

# **Statistics Canada's Statistical Registers Integration Project: A new platform for a new era**

Shujaat Ansari, Jamie Brunet, Gaétan St-Louis, Philippe Gagné, Patrick Mason

Presented to the Wiesbaden City Group on Business Registers,  
Neuchatel, Switzerland  
September 2018

## **Executive Summary**

Statistics Canada has embarked on a project to build and maintain a Statistical Registers Infrastructure (SRI) that comprises core interconnected registers for the population, buildings, businesses and activities. This is a key element of a modernized data management strategy for the agency that will facilitate the induction and enrichment of data inputs from new and traditional sources in a manner that protects privacy of information while at the same time ensuring data can be efficiently cross-linked to fulfill the needs for cross-cutting and timely statistics and analyses. Also, the creation of a central Statistical Building Register (SBgR) will fill important gaps for data pertaining to housing, property values and will ensure consistent geographic coding across the various statistical programs of Statistics Canada. The SBgR will also be a tool to confront, validate and improve the coverage of the Business and Population registers.

### **I. Introduction**

This paper will provide an overview of the SRI project, including the envisioned design and key tasks and deliverables, and discuss the benefits of the SRI.

We will report on the data methods and sources being assessed for the SBgR, which is a key component of the project, and elaborate on how it will integrate with the Statistical Business Register (SBR).

Another important component of the new infrastructure, the Register Matching Engine (RME), which will ensure that consistent and effective methods are used for reconciling data from outside sources into the Statistics Canada's data ecosystem, will also be discussed.

Finally, we will touch on some opportunities and further considerations for an integrated registers infrastructure, including some data linkage and possibilities for improved coverage.

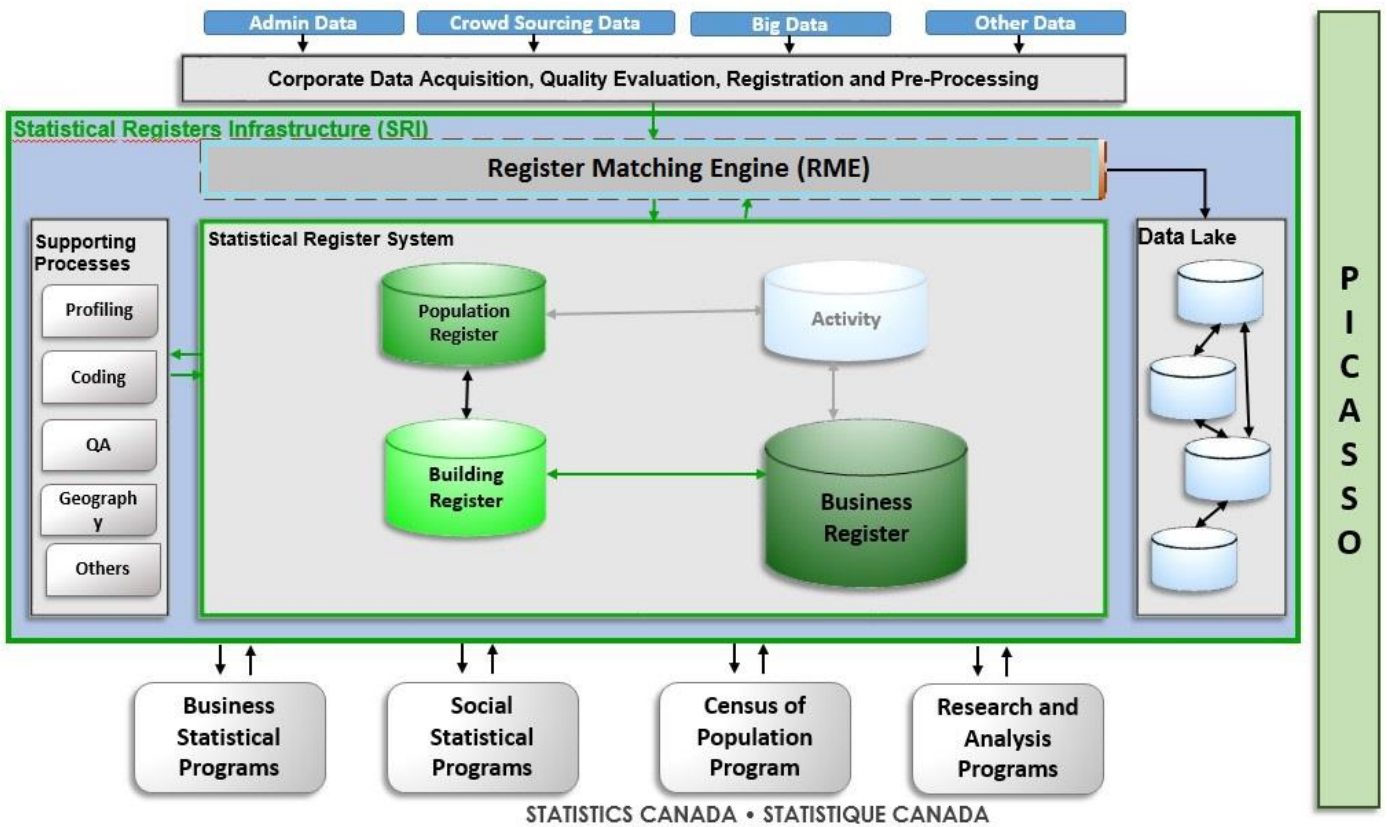
### **II. The Statistical Registers Integration: the Project Plan**

