The linkage of micro census data to vital records:
New perspectives for longitudinal studies

Hélène Vézina
Projet BALSAC, Université du Québec à Chicoutimi

Marc St-Hilaire
Centre interuniversitaire d'études québécoises, Université Laval

Claude Bellavance
Centre interuniversitaire d'études québécoises, Université du Québec à Trois-Rivières

Paper presented in the session: "Using micro census data: recent developments and new perspectives" at the First Conference of the European Society of Historical Demography (ESHD) www.eshd.eu

ABSTRACT

Some years ago, we initiated the development of a linkage program relying on systematic and automated tools and procedures to match microdata from the Canadian censuses to those from Quebec civil records. The development of this program is at the heart of the construction of the Integrated Infrastructure of the Quebec Population Historical Microdata (IQPM) which will integrate all available historical microdata on the Quebec population dating back to the beginning of European settlement, into a set of relational databases. In this paper, we will describe the development and implementation of the linkage program which contains three modules: one for linking census data to BALSAC families; one for linking BALSAC families to census data; and one for linking sets of data from distinct censuses. We will illustrate how it works by providing examples of linkage results and estimates of success rates from datasets on the city of Trois-Rivières and on the Saguenay region for seven modern nominal censuses (1852 to 1911). We will also emphasize how the linkage of census and vital data can make a major contribution to the construction of census-based longitudinal datasets. When completed, the IQPM will comprise data from these seven censuses on close to a million individuals belonging to 161,000 distinct households from two urban settings and three regions mixing rural and urban environments. All these data will have been linked to corresponding vital event records from the BALSAC database. To conclude, we will give an overview of the research opportunities that will emerge from the matching of the two types of data both for the social and the biological sciences as this vast array of biographical information will permit studies based on individual trajectories situated within families, households and communities and examined from a multigenerational perspective.

Introduction

The Quebec province in Canada has the advantage of an exceptional and remarkably well-preserved documentary resource with the wealth of birth, marriage and death certificates recorded by the Catholic Church since the early days of French settlement in the 17th century. More than 40 years ago, two major projects, the Programme de recherche en démographie historique (PRDH) and BALSAC, initiated the digital transcription and linkage of these parish records. More recently, a good share of Quebec data from the Canadian historical censuses prior to 1921 have also been digitized and formatted in databases by the Centre interuniversitaire d'études québécoises (CIEQ) and by the PRDH.

These two sources of data are fundamental for the study of two major categories of social facts. Information from vital events fosters research on family and kinship, such as reproduction, union formation, family strategies, migration and intergenerational transmission In turn, nominal censuses make possible studies related to the household (size and composition, roles of members, residential patterns, occupational structures, income and education) as well as micro-scale economic strategies.

Using continuous data found in civil registration it is possible to conduct longitudinal studies using family reconstitution techniques while sophisticated multivariate analytical tools have been developed to investigate census data which is by nature cross-sectional. Each source also has its limitations: vital registration provides none or very little indication on the socioeconomic conditions of individuals and families and none on living arrangements and neighborhood; censuses do not provide information on the role of extended kinship in economic and residential strategies.

In order to overcome these limitations and build on the strengths of the two types of data to expand research possibilities, we initiated in 2010 the development of a linkage program relying on systematic and automated tools and procedures to match microdata from the Canadian censuses to those from Quebec civil records and to link census data together. We aim to develop a research infrastructure that will take advantage both of the power of family and genealogical files based on civil registration and on the wealth of census data and to open up new avenues of research for the study of historical populations.

In this paper, we will describe the development and implementation of the linkage program providing some insights on the challenges and issues involved in this endeavour. We will illustrate how it works by providing examples of results and estimates of success rates. We will also discuss the possibilities and limits of our approach emphasizing how the linkage of census and vital data can make a contribution to the construction of census-based longitudinal datasets. Lastly, we will give an overview of the research opportunities that will emerge from the matching of the two types of data both for the social and the biological sciences.

Population reconstruction projects in Quebec

The Registre de la population du Québec ancien (RPQA) was created in 1966 by the Programme de recherches en démographie historique at the Université de Montréal. Its construction was inspired by the techniques of family reconstitution developed by the French demographer Louis Henry. It contains the longitudinal linkage of the entire Catholic population of Québec from the beginning of French settlement in Canada in 1608 (first record in 1621) to 1799 comprising 700,000 birth, marriage and death records.

In 1972, another population database was initiated at the Université du Québec à Chicoutimi. In the first phase, family reconstitution was performed on the population of the Saguenay-Lac-St-Jean, a region located 200 kilometers north of Quebec city (see Figure 1), from the beginning of French Canadian settlement in 1838 to 1971 relying on the 660,000 birth, marriage and death records. Since 1989, a second phase has been ongoing to perform the digital transcription and linkage of marriage records for the whole province of Quebec. The work is now completed up to 1965 and the database contains 2.2 million linked marriage records allowing for the automatic construction of genealogies.

The complete census of Canada for 1881 as well as samples for all other historical censuses (from 1852 up to 1911 with the exception of 1861) are formatted in databases and publicly available. In Quebec, complete count for the cities of Quebec and Trois-Rivières for seven decennial censuses (1852-1911) and partial count samples for many Quebec regions have been digitized in the course of previous projects[1].

