Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

Record linkage techniques to identify 2021 Canadian Census dwellings in the new Statistical Building Register

by Martin Lachance, Tanvir Quadir and Darryl Janes

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2022
Data Disaggregation: Building a more representative data portrait of society

Record Linkage Techniques to Identify 2021 Canadian Census Dwellings in the New Statistical Building Register

Martin Lachance, Tanvir Quadir, and Darryl Janes[1]

Abstract

The reconciliation of 2021 census dwellings with the new Statistical Building Register (SBgR) presented linkage challenges. The Census of Population collected information from various dwelling types. For a large proportion of the population, mailing addresses were at the centre: they were used for reaching out to people and collected as contact info. In parallel, the register environment has been evolving. The agency is transitioning from the Address Register (AR) to the SBgR holding both mailing and location addresses, while also covering non-residential buildings. The reconciliation was conducted using a combination of systems, notably the new Register Matching Engine (RME) for difficult cases. The RME holds an interesting range of sophisticated string comparators. A deterministic linkage approach was used, while incorporating some data knowledge like the entropy. Through metadata, the matching expert could also reduce the amounts of false positives and false negatives.

Key Words/Phrases: Metadata; String comparator; False positive; False negative.

1. Introduction

Statistics Canada collects information through a census every five years. It also maintains a system of registers that are updated regularly to complement the census information. The Address Register (AR) (Statistics Canada, 2019) is one such register used to support census and survey operations within the agency. Its focus has been on the collection and storage of address and contact information for private (Statistics Canada, 2022) and collective (Statistics Canada, 2022) residential dwellings. After the 2021 Census of Population, the AR was discontinued in favour of a new Statistical Building Register (SBgR) (Gagné et al., 2018). These registers will be discussed in Chapter 3.

As the SBgR is being developed, each building is assigned information about its location. There are several types of addresses in Canada, which makes the storage and linkage of addresses a challenge. Most addresses, especially in urban areas, are civic style with a street number and suffix, street name, street type, street direction, and unit number if applicable (e.g., “203 - 12 Main Street North”). However, some rural addresses use other methods such a description (e.g., “White house with green mailbox”), an intersection (e.g., “Highway 4 and Cherry Street”), or a mile marker (e.g., “KM 1606 Alaska Highway”). Many owners of rural properties simply provide mailing information such as post office boxes (e.g., “P.O. Box 205”) or rural route numbers. The challenge with post office addresses is that they do not indicate the property location and they are subject to change. In the prairie provinces of Manitoba, Saskatchewan, Alberta and British Columbia’s Peace River region, some rural addresses are identified through a Quarter-Section-Township-Range-Meridian (QSTRM) system (e.g., “SW-21-24-8-W3”). And in some far northern communities, addresses may be identified simply by building numbers (e.g., “Building B42”) without a street name.

The 2021 Census of Population is a great source of residential address updates. It covers both rural and urban areas in all provinces and territories. Following the SBgR initialisation, census dwellings were considered as the first source of updates for the SBgR. The novelty of both the register and the Register Matching Engine (RME) were key elements

[1]Martin Lachance, Statistics Canada, 170 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, martin-p.lachance@statcan.gc.ca; Tanvir Quadir, Statistics Canada, 170 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, tanvir.quadir@statcan.gc.ca; Darryl Janes, Statistics Canada, 170 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, darryl.janes@statcan.gc.ca

in the reconciliation process. This paper will briefly discuss the Canadian Census, AR, and SBgR registers. The tools and linkage methods used to perform the reconciliation will then be presented. The emphasis of this paper is on the methodology, more than the results.

2. The 2021 Census of Population

The Statistical Building Register (SBgR) is comprised of residential and non-residential buildings in Canada. It requires up-to-date address information for planning, designing, and execution of many statistical operations conducted by Statistics Canada. It contributes to survey frames and is used for data collection and data validation. Among the sources used to update the SBgR, the census was deemed a major contributor due to its significant impact on coverage, in particular for areas not well covered by administrative files. Some residential dwellings that should be represented on the SBgR cannot be identified from administrative sources; they are only found during census operations, especially in rural areas or in small multi-unit buildings. As part of the SBgR update and evergreening process, the 2021 Census of Population was linked with the SBgR to improve its coverage and update building and building unit attributes.