In the always-ongoing search for opportunities to produce more timely, efficiently-compiled and cross-cutting statistical information, Statistics Canada is transforming itself to be able to exploit

administrative/transactional/open source and big data and, where possible, eliminate the need for direct survey data.

The *Statistical Registers Integration* is central to this transformation. It is undertaking the development of a new infrastructure of core interconnected statistical registers that will be used as a coherent base by all the agency’s statistical programs to facilitate the induction and integration of the various types and volumes of data available in the digital era of today and tomorrow.

The new infrastructure will consist of four core statistical registers: the Statistical Population Register, the Statistical Building Register, the Statistical Business Register and the Statistical Activities Register.



For the **Statistical Business Register (SBR)**, the objective is to build on its existing system of data and functionalities that have established it as the centrepiece of Statistics Canada’s business and institutional data programs. Many of its components are well-suited to the new registers infrastructure and can be maintained, but there are some adjustments required.

Most importantly, we are developing a **Longitudinal Business Register (LBR)**: This is an important feature required for the SRI, as it will provide the basis for tracking businesses and institutions through time and also provide the basis through which register data can be historically revised through retroactive updates. The existing SBR system is designed to produce *current* point-in-time monthly snapshots, as this is what has been traditionally required for the SBR to fulfill its traditional role as a survey sampling frame. The LBR will fulfill the need for the SRI to support the development of current

and historical data pertaining to businesses and institutions. The LBR is now in development; the first implementation is slated for Fall 2018.

The **Statistical Population Register (SPR)** will provide listings of persons (along with necessary frame attributes), the target coverage of which will be all persons residing in Canada. It is being built by expanding on work that has been carried out in recent years to investigate the usage of administrative data (mostly tax data) to supplement -- and potentially replace in some cases -- census data. Through these efforts, an evergreen frame of persons living in Canada based on the census and other types of data can and is being constructed, though some work is still required to determine the optimal mix of data sources to be used. The first version of the SPR is being readied for the 2021 Census test planned for 2019. Subsequent to this test, a new version will be prepared in time for the 2021 Census, which will then form the basis for a pilot version of an administrative data-enhanced Census, which would be envisioned for implementation with Census 2026.

The **SPR** will evolve throughout this process, and Statistics Canada's social survey and analytic programs will be hooked in to use it for sampling and data integration purposes according to suitable time frames.

The **Statistical Building Register (SBgR)** will be a complete listing of residential and non-residential buildings in Canada. It is also being prepared to be able to support the census test in 2019, as it will be used hand-in-hand with the SPR in evolving to the administrative data-enhanced census. The two registers will be used in concert to help overcome one of the main challenges of using administrative data in lieu of census data: the need to ensure people are linked to the right places of residence at the right point in time.

The SBgR sources and methods and the purposes it will serve in the context of the SRI are expanded upon below.