A few attempts to link Quebec civil and census records were previously achieved relying on manual work performed on small datasets (Charbonneau *et al.*, 1970; Gauvreau, 1991; Gossage 1999). In some instances marriages records were used to facilitate linkage across censuses (St-Hilaire and Marcoux 2006; St-Hilaire, 2009; Olson and Thornton 2011). In 2010, members of our group initiated a pilot project to evaluate the feasibility of designing a program to link census data to marriage records found in the BALSAC database (Gauvreau *et al*. 2010; St-Hilaire and Vézina, 2010; Vézina and St-Hilaire, 2011). The development of this linkage program is now at the heart of the construction of the Integrated Infrastructure of the Quebec Population Historical Microdata (IMPQ) which was financed in 2013 by the Canadian Foundation for Innovation.

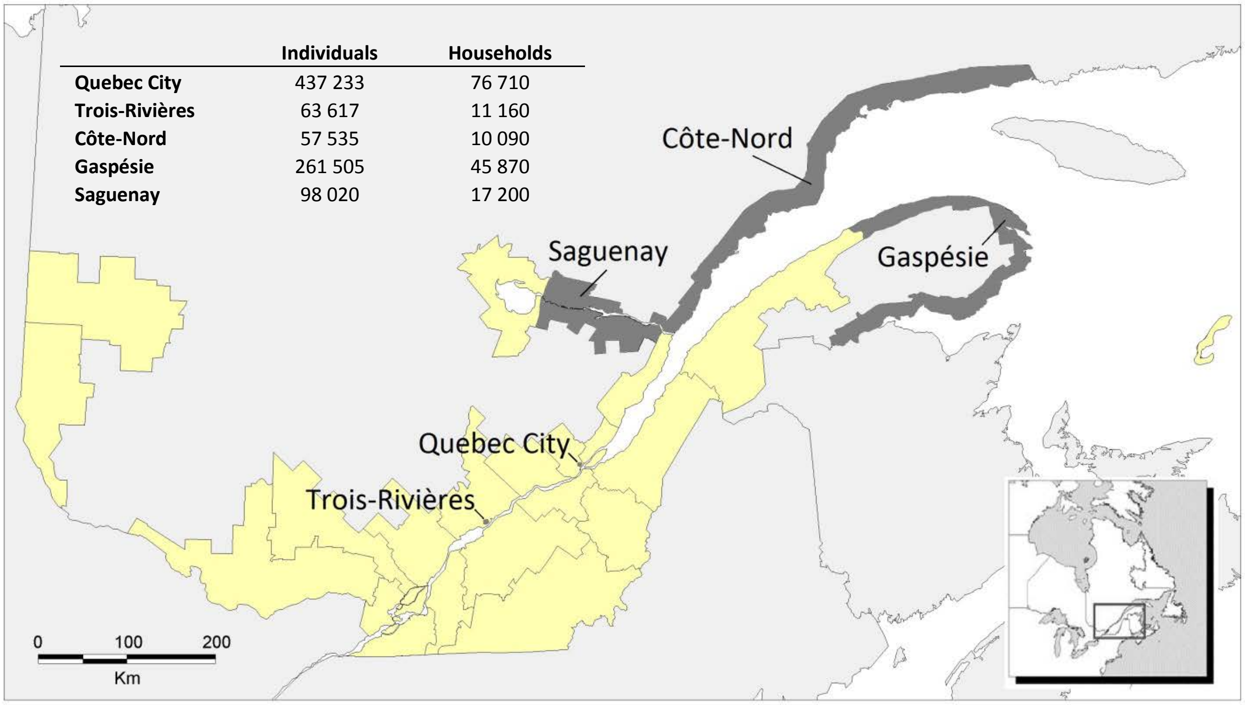
The IMPQ project

The IMPQ will integrate available historical microdata on the Quebec population dating back to the beginning of European settlement, into a set of relational databases. The objective is to preserve, highlight and develop this major historical and scientific heritage. The creation of the

[1] The 1881 census is available in the North Atlantic Population Project (NAPP) website (https://www.nappdata.org/napp/»). Data on Quebec City is available in the Population et histoire sociale de la ville de Québec project website (http://www.phsvq.cieq.ulaval.ca/index.php?p=accueil). The other datasets were constructed in the course of research projects and are not publicly available at this moment.

infrastructure will also facilitate integration of new data and development of new tools for linkage and analyses as well as promote training in relevant disciplines and collaborations both at the national and international levels (comparative research).

The construction of the infrastructure includes three components. First we plan a full fusion of the BALSAC and PRDH databases enabling simultaneous and continuous updating of the longitudinal data as well as integration and linkage of new vital records. The second part focuses on the harmonization of existing census data series and expansion of the geographical coverage to include two urban environments (Quebec City and Trois-Rivières) and three regions mixing rural and urban environments (Gaspésie, Côte-Nord and Saguenay). Together this represents 918,000 individual entries (see Figure 1). In the third component of the project, these individuals belonging to 161,000 distinct households are gradually linked to the BALSAC database and across censuses. The development of the linkage program which we present here is thus at the heart of the construction of IMPQ.



	Individuals	Households
Quebec City	437 233	76 710
Trois-Rivières	63 617	11 160
Côte-Nord	57 535	10 090
Gaspésie	261 505	45 870
Saguenay	98 020	17 200

Figure 1: Census data to be integrated in the IMPQ

The development of a program for the linkage of civil records and census data

The BALSAC linkage system of vital records as a starting point

From the start, we made the decision to use an approach based on the linkage procedures developed for the construction of the BALSAC database. The challenge was to adapt them to take into account the differences in the structures of the two types of data. We will first present briefly the BALSAC linkage system.