In order to perform a linkage between the 2021 Census and the SBgR, a list of census dwelling addresses was put together. Two major sources were combined: the Master Control System (MCS) and the Response Database (RDB). The MCS is a set of pre-collection addresses initialized from the Address Register (AR) that accounted for about 98.53% of the addresses. The RDB is a set of post-collection addresses containing responses entered by respondents, either electronically or on paper, and covering an additional 1.46% of the addresses. Of note, the 2021 Census Edit and Imputation (EI) universe is the final valid set of dwellings. It confirms the existence and validity of MCS or RDB addresses. About 0.01% were fringe cases that did not belong to either source (MCS or RDB), so they were not considered in the linkage. The snapshot of the Census contained approximately 16.7 million dwellings. There are three categories of census dwellings: private dwellings (97.42%), primary collectives (0.23%), and dwellings associated with collectives (2.35%).

The mode of collection in the Census is important as it could affect the completeness and accuracy of the addresses, and consequently the linkage strategies and linkage rates. Three primary collection methods were used in the 2021 Census of Population: mail-out (86.10%), mail-out-and-drop-off (MODO) (6.07%), and non-mail-out (list leave, canvasser, reserve, seasonal) (7.83%) (Ha et al., 2022). The definition of these categories is associated to the questionnaire distribution within a collection unit (Statistics Canada, 2022).

There were regional differences in the addressing. Across Canada, approximately 95 % of addresses are civic-style. Other variations are regional or found in rural areas, such as: QSTRM in western Canada, descriptive addresses often seen in some eastern provinces, and a building numbering system specific to areas in the northern territories.

3. The Registers

With respect to location and address information, Statistics Canada is transitioning from an Address Register (AR) to a Statistical Building Register (SBgR). The AR had been used within Statistics Canada for about three decades. It started with the use of telephone files to support surveys. During this period, it stored various information for residential and collective dwellings in Canada. This included mailing and location addresses, some geographical information, and phone numbers associated with the dwelling. It contained postal codes as well as municipality names as determined by Canada Post Corporation. The AR did not store any personal information about the occupants or owner of buildings at the address.

Each dwelling in the AR was assigned a unique address identifier, called AR_UID. Indeed, the AR was a register of addresses. The identifier could have represented a single dwelling residence, but it could have also referred to the individual units of a multiple-dwelling residence such as an apartment building or condominium. A building with 100 residential units, apartments, or suites would have 100 AR identifiers.

Until 2021, the AR was regularly updated with new information and addresses in areas of growth. The evergreen nature meant that it played a vital role for survey operations at Statistics Canada. The Canadian Census used it to

determine methods of collection and the addresses for questionnaires to be distributed. After a census, the AR would be re-initialized with content from the census.

With the evolution of technology, the SBgR was developed as a replacement for the AR due to the growing need of information and for a better way to link to other registers at Statistics Canada, such as the Business Register (BR). The SBgR started development prior to the 2021 Census. It will be the key register of building and address information for surveys going forward after the 2021 Census, along with preparing the 2026 Census of Population.

Unlike the AR that stored information based on an address, the SBgR was designed to store information for a physical building structure while containing information from the AR. In fact, nearly one year prior to the 2021 Census Day, the SBgR was initialised with selective administrative files, most of which were already used for maintaining the AR. The initial version of the SBgR was linked to the AR to assess its coverage. The SBgR contains the building name, type, and size. The register also contains geographic attributes related to a building's location, such as census geographies (Statistics Canada, 2022), coordinates, and QSTRM if applicable. Another advantage to the SBgR is that it will contain buildings that are residential, business-oriented, and mixed-usage, whereas the AR was limited to residential addresses.

Within each building, there will be one or more building units. For example, an apartment building with an office and 20 apartments will have 21 building units. The SBgR stores information about the building unit, such as known contact information, its status, and its address.