The goal is for these two registers (SPR and SBgR) to be continually updated ("evergreen") and dynamically interrelated. Ensuring the best possible coverage for both of these registers is a challenge in and of itself, but the challenges will be even greater when it comes to placing "the right people in the right dwellings at the right point in time". Achieving this goal, however, will allow the social and household survey programs at Statistics Canada to harness this interrelatedness to greatly improve sampling and collection outcomes. For example, characteristics of the individuals that compose households can be used to more precisely target populations, and rostering of respondents can be done prior to collection beginning in the field.

2019 will be an important year for the SRI: the SPR and SBgR will have been developed with complete integration between them. In 2019, we will also conduct a pilot survey that will test the usage of the new register infrastructure by social statistics programs for their activities: delivering the frame, compiling the selected units, providing the contact information for collection activities and processing respondents' feedback in order to modify or correct the attributes in the register.

Once the SPR and SBgR are complete and implemented (in 2019), the full integration of the SBR into the SRI will be completed by providing the Business Register with the Building Register linkage key and attributes. The challenge will be the same, i.e., maintaining this connection between the location of enterprises (and institutions) and the buildings in which they operate.

The plan calls for delivery of the final, integrated and fully tested platform that includes the SPR, SBgR and SBR by 2021.

In the longer term, the **Statistical Activity Register** will be added to the project's scope. The purpose of this register will be to record the interactions between individuals and institutions in society. In particular, the themes covered by this type of register are often justice, education, health, labour, etc. In the meantime, exploration work is under way, but it will be incorporated more formally into the project in a few years.

Finally, and as discussed at greater length further below, a **Register Matching Engine (RME)** is being developed in order to allow for the integration, de-identification and linkage of both the various registers and the data files that will underpin them and to allow users to link data back to the registers from data lakes, marts and warehouses. (The task of being able to find these data assets will be facilitated by **Picasso**, a new corporate tool, which will act as a centralized corporate registry of fit-for-use data assets, as well as a corporate repository for statistical metadata). The RME will be a key component of the SRI and a version has been released for User Acceptance Testing in July 2018; the RME in Production will be available in 2019.

### III. Benefits of the SRI

Maintaining an interconnected infrastructure of registers will allow Statistics Canada to create an environment where data can be received, processed, integrated, extracted, analysed and disseminated at a faster rate than would be otherwise possible. The process will also be coherent and consistent, saving time and effort by ensuring that staff are utilizing the same repositories, files and services. This infrastructure provides the spine for Statistics Canada's new data management and production strategy:

#### a) A less survey centric approach to statistics

By maintaining a robust set of characteristics about the people, buildings, geography and businesses in Canada, and by providing the keys to link these data together, the Integrated Statistical Registers (ISR) will allow survey programs to both eliminate questions from questionnaires (as that data will be available within the frame) and to link non-survey data more easily with the base registers in order to replace the need for collection of certain variables from respondents via a survey. The ISR will also allow for ingestion and processing of very large data files (payment processing, for e.g.) which can then be used in tandem with the information available on the base registers in order to structure the data to create Register-based statistics.

#### b) Data Integration

A system of Integrated Statistical Registers, combined with a Register Matching Engine (i.e. a suite of record linkage tools and processes), will allow for various data files (Administrative Data, Crowd Sourced Data, Big Data, survey data, etc) to be received by Statistics Canada and to then be processed and integrated into our data holdings in a systematic, secure and consistent manner. Data would be received by Statistics Canada, pre-processed, de-identified and then linked (by statistical identifier numbers) to the various base registers. High quality data integration is seen as a key component of the Statistical Registers Integration Project; as noted by Wallgren & Wallgren (p. 21-23, *Register-based Statistics*):

“System-oriented thinking is fundamental for register-based statistics. To improve quality, it is insufficient to look at one register at a time; instead the system should be seen in its entirety. Special attention should be paid to the quality of the base registers and the identifying variables that act as links between the different registers.... The register system ensures that microdata can be integrated and used effectively and opens new possibilities for quality assurance.”

