated statistical base registers would form the foundation of a register-based statistical system. It could then be expanded further and connected to other non-base statistical registers.

The planning was well underway when privacy concerns were raised with respect to specific administrative data collections and prompted the Agency to develop an innovative Necessity and Proportionality Framework in 2019. The Framework is anchored on the scientific approach in a manner that further enables transparency [14]. It is meant to guide collection of personal information and improve the privacy and security posture [15]. The Agency became more transparent about its collection and use of administrative data on its client facing "Trust Center" on its website, responding by the same token to new legislative requirements of the *Statistics Act* that had been amended in 2017 and asked that the Minister of Innovation, Science and Industry, who is responsible for Statistics Canada, be informed of new mandatory requests for information (subsection 8(3) of the Act) and that such requests be published (subsection 8(2) of the Act). These events, of historic importance for the Agency, had a major impact on statistical registers, with the result that the statistical building register moved to implementation, but the statistical population register idea was amended for the approach described next.

3. Adding a secure infrastructure for population data integration to the statistical register infrastructure

3.1. Driving forces

In 2020, the Canadian Statistics Advisory Council

[1] In Canada, there exists the Social Insurance Number, but its collection, use and disclosure are restricted by the Directive on Social Insurance Number (https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=13342).

(CSAC) tabled its first annual report, followed by a second report in 2021. CSAC was established as part of the suite of amendments to the *Statistics Act* in 2017 to provide advice to the Minister of Innovation, Science and Industry and the Chief Statistician of Canada on the overall quality of the national statistical system. Its mandate also includes the public release of an annual report. Very much influenced by the COVID-19 pandemic situation and the importance for decision-makers to have access to timely, consistent and disaggregated data, their report tabled in October 2020 called for a reconciliation between statistical needs and privacy of Canadians [16].

The COVID-19 pandemic has had unprecedented impacts on Canadians, and particularly on the most vulnerable populations. Social movements for Indigenous rights, racial justice, and economic equity are gaining momentum. More than ever Canada is having national conversations about equity, diversity and inclusion. In response, Statistics Canada developed the Disaggregated Data Action Plan (DDAP), which aims to fill gaps and set in motion a culture shift, where data disaggregation becomes standard practice at the Agency. The purpose of the DDAP is to achieve a more equitable Canada and to further collect, analyze and disseminate disaggregated data to improve insights and decision making. The Government of Canada announced in the 2021 Federal Budget: "We cannot improve what we cannot measure" [17], and proposed to provide funding to Statistics Canada to implement DDAP that will fill data and knowledge gaps.

Providing disaggregated data at the lowest level of geography possible, for Indigenous peoples, persons from racialized populations, persons with disabilities and by gender is almost impossible with survey data only. Aiming for intersectionality (e.g., young, Black, women) as opposed to binary interactions is simply an utopia under the current conditions, given the survey sample sizes that would be required and the response rates that keep declining. The only way to reach that goal is to prioritize integration of survey, administrative and alternative data and to adopt modern statistical methods. From the individual perspective, large integration of data is privacy intrusive and creates concerns when not well and fully articulated. With that in mind, a secure infrastructure for data integration (SIDI) is being designed to balance the statistical needs of a healthy and modern digital society with the privacy of Canadians.

While SIDI is in the planning stage, a multi-phased engagement and communication strategy will be deployed in 2022 to assure Canadians of Statistics Canada's commitments to protecting their personal information and using it in a privacy-responsible way, and to gain their social acceptability.

The external engagement activities also afford the Agency to clarify direction on the SIDI's program development and build a collaborative relationship with stakeholders to facilitate the sharing of information, concerns, knowledge and best practices.

The next section provides design features of SIDI and how they align with privacy principles. It is important to note that privacy legislation is under review in Canada at the time the SIDI was conceived and this paper was written. This review could lead to modifications or additions to privacy principles in place.

3.2. Secure infrastructure for data integration

The design of the Secure Infrastructure for Data Integration (SIDI) starts with the notion that a linked population file is the result of linking multiple data sources using personally identifiable information (PII) to obtain a file that represents a reference population of individuals with their associated characteristics. It involves input data sources (source layer), linkage and data attribution (transformation layer) and distribution of the resulting linked file to approved employees (provision layer).

The input data sources, as well as the resulting linked population file, can include attributes that can be classified into three categories: personal direct identifiers, statistical numbers and analytical variables. Personal direct identifiers (PDI) are a form of PII. They are any data element that used alone could uniquely identify a natural person. In other words, they are variables for which all or some values have the possibility to uniquely identify an individual. Names and the Social Insurance Number in Canada qualify as PDIs. Similarly one can refer to business direct identifiers (BDI) as data elements that used alone could uniquely identify a business.