The transition from the AR to the SBgR was facilitated using the 2021 Census. First, a traditional reconciliation took place between census dwelling addresses and the AR. A snapshot dataset holding AR and census information was then created as usual even though the AR was not re-initialised. Afterwards, the reconciliation of the 2021 Census dwellings with the SBgR took place. Census dwellings were directly linked to the SBgR, while also using their connection to the AR to take advantage of the AR-SBgR reconciliation. The reconciliation process between the 2021 Census of Population dwellings and the SBgR is discussed in the next chapters.

4. Reconciliation

There are two important reasons to reconcile the 2021 Census of Population dwellings to the Statistical Building Register (SBgR). The first is to ensure that the SBgR has excellent coverage of residential buildings and building units across Canada. Secondly, census and survey users need the best connection possible between census dwellings and the SBgR.

The reconciliation linkage was processed by province/territory, given the previously mentioned various addressing systems used across Canada. Once the linkage between the census dwellings and the SBgR was completed, unlinked census dwellings with sufficient location information that could not be found on the SBgR were "birthed" or temporarily added as new records on the SBgR until they have been confirmed. Unlinked census dwellings without sufficient information were to go through a process to identify the best geographic location possible (province, city, or lower level). This step is done because census users still need to know where these dwellings are geographically located, at a minimum. The geographic location identification process is not described in this document.

4.1 The systems

Statistics Canada uses two main systems to link addresses to the SBgR: the Address Processing Engine (APE) and the Register Matching Engine (RME).

The APE has been used for many years to perform address linkages of administrative data to the Address Register (AR). The emphasis of the system is on heavy pre-processing of data (i.e., parsing, standardisation) and the focus of linkage is mainly on groomed attributes matching perfectly. Linkages are done by street identifiers such as street name, type, and direction, civic numbers and suffixes, unit or suite numbers, municipality names and postal codes for the purpose of mailing, census geographies such as collection units and census subdivisions, and telephone numbers. Telephone numbers were not used in the SBgR reconciliation process when using the APE for linkage.

The APE is a system of SAS procedures that first identifies and parses the components of an address if necessary. For example, “Unit 201, 129A Saddle Road East, Red Fish, Alberta, T8G7K9” would be parsed into its components where Suite = “201”, CivicNumber = “129”, CivicSuffix = “A”, StreetName = “Saddle”, StreetType = “Road”, StreetDirection = “East”, Municipality = “Red Fish”, Province = “Alberta”, and PostalCode = “T8G7K9”. It then standardizes the components (e.g., “Road” becomes “RD” and “Alberta” becomes “AB”). It also applies a transformation of the street name and creates a derived variable to make it easier for matching. This transformed, coded street name is called a Road Attribute Search Key (RASK).

After transforming the address, the APE iteratively attempts to match the address components to the PCODE database containing Canada Post Corporation information, as well as Statistics Canada’s National Geographic Database (NGD) information. If the address is located, then the census geographic attributes are usually known for that record. Finally, the address and its geographic attributes are matched to the AR. The final output holds the parsed address information, various geographic and street information, match quality indicators, and the AR_UID.

Civic addresses can come in a variety of ways and formats. Abbreviations, punctuation, alternate municipality or street names, and ways of writing the unit or suite can cause difficulties for the parser. However, the APE performs very well at processing most of those types of addresses. It is also very good at processing addresses with postal codes.

The APE is a valuable tool for linking civic addresses to the SBgR, particularly those addresses found on the PCODE database. It was used to perform the linkage of most of the 2021 Census addresses to the SBgR very quickly. Unfortunately, the APE is not as successful at handling other types of addresses such as descriptive addresses, QSTRM, or northern community addresses. It also has difficulty when the address components have misspellings, truncation, or additional words (e.g., “129 Saddle Road Main Floor”). A second tool, called the *Register Matching Engine* was then used to perform reconciliation for the remaining cases that could not be linked by the APE.

The RME is a system that resides on the Statistics Canada Cloud. It has been developed specifically for the matching of administrative data to the system of registers, including the SBgR. It has a toolbox of sophisticated string comparators with built-in efficiencies to find match pairs. It also has metadata-driven tools to control the number of false positive and false negative pairs.