### **c) Common Frames, Common Processes**

Building an Integrated System of Statistical Registers means that Statistics Canada will be able to integrate data from various administrative and other sources into a coherent and consistent register system that will allow for the creation of register-based statistics, and for linkage and geocoding of various files. Files and data will be received into Statistics Canada via the same front doors, and will then be pre-processed and parsed by corporate tools; these files will then be sent to the Statistical Registers and Geography Division in order for them to be processed through our Register Matching Engine, which will assign statistical numbers to the records and allow for their de-identification and subsequent housing within a data mart.

Additionally, over the past 10 years, the Business Register at Statistics Canada has developed a set of collection and client facing processes and systems that have allowed for a high level of integration with the collection systems and subject matter areas at Statistics Canada. From a collection perspective, this means that collection staff and respondents have access to the latest frame information during collection, and that the information required for collection (respondent information, special arrangements, and data replacement strategies) is made available in a standard and generic format from the frame. A benefit of developing a common statistical registers infrastructure is that the newly developed registers (Building and Population) will be able to use a similar supporting infrastructure as the Business Register does for its production processes. This will allow Statistics Canada to streamline the way collection is managed across social, household and business surveys.

### **d) Timeliness of Data**

As Statistics Canada strives to make more relevant and timely data available to our partners, clients, and stakeholders the Statistical Registers System will facilitate the ability of the organization to quickly ingest timely data from various sources and make it available to our subject matter experts and methodologists for linkage, analysis, processing and dissemination.

### **e) Data Management, Data Security and Data De-identification**

One of the key elements underpinning an integrated and coherent Registers System will be the use of common data repositories, data access controls and information management principles and policies. Access to data and functional roles for the Statistical Registers Infrastructure will be based on the criteria that 1) a user is able to access features and data required to perform their job and 2) a user is unable to access features and data that they are not approved to use. A Roles and Permissions Matrix has been developed that outlines how data and functionality permissions are associated to various ‘personas’; individual users are then assigned to personas based on their role and job functions. Personas have various permissions to manipulate data, as well as various clearance levels to view different levels of detail for data. The Permissions and Clearance Levels can be found in Annex B. Along with strong data access control, the SRI will employ common data repositories: while data files coming into Statistics Canada for pre-processing and parsing may be housed in data lakes, the files

coming out of the Register Matching Engine will have been transformed and structured in a manner that makes it consumable for users. This structured data will be stored in a data warehouse and data marts that will allow for structured access. The data tables that make up the base registers themselves will be access controlled using our Corporate Access Request System (CARS).

Confidentiality of data will also be managed and secured by ensuring that all files and records are de-identified as they come into the SRI: names of individuals, addresses, and other information that can be used to identify specific people or places will be stripped from files entering the data lake and will be replaced by statistical identifiers. This not only facilitates record linkage (used for building the registers themselves, as well as linking files downstream in the statistical process), but guarantees that users of the Registers are only exposed to the information they require in order to perform their job-specific tasks.

Information management policies and directives at Statistics Canada will help determine the retention periods and archiving timelines for the various register outputs and data files. The Business Register has a robust and well-defined retention policy, and the Building Register has taken its lead from these rules in order to define its own retention periods.

#### **f) Response Burden Tracking**

The new base Registers (Building and Population) will also borrow many of the Response Burden tracking practices already employed by the Business Register. Response Burden compilation will track every unit that is sampled at both the household (for dwelling based surveys) and household member level, and will be compiled based on Survey ID, Sampled Unit ID, Reference Period, Original and Final Method of Collection and Final Outcome of the questionnaire (i.e. received or refused, etc). The response burden module will also be able to track records on the registers that have been excluded from collection (this could happen for various reasons: response burden thresholds exceeded, complaints, interviewer safety) and for what duration they are excluded. We are also examining the feasibility of implementing a response burden scoring function, which would assign a weighted response burden score based on a number of factors: number of surveys, time it takes to complete those surveys, intrusiveness of the survey questions, time since last sampled.

## **IV. The Statistical Building Register**

The Building Register currently being developed will be central to the SRI. In essence, it will **list and characterize the physical structures where people can or do reside, or where economic activity can or does take place.**

It will identify whether a building is a private dwelling, in which case it will be matched to its residents on the Population Register, or a place of business or employment, in which case it will be matched to a 'statistical business location' on the Business Register (which includes business and public and para-public institutions.) In some cases, a unit on the SBgR will be identified as both a place of residence and a place of employment, and therefore have links to both the SBR and the SPR. Examples here include business owners operating in their homes, retirement homes or prisons.