The BALSAC family reconstitution system, which processes vital data records, aims at grouping in the same file all mentions appearing in the records and referring to the same couple. Thus, the basic unit of information for linkage is the "couple mention" which contains four nominative elements namely the husband's name and surname and the wife's name and surname. All nominative information is processed to eliminate superficial name variations using FONEM, an automatic phonetization program (Bouchard *et al*., 1981a; Bouchard and Bouchard, 1981). As shown in Figure 2, FONEM can significantly simplify the name variations to be linked.

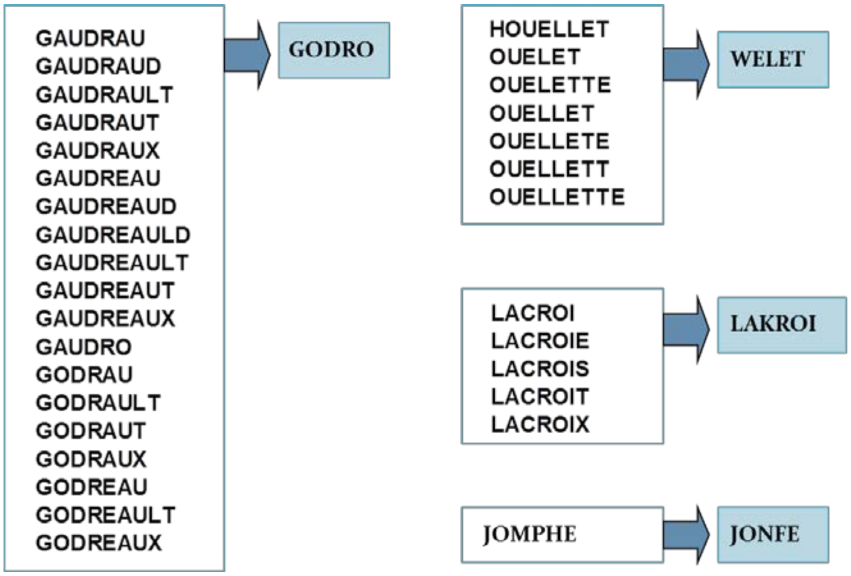


Figure 2: Examples of automatic phonetization with the FONEM program

Candidate couple mentions in BALSAC are created on the basis of at least two common nominative elements with the couple to be linked. These mentions are compared using three programs designed to detect and measure degrees and forms of similarity between two last names or two first names. ISG calculates a score based on the degree of similarity between nominative elements based on the position of same letters in names. INCL deals with truncated names by detecting suffix and prefix in names and deciding if one can be treated as being included in the other. ELM processes situation of multiple names and surnames and decide if two entities car be treated as equivalent or not. The decision making process relies almost exclusively on nominative information contained in the records but the coherence of dates in the family structure is also taken into account (Bouchard *et al.*, 1985).

The linkage operation consists of four different phases processing first the easiest links and moving progressively to the most difficult. The two first phases are performed immediately after the entry of a marriage record. The first one is entirely automatic: a couple is linked to an already existing family files only if all nominative information is perfectly identical and if there is only one potential candidate for the linkage. In the second phase, when the program has been unable to make the appropriate linkage within the established safety range**,** it provides a list of potential candidates for linkage and the entry clerk makes a selection if an unequivocal choice can be made. The third phase is performed at a later stage. It is also based on a list of candidates provided by the program and is operated by experienced staff who might use complementary sources such as genealogical repertories or websites to support their decision. Lastly, after these three stages, there remains a share of unsolved linkages which will eventually be submitted to a more in-depth investigation.

Linkage operations lead to the construction of a family file that includes all the records pertaining to a unique couple (their own marriage, the remarriage of a surviving spouse when existing, the marriage of their children[2]) (St-Hilaire, 1990). All links submitted to automatic linkage and those carried out at the stage of the computer-assisted manual linkage are immediately validated by automatic routines of coherences (for example: acceptable timespan between two events, chronological sequence of events, reported or calculated age, duplicate events or names)(Bouchard *et al*., 1981b).

Adapting the BALSAC method for census and civil records linkage

As mentioned in the previous section, the development of the program for the linkage of civil records and census data rests on the basic principles used at BALSAC for the linkage of civil records with nominative elements at the heart of the process. Essentially, the method remains the same and linkage is done using matching programs and comparison and decision algorithms developed at BALSAC. However, several adjustments were made to take into account differences in structure and content between the two types of data. Hence, before describing the method of linkage itself, the main issues raised by the differences in the two data sources and an overview of modifications brought to the BALSAC system are presented.

First, the structure of the basic grouping units (family vs. household) is different. The BALSAC family structure corresponds to the model of a "nuclear family" based on information found in civil registration events. Census household composition is determined by a survey at a specific point in time and is based on a residential criterion (housing). Several scenarios of household structure are possible: one or many single individuals, nuclear family (partial or complete), blended family, intergenerational family, etc. That being said, the similarities between the two types of data are real, since the household as a grouping unit often corresponds to a structure, if not identical, at least similar to that of a nuclear family.

Second, the quality of nominative information is not the same in both sets of data. Civil records are legal documents and priests, pastors, and other civil officers usually kept them with great care. They are very consistent trough time since there was little change in prescribed rules to record information (Bouchard and Larose, 1976). The context of census data collection is very different. It is a massive and complex operation conducted in a short period of time, at ten years intervals by people who were not always well-trained and sensitive to the cultural

[2] For Saguenay, family files also include spouses' death records, children's birth as well single children death record.

context of the people they enumerated (Bellavance *et al*., 2007). These factors impact directly on the quality of the data in general and of nominative data in particular.