An advantage that the RME has over the APE is that it does not rely on the PCODE or NGD databases to do the linkage. It can link addresses directly to the SBgR without concern of the postal status, and that is why it is a useful tool for performing linkages of location addresses. The RME can handle both, mailing and location addresses. It was used to process the leftover civic cases that the APE was unable to link – specifically those addresses that may not be on the PCODE or NGD databases, or may have some content that could be better handled by a more robust matching strategy and the toolbox of comparators. Section 4.2 will discuss the RME in more detail.

4.2 The Register Matching Engine

The Register Matching Engine (RME) used the hierarchical deterministic linkage SAS package called MixMatch (Lachance, 2014) as a foundation. MixMatch was developed internally at Statistics Canada. It formally started with a research project to handle various types of linkages such as for persons, addresses, or businesses. It has been used in linkages for agriculture frames and property ownership, among others. The software strength resides in the sophisticated string comparators and the metadata driven approach to maximise linkage, while preventing false links due to legitimate, but closely resembling names. However, MixMatch was not designed to handle large volumes of data. Some consideration was given to the choice of technology. A deterministic linkage approach for linking addresses that uses some of the beneficial APE features, while overcoming the APE deficiencies, was also needed. The RME is Java-based and runs on the Cloud. Most ideas from MixMatch were retained in the RME and the methodology of the software is still evolving with users’ needs.

The core steps of any linkage are the pre-processing of data, identification of potential links (i.e., candidate pairs) between two data sources, and the post-processing of links found. The post-processing consists of resolving links found by keeping the correct ones based on a series of criteria, regardless of whether a deterministic or probabilistic linkage approach is used. The RME is designed according to this flow.

The RME string comparator toolbox holds well known comparators such as the Levenshtein distance (Navarro, 2001), Jaro-Winkler distance (Jaro, 1989; Winkler, 2006) and Longest Common Subsequence. The latter is used as a foundation for more sophisticated comparators. A few comparators for numeric variables are being progressively added, notably the Euclidean distance necessary for calculating close geographical distances when dealing with coordinates in Lambert projection. Latitude and longitude coordinates can be converted and used in the comparator as well. A strong feature of the RME is its multi-word comparators. The RME currently holds families of conservative comparators and loose comparators. Conservative comparators will likely provide better quality links by reducing the number of false positive matches at the expense of increasing the number of missed links (false negatives). Loose comparators will provide more matches at the risk of allowing some false positive links. Multi-word comparators have already proven very useful when dealing with addresses, in particular when they are not parsed, parsed improperly or hold extraneous information.

Each comparator returns a quantitative score between 0 and 1, along with a descriptive score, called the P-R-I-C-E. The P-R-I portion indicates the level of agreement in sequence (Perfectly matching, otherwise Resembling, otherwise Included), whereas the C-E (Conflicts and Extra words) indicates the level of disagreement once the algorithm has ended. The P-R-I-C-E is very useful in accepting/rejecting links and breaking ties in address linkages.

Furthermore, the RME uses a metadata driven approach in its implementation and usage. For pre-processing, the RME user may provide utility tables to express a series of transformations. For example, a street type standardisation might occur if “STREET” becomes “ST”. Further to this, at the stage where potential links are identified, a user can use their expertise to set the RME linkage strategy. This can play a critical role when matching any name variable. The MixMatch concept “Exclusion Table” is used in the RME to “exclude” false links beforehand. The user specifies a list of pairs of values that look alike, but for which they do not want a link. For instance, (Martin, Marvin) could be such a pair when trying to match “Marvin Street” with “Martin Street”. In this case, there are two legitimate street names rather than spelling errors. Conversely, the user may want to link two essentially equivalent values that do not look alike. The user informs the RME to “include” the pair of values in the linkage through an “Inclusion Table”. For example, a rural area could be defined by two very different names. Another example is the amalgamation of municipalities, where citizens continue to use the former name. In order to get a match, the user would list the pair of equivalent names in an Inclusion Table. This approach helps to reduce the number of false negative links, the number of potentially good links from being discarded. Inclusion Tables can be also extended to matching variables different in nature, such as a town name with a rural postal code. In this case, a match on values would indicate that the addresses are in the same area. In this type of usage, the user has to be careful with the interpretation of a link.