Five main **functional roles** are envisioned for the SBgR once it is in operation:

**First**, the SBgR will be the primary mechanism through which the activities of people, businesses and institutions will be linked to a specific geographic location, and therefore associated with specific geographic areas (i.e. cities, provinces/territories, etc.). Currently, the geographic coding on the Business and social registers are developed and maintained somewhat separately. The SBgR will ensure consistency across statistical program areas and the maximum level of location specificity: there will be one central source for geographic codes, which will be easily acquirable by having the system of common identifiers on the SRI, and the data provided will include identifiers down to the Block Face level and also specify the latitudinal/longitudinal coordinates.

**Second**, the SBgR will be important for many types of statistical projects requiring record linkage. For example, in cases where it will be desirable to compile information to correlate demographic data with information pertaining to the nature of people's employment (e.g. industry of employment, size of business, etc), the location of employment would be identified through the SBgR, which would then link to the employer on the Business Register so that the necessary attributes could be acquired. Another example might be the need to select a sample of persons living in a specific type of dwelling; the sample would be drawn first from the SBgR, then residents living in those specific units could be further sub-sampled.

Further relevant examples of linkage opportunities can be found in Annex A

**Third**, the SBgR will be the basis for statistics and analyses pertaining to buildings as statistical units in and of themselves. *How many buildings are there and where are they located? What types of residential dwellings are people living in? What is the estimated value of the capital stock of buildings? What is the average square footage of buildings within a given area?* Having a central SBgR will provide a more coherent statistical picture in answering these questions.

**Fourth**, the SBgR will play a central role in the management of survey operations, as it will identify the precise residential, business and institutional locations where people reside, for the purposes of managing enumeration activities or survey visits.

**Fifth** and very importantly, the SBgR will greatly assist the SRI itself – and therefore Statistics Canada in general -- in making sure it accurately and completely covers the socio-economic make-up of Canada. By independently sourcing data from various sources on residential- and non-residential buildings, the SBgR will form a benchmark that should reconcile to and provide a quality assurance mechanism for the units observed on the other core registers.

### **Intended benefits for Business Register profiling**

The fifth role is particularly relevant for the statistical locations on the Business Register operating within complex enterprises. Having an evergreen SBgR that is built from a variety of administrative sources, independently of the Business Register, will provide an opportunity to confront the records on the SBR with the building units from the SBgR in order to close gaps in the SBR coverage. There are a few distinct ways this could help the SBR: 1) by allowing the SBR to assist its profiling activities with signals from the SBgR regarding units that may need to be added to various Enterprise structures (flagging structures that need to be profiled); 2) by allowing the SBR to automate certain births of locations, using various additional information sources along with the SBgR; 3) there has been an effort made over the past few years to improve the coverage of records on the SBR that help contribute to detailed and

robust allocation of data from survey and administrative sources. The Business Register produces a set of allocation factors that are used by subject matter and survey areas during processing in order to allocate their data to the lowest levels of geography possible; in order to assist this allocation the SBR has begun producing both 'T4' administrative records (locations on the SBR that seek to break out (i.e. delineate) some simple structures based on provincial employment remittances), as well as creating 'allocation entities' in order to reflect where revenues are being generated (for e.g. in the case of large rental landlords that may only have one record on the SBR, as its employees are all in one location, we can birth provincial receptacles to reflect where the rental income is being generated). The matching of SBgR building units against the SBR may present opportunities to expand the scope of this allocation work by allowing greater delineation of locations on the SBR, even when some of these units would not qualify as locations on the SBR under the traditional definition.

Likewise, the SBgR will permit verification that people have been correctly associated with places of residence. For example, if a new dwelling is observed, it could be verified that there are people on the SPR living in it, and we could make adjustments to the SPR as required.

### **How are we building the Building Register?**

In Canada, there is no specific administrative registry of buildings. The SBgR must therefore be built from various data sources. Statistics Canada is currently in the midst of testing new and existing sources and continues to make progress in building the SBgR.