Last but not least, whereas in the BALSAC database, first and last name are found for both spouses, in census data, the maiden name of the wife is almost always missing. As the whole linkage system relies on a couple mention comprising four distinct nominative elements, this represented an important issue for the development of our program.

Given these differences, some modifications were brought to the linkage program and new elements were introduced. One of the main changes in the method consists of a prior shaping of census data in order to extract nuclear families comparable to BALSAC families. To be considered for linkage, a household must contain at least two members since a minimum of three nominative elements is needed. Also, the head must be married or widow(er) as the basic unit of linkage is a family composed of at least two members.

To overcome the absence of the wife's maiden name in the households, it was decided to generate a new nominative unit of comparison (NUC) including the three other nominative elements available (father's last name, father's first name and first name of the mother) to which is added the first name of a child. It is then possible to compose as many NUCs as there are children mentioned. For each set of data, the program creates a table of nominative mentions whose number per household or per family varies according to the number of children.

Because of the changes brought in the composition of NUCs, it was necessary to relax sorting and matching criteria to bring more potential candidates in the phase of nominative pairing (see below). The adapted method of matching requires a single nominative element out of four to be perfectly identical between a NUC derived from the census and the one obtained from BALSAC. This also permits to overcome the difficulty to harmonize, standardize and compare nominative data from two different sources (civil registration and census). The relaxation of matching rules is however offset by the addition of a rating system of candidates which will be reported in the next section.

Another important change is the fact that elements of comparison other than nominative information are taken into account for linkage. Because of the variability in nominative information, it seemed important to involve other features in the process of comparison to support situations where the nominative context would be inappropriate. Thus, other elements found in the two sources were added for comparison, namely: residential data, occupations and information allowing for verification of concordance of dates and ages. These elements are not considered in the selection of household and family candidates, but are used in the comparison process and score assignment which quantifies the degree of similarity between the two sources. Hence, while linkage of civil records rests almost exclusively on nominative information, linkage of civil and census data relies on the comparison of both individual (nominative data) and contextual (family/household characteristics) elements.

Lastly, after performing some tests, we made the decision to exclude automatic linkage and to rely entirely on manual computer-assisted linkage. This means that all linkage decisions are taken by research assistants.