Table 4.2-1
Comparing complete street information using a multi-word comparator

		Complete street information				
Pair of records	Source	Street name	Street type	Street Dir.	RME Score	RME P-R-I-C-E
1	Census	52328, RANGE RD 234 BOX 156	ROAD		0.60	P=3 (RANGE, ROAD, 234), R=0 (no typo match), I=0, C=0, E=4 (52328, RD, BOX, 156)
1	SBgR	Range Road 234				
2	Census	Blue Ribon Lane by the Lake			0.65	P=2 (Blue, Lane), R=1 (Ribon), I=0, C=0, E=3 (by, the, Lake)
2	SBgR	Blue Ribbon	LANE			

The identification of potential links is based on a series of rules. For example, we could look for a perfect match on civic number, a partial match on street name, type and direction, or perfect match on city. For each condition of a rule involving an administrative variable and an SBgR variable (e.g., street name, type and direction all combined), the RME applies the same sequence of verifications. First, if there is a missing value on one side, a decision is made on the condition (passes/fails). If no values are missing, the RME checks if an Exclusion Table is being used. In that event, if we have a hit on a pair of values in the table, then the condition fails. If no such table is used or we do not have a hit on a pair of values, the RME applies the comparator selected by the user. The comparator may handle typos and may make use of an Inclusion Table. For instance, the use of a multi-word comparator could link streets for each pair of records presented in Table 4.2-1. The RME is also efficient in making use of indexes on the SBgR to handle perfect matches. The 2021 Census of Population presented addresses with various types of issues: misplaced

information, scanning errors for paper questionnaires, partial information, etc. The RME was of great help to handle the tougher cases.

4.3 The linkage process

The reconciliation of the 2021 Census dwellings with the SBgR was done using an agile approach. In a first wave, a large portion of dwellings with civic-style addresses were linked to the SBgR. A first snapshot of the SBgR was then made available to users. Next, in a second wave that led to a second snapshot, more complicated techniques were used, with the help of the RME, to try to find the still-unlinked civic-style address dwellings on the SBgR. The second wave also took care of the other types of non-civic addresses (e.g., QSTRM, North, descriptive) and collective dwellings (e.g., seniors home, jails, etc.). In this document, only the resolution of civic-style addresses is presented.

The first wave consisted of a combination of methods, with both the APE and the RME being used. The APE was used first to try to link as many dwellings as possible to the SBgR rapidly. At wave 1, direct links between the census dwellings and the SBgR were found, but some indirect links were also found. Direct links consisted of linking addresses directly to addresses on the SBgR, whereas indirect links consisted of linking to the SBgR through the AR. Indirect links were possible because of the connection between the census and the AR from the traditional reconciliation that took place between Census dwelling addresses and the AR, and then because of the connection between the AR and the SBgR from the SBgR initialisation. Direct links were considered to be of higher quality in the resolution process than indirect links.

Among the direct linkage methods, the APE was used to link census dwelling standardised addresses to the SBgR to try to identify building units. For some records where it failed due to address parsing issues, raw addresses were used in a second pass with the APE. Direct linkages also included the use of the phone number on the SBgR at the building unit level. When it matched the census dwelling's, the geography was compared between the census and the SBgR. A linkage strategy using the RME completed the direct linkages to the SBgR. The strategy was applied on the dwellings not linked by the second pass of the APE to take advantage of RME sophisticated tools when dealing with tougher cases. The remaining linkage methods provided indirect links. Some methods made use of phone numbers and some involved extra linkages using the APE. They were considered second, third and fourth quality level. Canada Post addresses from Point of Call (POC) files were an input source for both the AR and SBgR. Such addresses carry their own identifier that is stable over time, which strengthens the existing AR-SBgR connection obtained through linkage. A total of six indirect linkage methods were used.