Statistical numbers are permanent or persistent numbers that uniquely refer to a person, a building, a building unit or a business. Unlike direct identifiers, they are anonymous; They do not disclose the identity of a person or a business. Statistical numbers can replace sensitive direct identifiers on files and preserve the statistical utility of the data if assigned consistently across files, i.e., they retain the capacity to link data across files when needed, without re-identification, because individuals keep the same anonymous statistical number across files and over time. Statistical numbers do not fully anonymize files per se since files can contain analytical variables that once combined can present a

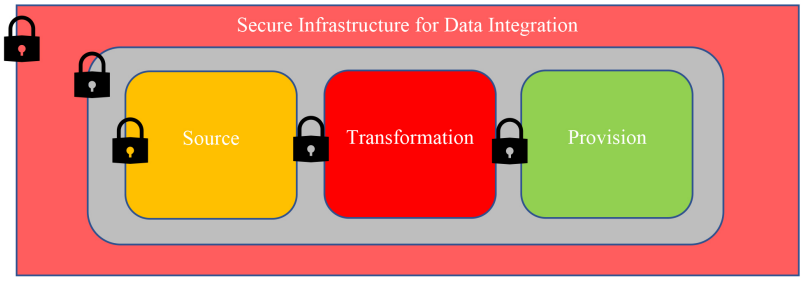

Fig. 1. High-level overview of the Secure Infrastructure for Data Integration.

disclosure risk. This being said, if a statistical organization takes great care in dissociating statistical numbers from direct identifiers in its information holdings, it can greatly reduce internal vulnerability threats such as unintentional disclosure or misuse of the information.

An important characteristic of the SIDI (see Fig. 1 for an illustration) is that each layer (source, transformation and provision) will constitute a separate compartment with its own security safeguards and access protocols tailored to the sensitivity of the data that it contains and manipulates. The transformation compartment for example, where data integration is performed using PDIs and where persistent statistical numbers are established and assigned, will have the highest security safeguards given the highly sensitive nature of the personal information (represented by the red color of the transformation compartment in Fig. 1). The access to this compartment will be limited to a very small number of record linkage experts and activities will be monitored to detect unauthorized accesses should they occur and suspicious behaviors of those who access the compartment. Other compartments or subsections of compartments may have medium to low security safeguards commensurate to the sensitivity of the data.

It is important to note that SIDI and its secure compartments are already within the "walls" of a Government of Canada approved Protected B IT infrastructure that can only be accessed by Statistics Canada employees based on strict protocols, who are themselves sworn under the *Statistics Act* to protect and not disclose the information. In other words, Statistics Canada plans to add extra security layers to a system that is already secure and plans to do so to reassure Canadians and itself that all measures have been taken to protect the information of Canadians. Each compartment is further described next.

3.2.1. The source compartment

Statistics Canada has obtained and has used administrative data for statistical purposes for more than 100 years. The act of obtaining administrative data has always been based on necessity and made in accordance with legislation in place. In 2019, Statistics Canada modernized its approach through the adoption of the Necessity and Proportionality Framework which was referred to earlier. This means that necessity must now be demonstrated with clearer evidences of the benefits to the public. Obtaining administrative data also has to be proportional, i.e., balanced against the privacy invasive nature of collecting and re-using personal information which was originally collected for other purposes. This also involves taking into consideration the sensitivity of the data, making sure that the approach is designed in a way that will achieve the desired outcome, that its benefits outweigh its privacy intrusiveness nature and that no other less privacy intrusive approach exists that could achieve the same goal. Since the implementation of the Framework in October 2019, any new data acquisition at Statistics Canada has been strictly tested against these necessity and proportionality principles.

With respect to the creation of a reference population file of individuals who live in the country at a specific period in time, administrative data sources that record demographic events are required in the absence of a population register or a recent census of population, i.e., natural growth (births and deaths), international migration (immigration and emigration in and out of the country) and internal migration (movements within the country). In Canada, the 10 provincial and 3 territorial governments register births and deaths occurring in their jurisdiction. This information is obtained and protected by Statistics Canada under the *Statistics Act*. In the same way, Statistics Canada receives immigration information on permanent and temporary residents from Immigration, Refugees and Citizenship Canada. For emigration and internal migration, there is unfortunately no single source that records these events in Canada. Statistics Canada obtains tax data sources (e.g., income tax reports, child tax benefits, statements of remuneration) from the Canada Revenue Agency and