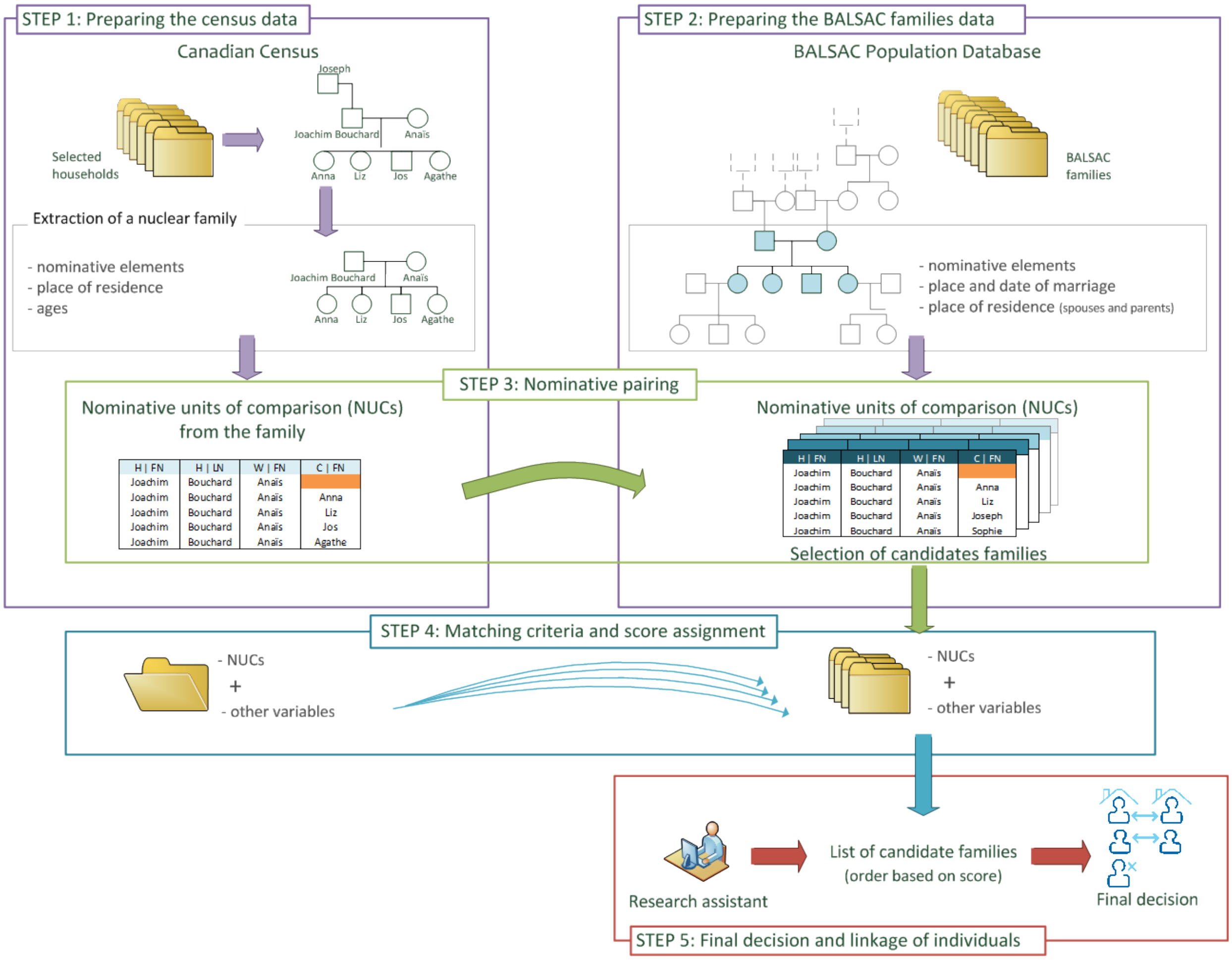
Linking census data to BALSAC families

The current version of the linkage program contains three modules: the first one for linking census data to BALSAC families; the second one for linking BALSAC families to census data; and one for linking households from distinct censuses. Here we present the linkage program using the first module to illustrate the whole process which is summarized on Figure 3. In the two first steps data from the census and from BALSAC are transformed into a common structure: the nominative unit of comparison (NUC) which contains the names of the spouses and is similar to the couple mention. In steps 3 and 4, data are sorted and paired to select potential candidates and a score is assigned to each candidate based on the selected matching criteria. In step 5, a final decision is made to proceed with linkage or not.

Step 1. Preparing the census data: To form the NUCs, nuclear families must first be identified and extracted from census households. More than one family can be extracted from a household. To compensate for the absence of the wife's maiden name in census data, the first name of each child is added to the NUC. Hence, there will be as many NUCs for a census family as there are children. The three-element combination containing the first and last names of the husband and the first name of the wife is also used and it is the only combination possible when a couple has no children recorded. Other members of the household are regarded as children of this couple if a) their surname is the same as the presumed father; b) their age is compatible with the age of the mother (minimum age difference of fifteen years); and c) they are single. In addition to the nominative data, other variables that will be used to link records and to calculate a score are kept (place of residence, ages).

Step 2. Preparing the BALSAC families data: Each potential candidate family from BALSAC is also organized in NUCs to provide a pool of candidates for each census family. First, a subset of families is extracted from the BALSAC database. These families are selected within a certain time interval based on the likelihood of having been enumerated at a given census. First and last name of husband and wife, name of children as well as other variables that will be used for linkage (place of residence, ages, date of marriage, occupations) are retrieved.

Step 3. Nominative pairing: Based on the programs of comparison of nominative similarity developed at BALSAC and described above, the application performs various sorting operations aiming at pairing each census family to candidate BALSAC families. The pairing criterion is that each NUC (from the census and from BALSAC) must contain at least one identical nominative element. A list of potential candidates for linkage with a specific census family is thus selected from the BALSAC families.


STEP 1: Preparing the census data
Canadian Census
Selected households
Joseph
Joachim Bouchard
Anaïs
Anna
Liz
Jos
Agathe
Extraction of a nuclear family
- nominative elements
- place of residence
- ages
Joachim Bouchard
Anaïs
Anna
Liz
Jos
Agathe
STEP 2: Preparing the BALSAC families data
BALSAC Population Database
BALSAC families
- nominative elements
- place and date of marriage
- place of residence (spouses and parents)
STEP 3: Nominative pairing
Nominative units of comparison (NUCs) from the family
Nominative units of comparison (NUCs)
Selection of candidates families
STEP 4: Matching criteria and score assignment
- NUCs
+
- other variables
- NUCs
+
- other variables
Research assistant
List of candidate families
(order based on score)
Final decision
STEP 5: Final decision and linkage of individuals

H | FN	H | LN	W | FN	C | FN
Joachim	Bouchard	Anaïs	🟧
Joachim	Bouchard	Anaïs	Anna
Joachim	Bouchard	Anaïs	Liz
Joachim	Bouchard	Anaïs	Jos
Joachim	Bouchard	Anaïs	Agathe

H | FN	H | LN	W | FN	C | FN
Joachim	Bouchard	Anaïs	🟧
Joachim	Bouchard	Anaïs	Anna
Joachim	Bouchard	Anaïs	Liz
Joachim	Bouchard	Anaïs	Joseph
Joachim	Bouchard	Anaïs	Sophie

Figure 3: The linkage of census data to BALSAC families

Step 4. Matching criteria and score assignment: Each of the BALSAC families selected in the pairing process is submitted to the comparison component of the program and a score is attributed to each potential match based on the degree of similarity with the census family being processes. While nominative data found in the NUCs still provides the most important linkage criterion, concordance of dates and ages as well as place of residence are also used in the comparison process. The goal is to take into account and make optimal use of the information available about the families in each dataset. Every element of the comparison receives a score weighted by its importance (for instance, logical concordance of dates is more important than similarity of children's first names) and the sum of these scores provides the total score.

Step 5. Final decision and linkage of individual: At this stage, the candidate BALSAC families are ordered according to their final score and manual work is involved to proceed with the linkage decision. In some instances, it might be necessary to consult external tools to confirm the choice. Based on the work presented in the section, we estimate that a final decision is taken to accept a candidate or abort the linkage attempt within two minutes on average and that depending on the region and on the census year, between 72% and 92% of the created links involve the candidate with the highest score. When a decision to link a census family to a BALSAC family is reached, the research assistant can also link individuals based on name, age and date consistency.

Linking families and individuals across censuses

The third module of the program allows for linkage across censuses. In the work we are currently conduction for the IMPQ project, linkage between two censuses takes place after both censuses have been linked to BALSAC using the first module. The whole linkage process is similar to the one described in Figure 4. In both censuses, nuclear families are extracted from household and NUCs are produced. Nominative pairing based on the NUCs lead to the selection of candidates from Census B for linkage to a specific family in Census A. Nominative and contextual information are used for score assignment and the final decision is taken by the research assistant.