In the resolution process for wave 1, each individual link was ranked according to the number of linkage methods providing the link, then the quality of the link. Following the ranking, a decision was made on the top link listed. A link of quality 1 found by two or more methods was considered a strong link. Weaker links required extra verifications. In this approach, the identification of the proper building was the first step. Next, links in which the building could be confirmed went through a similar validation process to try to confirm the building unit. In some instances, only links to the building could be validated. Throughout this two-step validation process, potential duplicates on the SBgR were identified via multiple links between census dwellings and SBgR buildings. They were put aside for review. Pairs of potential duplicates on the census side were also identified, for instance, when multiple census dwellings linked strongly to the same SBgR building unit. However, as opposed to the SBgR, the census data was frozen in time and potential census dwelling duplicates could only be flagged. The resolution of many-to-one and one-to-many links involved complex rules, while taking into account that some streets are known by more than one name, or aliases. More than 15 million civic-style address census dwellings were linked to a building unit at wave 1. A little over half a million outstanding census dwellings were pushed to wave 2.

The second wave of the reconciliation was done using only the RME. This wave targeted the most difficult civic-style addresses. Linkage strategies had to be more elaborate than the ones used at wave 1. A two-step approach was used this time. Instead of linking directly to the building units right away, attempts were made to find the proper building first, then another linkage strategy was used to find the proper unit inside the building.
Once links between census dwellings and SBgR buildings had been found, the resolution of those links was undoubtedly the most challenging part. Rules were applied, relying on multiple variables, each with various levels of matching precision.

First, for each link obtained, each pair of variables being compared (e.g., census street names and SBgR street names) was scored through a comparator and a P-R-I-C-E was obtained. For instance, link #1 could score 0.67 on the street name comparison, while link #2 could also score 0.67, but with a different P-R-I-C-E. Using this descriptor, we reject the links involving Conflicts as they are mainly bad links (e.g., "Blue Jays **Street**" vs "Blue Jays **Lane**"). If we were much stricter in accepting pairs, we would also reject links that have the most noise, represented by counts in the R-I components of the P-R-I-C-E. This was especially true if the Exclusion Table was significantly incomplete and resulted in linking on false typos (e.g., "**Annie** Street" vs "**Angie** Street"). Finally, Extras may represent added precisions on address components with minimal negative impact. These were accepted when matching full street names. The quantitative score (here, 0.67) is calculated using a ratio of the weighted agreement portion (P-R-I) over the weighted P-R-I-C-E components.

Although a deterministic approach was used for linking census dwellings to the SBgR, some data knowledge was introduced. This played a role in the second wave. Each variable entering a resolution rule was weighted using a measure of *entropy*, which is an attribute of a variable. A variable with a high diversity of values has great discriminatory power, which translates into high entropy. Variables with high entropy were used in the matching rules. However, spelling errors can increase the number of distinct values of a given variable, which inflates the entropy. In order to minimise such bias, the entropy of variables is calculated from the SBgR, which is considered the "cleaner" reference. Furthermore, the frequency of a specific value occurring within a variable should be used in combination with the entropy. For example, the street name variable has high entropy. However, some values like "Main", in "Main Street", have high frequencies because there is a main street in most municipalities. For the value "Main", the discriminatory power of the street name brought by the entropy is canceled out. Unfortunately, frequencies could not be implemented in time for the second wave of reconciliation. Only the entropy was used.

After having taken all of the above into consideration, the strength of a link for civic-style addresses mainly relied on three key components: (1) a perfect match on the civic number, (2) some matching street information and (3) a match on at least one geographical area. A sequence of rules were applied, classifying links from strongest to weakest. Finally, a customised multi-link resolution was performed, taking into consideration the strength of the link.