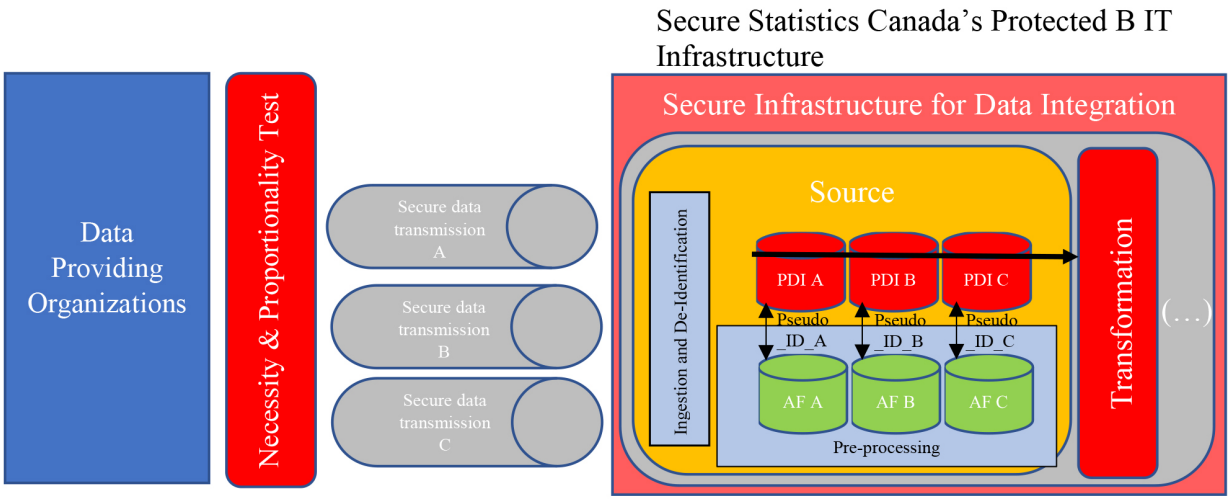

Fig. 2. Overview of administrative data acquisition and SIDI source compartment.

driver's licence information from most provincial and territorial governments that assist Statistics Canada in predicting which individuals appear to be in or out of the country, as well as establishing the most plausible usual place of residence of those who still live in the country. These governmental administrative data sources are critical constructs of reference populations and are called contributor files. Note that in Canada there is no universal personal identification number that can link individuals across governmental data sources.

The demographic concept of a base population is important here. It is a population established at a specific point in time that will form the basis to establish the next one. Thus it is iterative in nature. Normally, the population is "rebased" every census year because the statistical organization benefits from both the complete enumeration of the census and the contributor files described above to correct for any census coverage deficiency. In between censuses, this base population is updated using new contributor files to create the next base population until the next census is run and the population is fully rebased.

An assessment of the needs for reference populations within the Agency has revealed that SIDI should create reference populations twice a year for now, especially as of July 1. The date of July 1 is chosen because it is the current reference point of Statistics Canada's annual censal and post-censal population estimates. January 1 could be the second date chosen to provide statistical programs with an updated reference population six months later, e.g., to serve as a survey frame for social surveys. In terms of implementation, the transformation compartment of SIDI will keep information on the previous base population that was created. Only the new contributor files that are needed to update this population will be brought into the source compartment. As with any data file that Statistics Canada obtains from providing organizations, contributor files are sent using approved standard secure transmission procedures and systems. A data custodian, typically the director of the most relevant statistical program, is responsible for the secure storage, use and retention of the file once at Statistics Canada.

With the implementation of SIDI, it is planned that each contributor file would be brought rapidly into the source compartment where PDIs would be separated from the analytical variables that may be present on the file (see Fig. 2). This process would be executed as soon as the contributor file is received by Statistics Canada to maximize privacy protection. At this phase of data ingestion where an initial de-identification occurs, a pseudo record identifier (Pseudo_ID in Fig. 2) would be assigned to each record on the file and remains the only variable common to both parts of the file (the "PDI" file and the analytical file (AF)). Some efforts could be made to assign the same pseudo record identifier to the same individual across different vintages of the same contributor file. For example, it is possible to associate the same pseudo record identifier to the same driver's licence number each time a new provincial driver's licence file comes in. The driver's licence number is a PDI of no usefulness for linkage to other contributor files, since it is solely used for the administration of the driver's licence program within a province. However, it would be of interest to the linkage of different vintages of that contributor file over time. Statistics Canada's