There is however a prior selection in Census A of households depending if they have been linked to a BALSAC family or not. This selection process is displayed in Figure 4. It shows that in the preparation of data, some households will be processed and others will not based on the criteria described previously (at least two people in the household and the head being married or widowed). Among processed households, some of them will then have been linked to BALSAC while for others linkage will not have been successful. Direct linkage to Census B will be attempted on all households from Census A which have not been linked to a BALSAC family. Moreover, among linked households from Census A, some will be connected to a household in Census B through their linkage to the same BALSAC family but others will not. Direct linkage to Census B will also be initiated on this latter set of households.

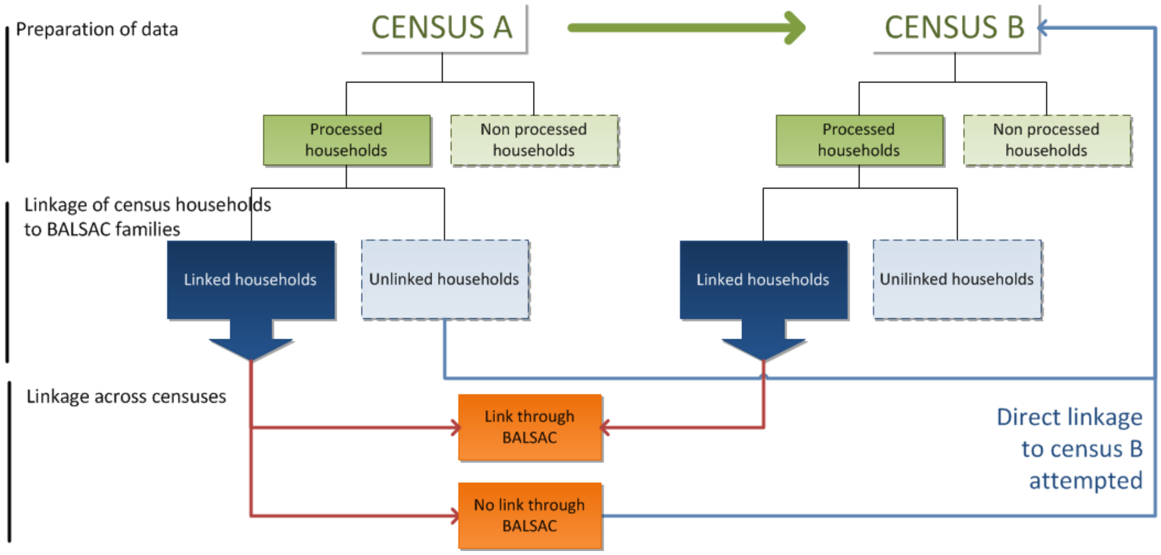


Figure 4: Linkage across censuses

Ongoing work on Quebec historical microdata

Table 1 shows the results of the linkage of census data from Trois-Rivières with the BALSAC families. For this city, only marriages are available in the BALSAC database and therefore families are composed of a couple and their married children. Linkage is completed for the seven historical censuses to be included in the IMPQ. Between 88 and 97% of households were processed meaning that a family could be extracted from the household data. The table also indicates that between 72 and 89% of households were matched to a BALSAC family. If we consider only processed households, this gives a success rate ranging from 81% for the 1861 census to 92% for the 1891 census. Among members of these linked households, between 59 and 73% were located in the BALSAC database. On average, each linked household contains from 5 to 7 people and between 3.5 and 4 of these individuals were linked to the BALSAC database.

Linkage results from one census to another show significant variability. Success rates are dependent mostly on the quality of nominative data and on household structure. For instance, the lower figures for 1861 and 1871 in Trois-Rivières are at least in part due to the fact that enumeration was based on self-declaration which might have lowered the quality of data.

Table 1 : Results of the linkage of census data from Trois-Rivières to BALSAC families

	Census year						
	1852	**1861**	**1871**	**1881**	**1891**	**1901**	**1911**
Total number of households	787	992	1714	1760	1700	2223	2993
Processed households	96%	93%	88%	94%	96%	96%	97%
Households linked to a BALSAC family	83%	76%	73%	84%	89%	84%	87%
Total number of individuals in linked households	4058	5367	6736	8079	8276	10455	14472
Individuals linked to a BALSAC family	63%	59%	66%	66%	71%	71%	73%
Mean number of individuals in linked households	6,3	7,1	5,4	5,5	5,5	5,6	5,6
Mean number of linked individuals in linked households	4,0	4,2	3,6	3,6	3,9	4,0	4,1

For the Saguenay region, the linkage operations are completed for the 1871 and 1881 censuses (Table 2). Since for that region we rely on the complete set of vital events (as opposed to marriages only), we expected higher success rates. The results are indeed quite spectacular: for the 1871 census, a family was extracted from 94% of households and 98% of them were linked to BALSAC which means that 92% of the total number of households was linked. For 1881, the figures are even higher with 94% of all households connected to a BALSAC family. Among individuals living in a linked household, 93 and 94% respectively could be found in BALSAC. This clearly demonstrates that availability of birth and death records allows the program to produce optimal linkage results.