Once a link to a building was confirmed, attempts were made to find the proper unit inside the building on the SBgR. The first attempt was to find a perfect match on suite or apartment numbers. Unfortunately, as expected, we did not have much success as we were dealing with the tougher cases. In particular, small multi-unit buildings are very difficult to resolve. They are usually existing buildings that have been subdivided into smaller units. For example, civic number "20" is now either 20 and 20A, or 20A and 20B, and so on. These units often tend not to show up on administrative files. Most appear to be found only by census field operations. They are also very difficult to resolve in an automated fashion. Ideally, for such complex cases, if we had had person-level information at the SBgR building unit level, we might have been able to link census dwellings to SBgR building units. But it was not possible when the reconciliation was conducted. Another problem resulted from the misplacement of civic suffixes or unit/suite numbers. For example, if the 'A' in '20A' is a civic suffix, then the address represents a building on the SBgR. If it is a unit, then the address represents a unit within a building. A special three-step strategy was developed to handle this. First, we combined the civic number, civic suffix and suite number together into a single variable. This was done on both sides: census and SBgR. Next, we would split the information into two new variables, for both: numbers only in a first variable, the rest in the other. Finally, we would request a match the character and the numeric variables, using multi-word comparators to handle extra pieces of information. If this method only brought about 10,000 building unit level links out of the new 66,700 building level links obtained, our confidence was high in the quality of the links.

5. Conclusion

The simultaneous transition to the new Statistical Building Register (SBgR) and a new linkage system, the Register Matching Engine (RME), offered challenges, learning opportunities, but mainly permitted the development of techniques to improve the quality of the frame. First, while developing a linkage strategy, it is essential to explore and understand the data. The 2021 Canadian Census dwellings contained a collection of addresses not seen on administrative files. Understanding the SBgR data structure was also crucial. Secondly, linkages involving addresses can be far more complex than linkages for businesses, which themselves are considered more difficult to handle than person-level linkages. This was the biggest lesson learned in doing the reconciliation. For person-level linkages, multiple person attributes and other variables are usually available, including address components or summarized

locational information. For business linkages, a shortage of variables is frequent, where often only a business name is available, without factoring in repetitions due the multi-level organisational structure of businesses. But there are multiple ways of expressing an address in Canada and they hold the most “names”. In fact, for all types of linkages, the “names” are the most challenging variables. At a minimum, there are two names when dealing with addresses in Canada: street name and municipality name. Fortunately, legal constraints at the provincial and city level prevent repetitions. In the end, the RME offered flexible and powerful tools for linking addresses, especially names. It offered multi-word comparators and a metadata driven approach to control amounts of false positives and false negatives.

References

Gagné, P., Pignal, J., Quadir, T., and Wolfe, C. (2018), “Towards a Register-Centric Statistical System: Recent Developments at Statistics Canada”, paper presented at the Statistics Canada International Methodology Symposium 2018, Ottawa, Canada.

Ha, B., Lee, M., Mayda, M., and Wong, J. (2022), “2021 AR Snapshot and Reconciliation Project”, unpublished report, Ottawa, Canada: Data Integration Infrastructure Division, Statistics Canada.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, 84(406), 414-420.

Lachance, M. (2014), “Useful Functionalities for Record Linkage”, paper presented at the Statistics Canada International Methodology Symposium 2014, Ottawa, Canada.

Lachance, M. (2014), MixMatch 1.2 – User Guide, unpublished report, Ottawa, Canada: Statistics Canada.

Navarro, G. (2001), "A Guided Tour to Approximate String Matching", *ACM Computing Surveys (CSUR)*, 33(1), pp. 31-88.

Statistics Canada (2019), “Surveys and Statistical Programs – Address Register (AR)”, an article published in the Statistics Canada website. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1260700

Statistics Canada (2022), “Directory, Census of Population, 2021 – Private Dwelling”, an article published in the Statistics Canada website. https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/Definition-eng.cfm?ID=dwelling-logements005

Statistics Canada (2022), “Directory, Census of Population, 2021 – Collective Dwelling”, an article published in the Statistics Canada website. https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/Definition-eng.cfm?ID=dwelling-logements002

Statistics Canada (2022), “Census of Population – Census Geography”, an article published in the Statistics Canada website. https://www12.statcan.gc.ca/census-recensement/2021/geo/index-eng.cfm

Winkler, W. E. (2006), "Overview of Record Linkage and Current Research Directions", *Tech. Report. Statistics #2006-2*. Statistical Research Division, U.S. Census Bureau.