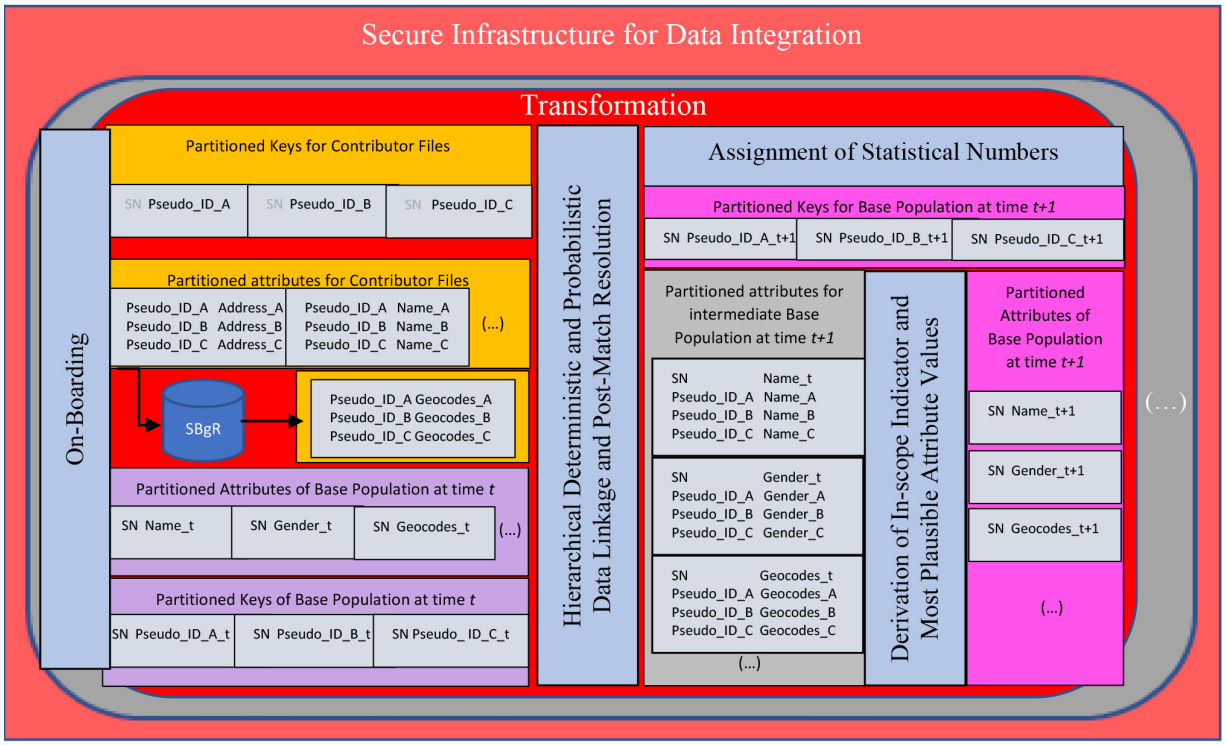

Fig. 3. Overview of SIDI transformation compartment.

goal is to design SIDI with privacy at the forefront and for that reason the intention is to bring only the latest contributor files into the transformation compartment. The pseudo record identifiers would replace administrative program numbers (such as the driver's licence number) and because of their longitudinal nature (except for errors made in the management of the administrative program number), they would help recognize records that refer to the same individuals on a given type of file over time.

Finally, the process in the source compartment is completed when the PDI files are made available to the transformation compartment. The analytical files are made available to pre-processing where basic cleaning, coding or transformation of the analytical variables occur. Figure 2 illustrates and summarizes the process in the source compartment. It presents the case where three contributor files (A, B and C) are brought into the source compartment. In reality there will be more files (e.g., birth, death and driver's licence files for each province and territory, national tax and immigration files). Also, an individual is not expected to appear on all files, which has not been made explicit in the figures presented in this paper to avoid overloading the illustration of the processes.

3.2.2. The transformation compartment

Four main processes are performed in the transformation compartment: on-boarding of files, data linkage, assignment of persistent anonymous statistical numbers, and in-scope and variable derivation to attribute the most plausible values to key basic demographic variables such as age, sex and usual place of residence. They are represented by the light blue rectangles in Fig. 3.

The on-boarding process takes the PDI files available from the source compartment and creates a partitioned environment for attributes within the transformation compartment. In other words, each PDI (names, addresses, dates of birth, gender, etc) constitutes a partition of that environment. The pseudo record identifier is kept in each partition. Processing also occurs on each PDI to prepare and standardize them for data linkage (not explicitly shown on Fig. 3).

Similarly, a partitioned environment for keys is also created where each contributor file has its partition. The keys included in each partition are the pseudo record identifier for that file and a provision for the persistent and anonymous statistical number (referred to as "SN" in Fig. 3) assigned permanently to each individual. Such partitioning of keys and attributes constitutes an addi-

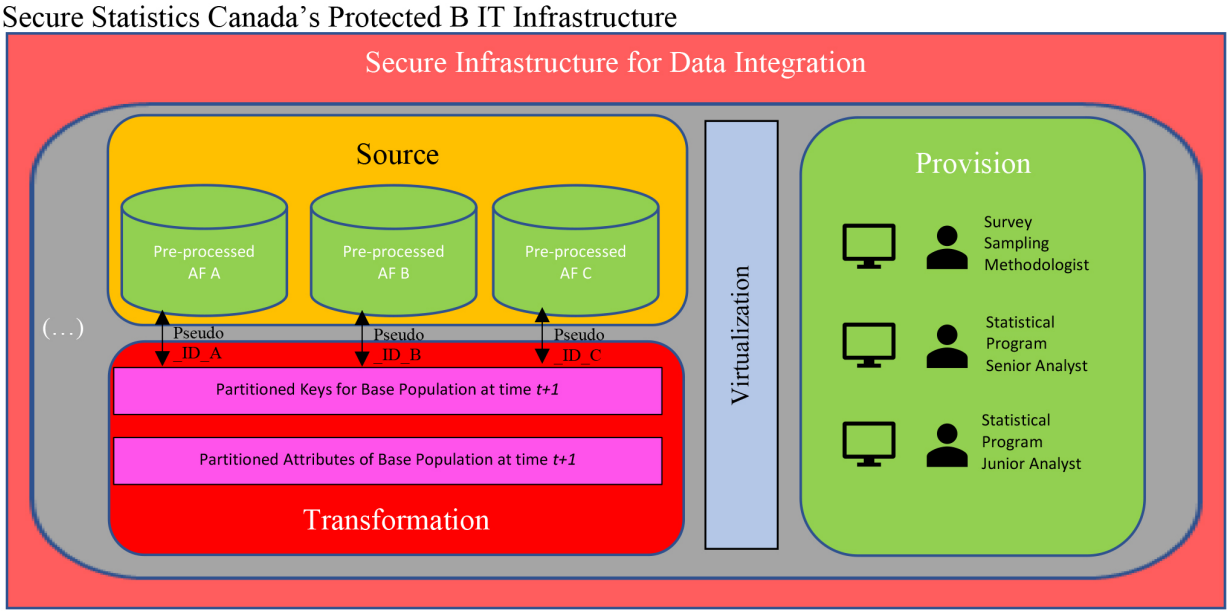

Fig. 4. Overview of SIDI provision compartment.

tional security safeguard offered by SIDI. Access to these partitions will be controlled by different personas (i.e., employees' roles) within the infrastructure and use different security groups.