Table 2: Results of the linkage of census data from Saguenay to BALSAC families

	Census year	
	1871	**1881**
Total number of households	1994	2074
Processed households	94%	96%
Households linked to a BALSAC family	92%	94%
Total number of individuals in linked households	11399	13508
Individuals linked to a BALSAC family	93%	94%
Mean number of individuals in linked households	6,2	6,9
Mean number of linked individuals in linked households	5,8	6,5

Results of linkage across censuses are displayed in Table 3 for two pairs of censuses in Trois-Rivières and one in Saguenay. As shown in Figure 4, there are two ways to link a family from a given census to a family in the next one. The first one is through BALSAC when the families from the first and second censuses have been linked to the same BALSAC family. The second one is direct linkage to the next census for families which could not be linked through BALSAC. About 60% of Trois-Rivières families from the starting census could be linked to the next one and about 70% in Saguenay. The higher success rate for Saguenay is explained by a higher proportion of families linked through BALSAC (58% vs 42 and 48% for Trois-Rivières). Results obtained from direct linkage to the next census are very similar for the three pairs of censuses.

Table 3: Results of linkage across censuses for Trois-Rivières and Saguenay

	Trois-Rivières				**Saguenay**	
	1852>1861		**1891>1901**		**1871>1881**	
	N	**%**	**N**	**%**	**N**	**%**
Processed households	754		1634		1875	
Linked households						
through BALSAC	317	42,0	778	47,6	1094	58,3
through direct linkage to next census	124	16,4	203	12,4	239	12,7
Total	441	58,5	981	60,0	1333	71,1

Discussion

Possibilities and limits of the approach used for linkage

Our linkage approach is based on the comparison of two sets of nuclear families: the first one reconstituted from vital events in the BALSAC database and the second one extracted from households in Canadian historical censuses. It relies mostly on nominative information but contextual information on family members' characteristics is also used. Our objective is to link as many families and individuals from the census population as possible with the most efficient method in terms of reliability and efficiency using a three module program developed by our group.

Because of the differences in the two data sources, the program does not allow for a perfect match, but rather a match with various degrees of similarity between BALSAC families and households or among households. Family structures from BALSAC may be incomplete (in fact except for Saguenay they contain only couples and their married children) and those extracted from households can be very different (recomposed families, widowhood, household members outside the family). However the program can still deal with situations where the nuclear family cannot be fully restored as long as the structure is not completely fragmented. The success rate is obviously dependent on the degree of alteration of the structure.

From the viewpoint of census data, the linkage with civil records, especially marriages, is of tremendous help to link censuses and build census-based longitudinal datasets. Through the BALSAC family file, it is possible to connect a single male child within his parents' household and the married man appearing as a household head ten years later Marriage records do not only confirm links based on individual's nominative information, but are also necessary to solve a large number of homonymy problems. In addition, due to the loss of maiden names after marriage, they are the only source which could allow the linkage of a girl within her parents' household with the woman being a member of a couple in a subsequent census.

Considering couple (or couple and child) mentions as the unit of comparison, preferably to a unit based on individual information, minimizes the negative impact of nominative variations on linkage performance and increases success rates. Using this approach, all households in a given census will be processed for the purpose of a linkage, except households for which it is impossible for a member to compose nominative mentions, and households whose members are all single (such as religious communities, hospitals, etc.). Thus, our linkage procedures will yield longitudinal data biased in favour of couples and families.

Also, we use the BALSAC population database as the reference data for vital events. The advantage of using BALSAC is that it covers the whole territory of Quebec for the entire period for which nominal censuses are available and it is fully family-structured. Except for the Saguenay region, it contains at the moment only marriages so families are composed of parents and their married children in most instances. This also plays in favor of more linkage for couples and families.

Some possible biases are brought about by our linkage method. There as many nominative units of comparison in a family as there are children therefore the probability of linkage increases with the number of children. We rely on places of residence not for the selection of candidate families but in the calculation of the score to rank the candidates. Thus stable families might have a certain advantage over those who move.

The program enables the linkage of individuals. However, the correspondence of individuals is more difficult to establish and following individual fates is complex. From a census to the next, for example, an individual child can either remain in the parental household, or be part of another type of household as a single-adult or married-adult (in the case of girls, they will then lose their name to take the name of their husband), or be lost to observation because of death or migration. Moreover, because of the linkage methodology focused on the selection and extraction of nuclear families, household members who are not part of this type of family unit are largely excluded from the process.

Data available for research in the IMPQ

Once completed, the infrastructure will make the following data available for research: 1) all marriage certificates since the implementation of parish registers in Québec in the early 17th century to 1965. This will enable automatic reconstruction of the genealogy of the Quebec population over a period of three-and-a-half centuries; 2) linked data based on three types of vital event records (birth, marriage and death certificates) from the beginning of the 17th century to 1849. The 2.3 million certificates will allow researchers to explore the historical demography of families over a 250-year period; 3) complete-count census microdata covering two urban settings (Quebec City and Trois-Rivières) and three regions mixing rural and urban environments (Gaspésie, Côte-Nord and Saguenay) across seven modern nominal censuses (1852 to 1911). These microdata (close to a million individuals belonging to 161,000 distinct households) will be linked across censuses and to corresponding vital event data.

This vast array of biographical information will permit studies based on individual trajectories situated within families, households and communities and examined from a multigenerational perspective.