### **Usage of existing data and methods**

Since the 1990s, Statistics Canada has maintained a register of mailing and locational addresses -- the Address Register (AR) -- which has been used primarily for providing listings of occupied and unoccupied dwellings for the census and social surveys.

The sources for compiling the AR have evolved over time, but today, the AR is constructed using a 'Point-of-Call' file of addresses obtained from the postal office (Canada Post), telephone billing files, online directory files such as Info Direct, the Yellow Pages telephone listings, and also data from personal tax returns. Listing exercises by field officers are also often used. Recently, online satellite imagery to identify growth areas and missing dwellings have been used to support and in some cases reduce the need for these field operations.

The Point-of-Call file is particularly useful as it provides distinctly-identified valid postal addresses. This is an improvement from the past. Up until recently, Canada Post could only provide *ranges* of valid street addresses. Nonetheless rural areas remain a particular challenge, and as postal addresses often do not refer to specific locational addresses, field listing are often required.

### **New data sources being evaluated**

The existing data sources provide a very good starting point for the private residential dwellings that will be on the SBgR. To build the SBgR, we will also need to acquire and add some variables necessary for it to be able to perform its intended roles. Also, we will need to add in records pertaining to non-residential (business and institutional) listings.



One of the challenges is to identify a source of information that will independently identify whether a building is residential or non-residential. This would be important for the SBgR to be able to perform the fifth functional role mentioned above (i.e. a benchmark for business locations and private dwelling listings).

Recently, Statistics Canada has gained access to **911 emergency** listings that identify precise locational data. Testing of these data is showing promise, particularly given their ability help precisely pinpoint units where the mailing address does not relate to a specific locational address, which happens often in rural areas.

We have also acquired **property valuation data** for specific provinces that commission market-value assessment for property tax purposes. Results are indicating that these will be useful for identifying the building types, and also for assigning a measure of property and asset value. This is especially important given the need for Statistics Canada to produce data to support policy analysis of foreign property ownership.

### **Usage of open data platforms**

There are also pilot projects to explore the potential usage of various types of internet information to assist with the building of the SBgR. This includes web scraping techniques that would harvest information from various types of web sites, such as (for example) commercial property listings or data available from online mapping tools.

We also piloted a crowd-sourcing approach to acquiring building data through the usage of the Open Street Map platform. Results indicated that this may be useful in some cases, but may also have some limitations given the need for incentives to participate.

It is important to note that these innovative new methods also come with the need to investigate and consider the legal and licencing ramifications that come with using third party internet data. A framework for these aspects is currently being developed.

## **V. Register Matching Engine**

Amalgamating several data sources without having a unique and universal administrative identifier requires developing a high-performance record linkage engine. This is crucial for the viability of the proposed infrastructure. The following are some fundamental elements that must be provided by such a tool:

- Assigning a unique statistical number to each new statistical unit (person or building).
- De-identifying units by concealing the raw personal identifiers taken from the register and storing these 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. This is probably the most complicated aspect. The situations to be addressed will be of varying complexity. For example, this can go from determining whether "William Smith", "Bill Smith" and "W. Smith" are the same

- person to determining whether two administrative records refer to the same building by comparing their respective GPS coordinates.
- The role of the linkage engine will also be to place individuals in their dwelling. There will be continuous queries (in the form of messages) from the person register to the building register to obtain the unique identifier. This requires a solid geospatial database that contains information on the road network, address ranges and all standard geographic borders that are used at Statistics Canada.
  - We will also provide a Record Linkage Service –custom processes and post-processes for individual clients (may include frame-reconciliation for surveys using lists from external sources, etc.)

The Register Matching Engine is composed of a suite of tools to process data and perform matching. It makes use of a Deterministic rule-based approach which:

- can lead to better tracing of logic and reasoning for matches
- allow for use of heuristic knowledge that has been developed
- is suited to geocoding as hierarchical matching can be employed
- designed and developed with the needs of registers (performance) in mind

There are drawbacks using only this approach, but overall it was preferred for Base Register maintenance: we wanted to ensure more consistent matching results and we needed to keep performance in mind... at each step of the way, we are looking at the performance to ensure the process is fast and efficient.