The previous base population (created at time t) sits also in the transformation compartment and is partitioned the same way. The partitioned environment of its attributes records the most plausible value of each base population attribute established at time t (the process is described later in this section). The partitioned environment of keys for the base population includes a partition for each contributor file that was used to validate or update it at the time of its creation and includes the pseudo record identifier and the statistical number. For illustration purposes, Fig. 3 presents the connection of the three contributor files (A, B and C) to the SN via the pseudo record identifiers, which may actually be blank in certain instances.

The linkage process is decoupled. First the address partition is linked to the Statistical Building Register (SBgR) using the Register Matching Engine.[2] The address is thus associated to the most plausible statistical building unit number present on the SBgR, which also comes with the standard geographical classification that provides standard codes for the geographical regions of Canada such as provinces and territories, census divisions (counties, regional municipalities) and census subdivisions (municipalities). This linkage can be referred to as address geocoding. The address is an important linkage key in population data linkage. In Canada, no unified national standard exists for location addresses. Using the address geocodes in the linkage rather than the "raw" addresses helps remove the "noise" that could be created during the linkage and reduces linkage errors. Moreover, from that point on the address geocodes, and not the detailed address, are passed to the second step of linkage, and that further increases privacy and data security, since the other individual's attributes are decoupled from the complete address information.

The second stage of linkage uses the partitioned keys, the address geocodes and other attributes of the previous base population and the contributor files. A linkage strategy[3] that consists of a succession of hierarchical deterministic or probabilistic record linkage techniques

[2] Although not described in this paper, the design also allows splitting business information present on a file to be linked to the Statistical Business Register using the Register Matching Engine to obtain statistical business numbers, to which business characteristics can be later associated if needed.

[3] Part of that linkage strategy might involve reusing linkage keys that already exist in the Statistics Canada's Social Data Linkage Environment (SDLE) [18]. The Derived Record Repository of the SDLE contains a SDLE identifier (very similar to the notion of a statistical number described in this paper) and its connection to a number of files, many of which are the contributor files needed to establish reference population files.

Table 1
SIDI responsiveness to privacy principles[4]

Privacy principle and its description	Particularities of SIDI
Accountability: An organization is responsible for personal information under its control and shall designate an individual or individuals who are accountable for the organization's compliance with the following principles.	SIDI will be set up in accordance with a Privacy Impact Assessment (PIA) approved by Statistics Canada's Chief Privacy Officer and the Chief Statistician. Clear ownership and accountability for SIDI will be given to the director of an organizational unit within Statistics Canada.
Identifying purposes: The purposes for which personal information is collected shall be identified by the organization at or before the time the information is collected. Limiting collection: The collection of personal information shall be limited to that which is necessary for the purposes identified by the organization. Information shall be collected by fair and lawful means.	The collection of personal information is performed in accordance with the *Statistics Act* and privacy legislation in place. In collaboration with the Office of the Privacy Commissioner of Canada, Statistics Canada has developed a Necessity and Proportionality Framework that will continue to be used to demonstrate the public value of collecting personal information and integrating this information using SIDI.
Limiting use, disclosure, and retention: Personal information shall not be used or disclosed for purposes other than those for which it was collected, except with the consent of the individual or as required by the law. Personal information shall be retained only as long as necessary for fulfilment of those purposes.	The use of personal information is for statistical purposes only and performed in accordance with the *Statistics Act*. The purpose of periodically creating reference population files from collected personal information with SIDI (e.g., person-level survey frames) will be clearly demonstrated in the approved PIA. Programs at Statistics Canada that have needs for other population files will need to measure their request against the Necessity and Proportionality Framework, including the absence of other less privacy intrusive alternative to achieve the same goal, and receive approval to have their population files created by SIDI. If the particular request is deemed sensitive then further privacy impact assessments may be requested. Any population file created by SIDI will be limited in content (in terms of records and variables), created and retained for the identified and approved internal use. Statistical information derived from population files and released externally will be done in accordance with the *Statistics Act* and will be in the form of aggregates that do not contain identifiable personal information.
Accuracy: Personal information shall be as accurate, complete, and up-to-date as is necessary for the purposes for which it is to be used.	SIDI will be designed with methods and processes that treat personal information in a way to ensure cross-sectional, horizontal and longitudinal statistical quality (i.e. within and across statistical programs, and overtime). Statistics Canada will ensure that personal information is as accurate and up-to-date as practically reasonable.
Safeguards: Personal information shall be protected by security safeguards appropriate to the sensitivity of the information.	SIDI will periodically produce a base population that can be used to perform consistent and Agency-wide de-identification of personal administrative data files held at Statistics Canada.[5] The SIDI also implements a number of modern security measures: – Regular vulnerability and penetration tests – Compartmentalization of data and processes – Access to the most sensitive data by only a handful of employees – Secure interfaces – Data encryption at rest and in transit – Role-based and controlled accesses – Application of a framework that monitors accesses, detects unauthorized accesses and suspicious behaviors of authorized accesses, and reinforces prevention and safeguards as needed.