Emerging research opportunities

The interconnection between civil records and the censuses, along with linkages across the censuses themselves, will substantially broaden and enrich the avenues of research in both the social and the biological sciences. It will be possible to conduct detailed studies on a crucial period in the history of the Quebec population (mid-19th century to the first decade of the 20th century) focussing on the evolution and long-term consequences of phenomena such as cultural diversity, social mobility and intercommunity relationships. During this period Quebec's inhabited space[3] expanded as a result of new agricultural settlements and broadened maritime and forestry activity; at the same time, Québec was transitioning to an industrial economy, and urbanization was accelerating. One of the most spectacular results of the linkage process will be the capacity to update the too often overshadowed life histories of one half of the population: women. As most censuses do not give the maiden names of married women, it has not been

[3] We mean the territory occupied by sedentary populations of mostly European descent.

possible up until now to use census data to investigate how women's living conditions evolved over their lifetimes.

From a population genetics and biodemographical perspective, the infrastructure will considerably enrich research on the transmission of biological and sociocultural characteristics, on the genetic diversity in Quebec regional populations and on the factors that have shaped this diversity. These will translate into original studies that can contribute to a better understanding of the genetic determinants of health.

Conclusion

Our goal was to develop tools and procedures for systematic and automated linkage involving vital records and census data. Development and testing operations are largely completed and we now have at our disposal a powerful program composed of three modules providing the ability to select the desired type of linkage (census to BALSAC, BALSAC to census and census to census).

Our linkage results are variable depending on the region and on the census year and further work could certainly be done to attempt manual linkage on individuals and households that could not be linked with the program. However, already we can provide to the research community original datasets on the Quebec population combining information on the two types of data which is unprecedented. In the course of the IMPQ project, a portal will be developed to provide access to these data.

We are also currently working on the description and analysis of the linkage results by comparing linked and unlinked households on various attributes. Notwithstanding the limits and potential biases mentioned above, we want to verify to what extent the linked population is representative of the whole population.

Lastly, the software could also be improved. It would be desirable to come to greater automation of decision-making and to further refine the process of individual linkage. But even it its current state it provides an innovative tool with the ability to combine longitudinal information from BALSAC to census data and to create census-based longitudinal datasets.

Acknowledgements

This project is funded by the Canadian Foundation for innovation, the Quebec ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie as well as the Université du Québec à Chicoutimi, Université du Québec à Trois Rivières and Université de Montréal.

Other researchers members the IMPQ project include Danielle Gauvreau (Concordia University), Bertrand Desjardins, Lisa Dillon, Alain Gagnon, Damian Labuda (Université de Montréal), Richard Marcoux (Université Laval), France Normand (Université du Québec à Trois Rivières), Marc Trembaly (Université du Québec à Chicoutimi). We also acknowledge the support of research professionals, clerks and technical staff especially Mario Bourque, Dominic Desbiens and Ève-Marie Lavoie.

Bibliography

Bellavance C, Normand F, Ruppert ES (2007) Census in Context: Documenting and Understanding the Making of Early-Twentieth-Century Canadian Censuses. *Historical Methods,* 40, 2: 92-103.

Bouchard G, Bouchard M (1981) *Ajouts au code de transcription phonétique FONEM*. Document de travail BALSAC II-C-66.

Bouchard G, Brard P, Lavoie Y (1981a) FONEM: un code de transcription phonétique pour la reconstitution automatique des familles saguenéennes, *Population*, no. 6 (novembre-décembre), pp. 1085-1104.

Bouchard G, Casgrain B, Roy R (1981b) *Tests de validation des fiches de couple par ordinateur*. Document de travail BALSAC II-C-67.

Bouchard G, Larose A (1976) La réglementation du contenu des actes de baptême, mariage, sépulture, au Québec, des origines à nos jours. *Revue d'histoire de l'Amérique française*, 30-1: 67-84.

Bouchard G, Roy R, Casgrain B (1985) *Reconstitution automatique des familles. Le système SOREP.* 2 volumes. Chicoutimi, Université du Québec à Chicoutimi, 745 pages.

Charbonneau H, Lavoie Y, Légaré J (1970) Recensements et registres paroissiaux du Canada durant la période 1665-68. Étude critique. *Population*, 25, 1: 97-124.

Gauvreau, D (1991) *Québec, une ville et sa population au temps de la Nouvelle-France*. Sillery, Les Presses de l'Université du Québec.

Gauvreau D, Thornton P, Vézina H (2010) Le jumelage des recensements aux mariages du fichier BALSAC présentation de l'approche et étude exploratoire des enfants de couples mixtes à la fin du XIXe siècle. *Cahiers québécois de démographie,* 39: 357-381.

Gossage P (1999) *Families in Transition: Industry and Population in Nineteenth-Century Saint-Hyacinthe.* Montreal/Kingston, McGill-Queen's University Press.

Olson S, Thornton P (2011) *Peopling the North American City: Montreal 1840-1900.* Montreal, McGill-Queens Press.

St-Hilaire M (1990) *Description du format d'impression d'une fiche de couple*. Document de travail BALSAC I-C-96.

St-Hilaire M (2009) Filling the gap: The use of marriage records to help with inter-censuses linkage of young adults in Quebec City (1851-1911). Paper presented at the 2009 *RECORDLINK* workshop, University of Guelph.

St-Hilaire M, Marcoux R (2006) Dynamique démographique dans une capitale en reconversion économique : Québec, 1851-1901. Communication au congrès de la Société historique du Canada (London, Ontario).

St-Hilaire M, Vézina H (2010) Between household and family: The use of marriage records to link census data (Quebec City, 1852-1911). Paper presented at the 2010 *RECORDLINK* workshop, University of Guelph.

Vézina H, St-Hilaire M (2011) Entre famille et ménage : le jumelage des données d'état civil et de recensement dans la population québécoise. Colloque « La fin des recensements ». 79^{e} congrès de l’ACFAS, Sherbrooke, 9-13 mai.