The Deterministic Rule-based logic defines a series of rules describing the matching of different sets of variables. To do the matching, a series of score functions/various comparators are used along with weights, and entropy/frequency measures to arrive at potential matches. We also make use of conversion tables, search keys, various string comparators, etc. We also created the concept of linked records to make use of household/building information. Efficiencies have been built into the matching process to limit comparisons to aid in performance (indexes, logic...). The RME is highly parameterized, allowing users more control; this fosters an environment of critical analysis. And under a continuous improvement model, the expectation is that future development will add to the toolbox and allow additional methods to improve overall linkage.

Once a file is processed, the RME returns multiple results each with a score measure to enable post-match resolution based on the requirements of the post-process. Thus, the multi-link resolution will be performed based on the anticipated use of results and the existing register linkages. Similarly, the weak-link resolution will be use-dependent depending on client requirements. There will also be a series of tools/algorithms developed to guide the post-match resolution process. The Register Matching Engine will also be used to link business entities on the Business Register to the building units on the Building Register.

## **Annex A: Further Linkage Opportunities that the SBgR would support**

Having an Integrated Registers System with a SBgR at its core will ultimately allow for the creation of robust and multi-faceted statistical outputs, from a combination of register-based and survey-based sources. A couple of examples of programs and outputs that could result from this infrastructure follow.

#### **a) The Homecare and Residential Care Facilities Surveys**

These surveys are censuses of all units in their respective NAICS (North American Industrial Classification System) using only tax data; there is no response burden. They will each pull their population from the Business Register and will thus have a link to a statistical number. While the initial plan is to produce estimates for the enterprises within these industries, the Integrated Registers will allow this data to be cross linked to the collectives (buildings) and people using these services, in order to create characteristics of these dimensions and to better understand which populations use which services, and how these track over the private and public & for-profit and non-profit spheres.

#### **b) SCIEU**

The Survey of Commercial and Institutional Energy Use was run using the Business Register as its frame in 2014. There were some limitations of doing this, as the survey targets the energy use in buildings and is less interested in the establishment and enterprise as concepts for sampling. For example, the survey was interested in targeting individual residential buildings, which are not located on the SBR. So it employed a strategy of sampling the building owner and then creating a roster of their buildings during collection. Likewise with individual primary schools: the survey was interested in the schools themselves, whereas the SBR only has the school boards (which are the administrative units representing numerous schools at the local level) on the frame. For its next iteration in 2019 SCIEU will be using the Building Register and its attributes to draw its sample. This will allow it to skip having to roster buildings during collection. However, thanks to the link between the building units and the Business Register for non-residential buildings, as well as the possibility of using auxiliary information to link the residential building back to the SBR, SCIEU will still be able to pull in the SBR attributes for processing and dissemination of its data, as required.

## **Annex B: Permissions and Clearance Levels**

### *Permissions*

C - Create Permission - allows for someone to author or add an item (ex. create a query)

M - Modify Permission - User is allowed to modify existing item or item(s)

R - Read Permission - Allows user to read information on the page

D - Delete Permission - Allows a user to delete item(s) if required.

A - Access to Function - Function is visible to the user and accessible.

E - Execute Permission - Allows user to perform action on the page (ex. upload and overwrite Help Files, Commit Sample Extractions)

### *Clearance Levels*

Determines view of data based on the data sensitivity and role. It outlines each attribute available in the Base Registers. Currently there are four categories of clearance as follows:

- CLF - Full Clearance
- CLM - Medium Clearance
- CLL - Low Clearance

- CLMin - Minimum Clearance

An example of how the clearance levels would determine data accessibility would be that a user with a persona with Medium Clearance would be able to see the day/month/year of the date of birth of a person on the population register (keeping in mind the name of the person would not be exposed, but simply a statistical number), whereas a user with a persona with Low Clearance would see month/year only and a user with a persona with Minimum Clearance would see year only.

Some of the various personas that have been determined so far are:

- Data Source Analyst
- Quality Assessment Specialist
- Matching Expert
- Resolutions Clerk
- Register Manager
- IT Support Technician
- Methodologist
- Subject Matter Expert
- Sampling Methodologist
- Analyst