[4]To ensure a responsible privacy approach, the application of the privacy principles to SIDI presented in this table were guided to some extent by those of the OECD (http://oecdprivacy.org/). In addition, Statistics Canada has a Chief Privacy Officer in place whose function is to ensure that the organization is at all times privacy compliant with the Canadian *Privacy Act*.

[5]Any personal administrative file can actually follow the process described in the source, transformation and provision compartments presented in Figs 2 to 4, with the main difference that the file is not necessarily a contributor file. The administrative file is linked to the base population in the transformation compartment to determine the statistical number that corresponds to and can replace the PDIs in the file, which is then made available in the provision compartment in a de-identified way. The statistical value and utility of the data is thus preserved by the presence of persistent anonymous statistical numbers. Data files created for surveys using SIDI person-level frame will use the statistical numbers rather than PDIs and will be de-identified "by design".

Table 1, continued

Privacy principle and its description	Particularities of SIDI
Openness: An organization must make detailed information about its policies and practices relating to the management of personal information publicly and readily available.	The SIDI will gain social acceptability by engaging with Canadians early. The objectives of SIDI Engagement and Communication Strategy will be to: – Clearly define, convey and articulate the Necessity and Proportionality of the SIDI scope and approach – Obtain feedback from Canadians and adapt as required – Reaffirm and reassure Canadians of Statistics Canada's commitments to protecting their personal information, e.g., via Statistics Canada's Trust Center.
Individual Access: Upon request, an individual shall be informed of the existence, use and disclosure of his or her personal information and shall be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate.	SIDI will improve the experience of Canadians who file privacy requests to Statistics Canada. The personal information provided by Canadians as part of their request can follow the source, transformation and provision compartment described in Figs 2 to 4. It will be associated to the information that was requested by the individual via the persistent statistical number and returned to the individual without the statistical number. Canadians who wish to challenge the accuracy and completeness of the information in administrative data sources will be encouraged to contact the original organization that has provided that information to Statistics Canada.

is executed, followed by an extensive post-match resolution exercise to determine the valid matches. The third stage confirms the statistical numbers for individuals who were present on the previous base population and their association to the most recent contributor files and assigns new statistical numbers to the new individuals appearing on contributor files. The fourth and final stage analyses the linkages that were made between the last base population and the contributor files to determine if the individual is still considered part of the reference population, i.e., "in scope". The attributes in the previous base population and in the contributor files are also analysed to determine the most recent plausible attribute to be recorded on the newly created base population (at $t + 1$). These last three stages may sound trivial but they are not. The absence of a universal personal identification number makes the linkage strategy and the attribution of persistent anonymous statistical number significantly more complex. They could be the subject of a paper on their own.

3.2.3. The provision compartment

The provision compartment is a platform or hub where the reference population files are securely distributed to employees who need access to the files in their statistical programs. In addition to the base population attributes derived in the final stage of the transformation compartment, the partitioned keys of the base population allow the connection to the analytical files (AF in Fig. 2) that are kept in the source compartment. The provision compartment actually does not hold files physically per se. It uses virtualization, i.e., it provides a virtual rather than an actual representation of the reference population file that a given employee needs (in terms of records and variables) and the views vary based on the employee's role (aka personas). This is illustrated in Fig. 4.

3.2.4. SIDI's alignment with privacy principles

Many of the responsible privacy elements of SIDI were covered earlier in Section 3. However it is useful to recapitulate what they are and touch on others that may not have been covered explicitly. Table 1 above provides that summary.

4. Concluding remarks

This paper has illustrated how the design of Statistics Canada's register infrastructure has evolved with the privacy landscape, and more specifically how the design of a secure population data integration infrastructure to create reference population files was elaborated with privacy at the forefront. The SIDI is expected to close the gap left by the absence of a statistical population register that, when interconnected to a statistical building register, that is itself interconnected to a statistical business register, creates the necessary statistical foundation to study people's demographics in relation to where they live and where economic activities are performed.

Experts in official statistics will undoubtedly see the benefits of a system of integrated statistical registers. However, it is important to demonstrate to Canadians that the proposed infrastructure will ensure the privacy of their data while meeting the data needs of society, i.e., balancing individual and collective rights.

Along with privacy, official statistics are key pillars of democracy. They provide objective information on the state and progress of the economy, society and environment to support the democratic process and to guide public and private decision making. This information is increasingly needed at the community level and for population subgroups. It no longer suffices to

tell Canadians how their country, its people and its businesses are doing; Canadians want to know how their own neighbourhood and the people to which they relate are doing.

There is no simple solution to achieve the production of that wide range of disaggregated statistics with the quality that governments and Canadians expect from their national statistical organization. This requires a more nuanced and interrelated approach that takes into account the balance between producing statistics and respecting individual personal information.

The traditional way is to conduct surveys to collect the required information. Selecting a sample from a reference population produced by SIDI is likely to yield statistical results that are less exposed to biases detrimental to marginalized populations than doing so by reaching out to people via a sample of dwellings. The reason is that the basic demographic attributes that SIDI will include in the reference population will inform the sample design and selection and ensure that no subpopulation of interest is left behind. This also represents a more direct way to reach Canadians with the hope that it will alleviate the declining response rates observed in social surveys.

The other way of meeting the need for disaggregated statistics is to obtain and use administrative data. There are however two primary challenges with sources of administrative data taken one at a time. First, it is not always clear how well it covers the population of interest. Second, a single source often does not contain all the attributes that can meet the multifaceted needs for disaggregated statistics. Connecting each administrative data file to the reference populations created by SIDI (via the persistent and anonymous statistical numbers) is a solution to both problems. First, it allows the administrative data file to be assessed for coverage issues against the relevant reference population, which in turn can offer corrective measures if the administrative data file is planned to be used alone to derive statistics, thus preventing biases against certain subpopulations. Second, by having all necessary administrative data files de-identified but connectable by statistical numbers allows attributes to be selected across files and viewed together in the SIDI provision compartment. It is thus possible to conduct multifaceted or intersectional analyses that would otherwise be more challenging to do in a timely fashion.

SIDI is still at the planning stage and the design of SIDI presented in this paper may evolve as Statistics Canada engages with Canadians to obtain feedback and adjust. The goal of SIDI is to design a solution that reconciles increasing needs for trusted statistical information and privacy of Canadians. This objective has guided so far Statistics Canada's design of SIDI and will continue to do so as the Agency further refines it toward implementation.

Acknowledgments

The author would like to express sincere thanks to two colleagues who have been a source of inspiration for this paper. First, immense gratitude goes to Jean Pignal, who was a pioneer in establishing the vision for a system of integrated statistical registers at Statistics Canada and first led the elaboration work of a statistical population register. He unfortunately passed away in April 2021 after a battle with cancer. Second, outstanding recognition goes to Patrick Mason who took the baton and is the great mind behind the new Secure Infrastructure for Data Integration.

Thanks also go to André Loranger and Eric Rancourt, respectively assistant chief statistician and director general at Statistics Canada, for valuable comments made as part of the official internal peer and institutional reviews of this paper, and also to Antonio Bakopoulos, Sylvain Delisle, James Falconer and Philippe Gagné who provided many suggestions that improved the paper.

Finally, the author would like to thank the referees of the paper for final recommendations before the paper's publication.

References

[1] United Nations Economic Commission for Europe (UNECE). Register-based statistics in the Nordic countries: Review of best practices with focus on population and social statistics. United Nations Publication. 2007. Available from: https://unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf (accessed January 15, 2022).

[2] United Nations Economic Commission for Europe (UNECE). Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses. United Nations Publication. 2008. Available from: https://unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20184.pdf (accessed September 29, 2021).

[3] Wallgren A, Wallgren B. Register-Based Statistics – Administrative Data for Statistical Purposes. John Wiley & Sons, Ltd. 2014.

[4] Statistics Canada. Business Register (BR). 2021. Available from: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=1105 (accessed January 15, 2022).

[5] Statistics Canada. Address Register (AR). 2019. Available from: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5310 (accessed January 15, 2022).

[6] Office of the Privacy Commissioner of Canada. PIPEDA Fair Information Principles. 2019. Available from: https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/ (accessed January 15, 2022).
[7] Statistics Act, Revised Statutes of Canada, c. S-19. 1985. Available from: https://laws-lois.justice.gc.ca/eng/acts/s-19/fulltext.html (accessed January 14, 2022).
[8] Statistics Canada. Final Report on 2016 Census Options: Proposed Content Determination Framework and Methodology Options. 2012. Available from: https://www12.statcan.gc.ca/census-recensement/fc-rf/reports-rapports/r2_index-eng.cfm (accessed January 14, 2022).
[9] Trépanier J, Pignal J, Royce D. Administrative Data Initiatives at Statistics Canada. Proceedings of the 2013 Research Conference of the Federal Committee on Statistical Methodology. 2013. Available from: https://nces.ed.gov/FCSM/pdf/G1_Trepanier_2013FCSM_AC.pdf (accessed November 27, 2021).
[10] Rancourt E. Admin-First as a Statistical Paradigm for Canadian Official Statistics: Meaning, Challenges and Opportunities. Proceedings of Statistics Canada 2018 International Methodology Symposium. 2018. Available from: https://www.statcan.gc.ca/en/conferences/symposium2018/program (accessed December 17, 2021).
[11] Statistics Canada. Statistics Canada Report on Plans and Priorities 2014/2015. Catalogue no. 11-018-XWE ISSN 2292-5252. 2014. Available from: https://publications.gc.ca/collections/collection_2014/statcan/CS1-2-2014-eng.pdf (accessed November 27, 2021).
[12] Statistics Canada. First report on the Census Program Transformation Project: Researching a new approach to census-taking. 2017. Available from: https://www12.statcan.gc.ca/census-recensement/fc-rf/98-506-x/98-506-x2017001-eng.cfm (accessed January 15, 2022).
[13] Gagné P, Pignal J, Quadir T, Wolfe C. Towards a Register-centric Statistical System: Recent Developments at Statistics Canada. Proceedings of Statistics Canada 2018 International Methodology Symposium. 2018. Available from: https://www.statcan.gc.ca/en/conferences/symposium2018/program (accessed November 12, 2021).
[14] Rancourt E. The scientific approach as a transparency enabler throughout the data life-cycle. Statistical Journal of the IAOS. 2019; 35(4): 549–558. doi: 10.3233/SJI-190581. IOS Press.
[15] Statistics Canada. Principles of Necessity and Proportionality. 2019. Available from: https://www.statcan.gc.ca/en/trust/address (accessed November 12, 2021).
[16] Canadian Statistics Advisory Council (CSAC). Canadian Statistics Advisory Council 2020 Annual Report – Towards a Stronger National Statistical System. 2020. Available from: https://www.statcan.gc.ca/en/about/relevant/CSAC/report/annual2020 (accessed November 12, 2021).
[17] Government of Canada. Budget 2021 – A recovery plan for jobs, growth, and resilience. 2021. Available from: https://www.budget.gc.ca/2021/report-rapport/p3-en.html#chap7 (accessed November 12, 2021).
[18] Statistics Canada. Social Data Linkage Environment – Privacy impact assessment. 2021. Available from: https://www.statcan.gc.ca/en/about/pia/sdle (accessed January 15, 2022).

Glossary

Administrative data census: A census based on data held in various administrative registers or other administrative data sources that are transformed and integrated into a system of base and non-base statistical registers to provide the necessary data to a population and housing census. It is possible to complement the information with unit record data (microdata) from already-existing sample surveys, other administrative or alternative data sources. All administrative data censuses have in common the fact that no specifically designed census questionnaires are used to collect information about the population.

Administrative data sources: Data holdings containing information which is not primarily collected for statistical purposes. This type of data is collected by government departments and other organizations for the purposes of registration, transaction and record keeping, usually during the delivery of a service.

Administrative registers: Registers primarily built and maintained for administrative purposes.

Base statistical register: Statistical register that defines important object types (statistical units) and object sets (standardized populations) in a national statistics system. It contains links to objects in other base statistical registers and to statistical registers that relate to the same object type. There are four base statistical registers: population, buildings, business and activity.

Building: 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. A building may be used or intended for residential, commercial, industrial or institutional purposes, including the provision of services. A building can be entered by persons or animals and is suitable or intended for protecting them and objects. For technical reasons, "building" also includes a separately usable underground construction.

Building unit: Part of a building, and is either residential or non-residential. It must have its own entrance, either through an outer door or through an interior door in a shared hallway. The building unit should have its own identifier within the building. If such an identifier is not available, a unit description can be used to identify the building unit.

> A *residential building unit* is a structurally separated and independent place constructed, built, converted or arranged for human habitation, whether occupied or not. It can be used to identify private dwellings.
> A *non-residential building unit* is a structurally separated and independent place constructed, built, converted or arranged to be occupied or used for

commercial, industrial or institutional purposes, including the provision of services.

Combined census: A census that relies partly on base and non-base statistical registers, administrative, alternative or existing sample survey data, and partly on a limited collection of data from a field enumeration of the population for specific variables.

Economic units: Comprised of two separate but related types of economic units, i.e., the legal units and the statistical units.

> *Legal units* are units registered in administrative registers, e.g., taxation registers, including non-market units such as government departments and non-profit institutions.
> *Statistical units* are delineated for statistical purposes using legal units. Enterprises and establishments are two important statistical units and the statistical business register records the links between them.

Personally Identifiable Information (PII): Any information that relates to an identified or identifiable natural person (as opposed to a legal person), either directly or indirectly, that can be used to distinguish an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual.

Personal direct identifiers (PDI): They are a form of personally identifiable information that used alone could uniquely identify a natural person. They are variables for which all or some values have the possibility to uniquely identify an individual.

Register: Systematic collection of unit-level data about a specific group of units uniquely identifiable that is organized in such a way that updating and expansion with new variable values for each unit is possible.

Statistical activity register: A base statistical register of activities, i.e., a relational object that connects individuals to organizations. Activities can be of different kinds: job activity, study activity or other activity a person may have with different welfare institutions (e.g., military service, employment insurance, disability pension).

Statistical building register: A base statistical register of buildings and building units in a country. It covers buildings that are residential, non-residential or a mix of the two.

Statistical business register: A base statistical register of economic units that are resident of a country.

Statistical numbers: permanent numbers that a National Statistical Organization creates for statistical purposes to uniquely identify a person, a building, a building unit or an economic unit on its statistical registers, databases or files. When data are extracted from the register and used in other statistical activities (e.g., linkage to other data, analysis), these numbers are anonymous and can replace more sensitive direct identifiers associated to people and businesses.

Statistical population register: A base statistical register of residents of a country that compile their life events and basic socio-demographic characteristics in such a way that the resident population at a point in time or over a specific period of time can be extracted and the size and characteristics of that population can be compiled.

Statistical registers: Registers created for statistical purposes. They are typically created by transforming and integrating data from administrative registers and/or other administrative data sources. Other alternative data sources can also be used.

Traditional census: A census approach that collects basic information about population and dwelling characteristics from all individuals at a specific point in time. More detailed characteristics are collected either from the whole population or on a sample basis. The census collects information using self-completed Internet or paper questionnaires, or in-person or telephone interviews.