An Overview of the BALSAC Population Database. Past Developments, Current State and Future Prospects

By Hélène Vézina and Jean-Sébastien Bournival

To cite this article: Vézina, H. & Bournival, J.-S. (2020). An Overview of the BALSAC Population Database. Past Developments, Current State and Future Prospects. *Historical Life Course Studies, 9*, 114–129. https://doi.org/10.51964/hlcs9299

HISTORICAL LIFE COURSE STUDIES

Content, Design and Structure of Major Databases with Historical Longitudinal Population Data

VOLUME 9, SPECIAL ISSUE 5, 2020

GUEST EDITORS
George Alter
Kees Mandemakers
Hélène Vézina

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at http://www.ehps-net.eu/journal.

Co-Editors-In-Chief:
Paul Puschmann (Radboud University) & Luciana Quaranta (Lund University)
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 9 (2020), published 25-08-2020

An Overview of the BALSAC Population Database

Past Developments, Current State and Future Prospects

Hélène Vézina
BALSAC Project, Université du Québec à Chicoutimi

Jean-Sébastien Bournival
BALSAC Project, Université du Québec à Chicoutimi

ABSTRACT

The BALSAC database, developed since 1971, contains data on the Quebec population from the beginnings of European settlement in the 17th century to the contemporary period. Today, BALSAC is a major research infrastructure used by researchers from Quebec and elsewhere, both in the social sciences and in the biomedical sciences. This paper presents the evolution and current state of the database and offers a perspective on forthcoming developments. BALSAC contains marriage certificates until 1965. Coverage is complete for Catholic records (80 to 100% of the population depending on the region and the period) and partial for the other denominations. Birth and death certificates from all Catholic parishes have been integrated for the period 1800–1849 and work in underway for 1850–1916. All the records entered in BALSAC are subject to a linkage process which, ultimately, allows the automatic reconstitution of genealogical links and family relationships. The basic principle has remained the same since the beginning, namely to match individuals based on the nominative information contained in the sources. The changes made in recent years and the resulting gains are mostly related to IT advances which now offer more flexibility and increased performance. Future perspectives rest on the diversification of the sources of population data entered or connected to the database and, as a corollary, by continuous optimization of data processing and linkage procedures. In the era of 'big data', BALSAC is gradually moving from a historical population database to a multifaceted infrastructure for interdisciplinary research on the Quebec population.

Keywords: Family reconstitution, Population database, Quebec population, Record linkage, Vital records

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs9299
The article can be downloaded from here.

1 INTRODUCTION

The BALSAC database, developed since 1971 at the Université du Québec à Chicoutimi, contains data on the Quebec population from the beginnings of European settlement in the 17th century to the contemporary period. These data come from the digitization of civil records and have been linked together to reconstruct families and genealogical lines over almost 400 years. Today, BALSAC is a major research infrastructure used by researchers from Quebec and elsewhere, both in the social sciences and in the biomedical sciences.

BALSAC will be entering its fiftieth year shortly. Although the formal structure ensuring its management took various configurations and names[1], the main mandates have remained centered on the preservation and development of the database and on the promotion of its exploitation by the scientific community. Technological advances have, however, brought about considerable transformations in the operations surrounding these activities, both conceptually and technically.

This paper is an opportunity to describe the evolution and current state of BALSAC and to offer a perspective on forthcoming developments. We start with a brief overview of settlement history on the Quebec territory for a better understanding of the content of the database. Next, we trace the main steps in the construction of the database and present the work underway which will enrich it substantially. Then, we provide a detailed description of the content, structure and linkage methodology. Finally, we conclude by outlining development prospects for BALSAC.

2 A BRIEF OVERVIEW OF SETTLEMENT HISTORY

The Quebec population has characteristics conducive to the construction and exploitation of a population database focused on history, demography and genealogy. One of its main features is its recent formation following European exploration, imposing well-defined temporal and geographical limits for the reconstruction and investigation of its genealogical, family and even genetic heritage.

European settlement on the Quebec territory started with the arrival of French pioneers in the early 17th century (Charbonneau et al., 1993; Charbonneau, Desjardins, Légaré, & Denis, 2000). Approximately 10,000 immigrants settled and experienced a family life in the St. Lawrence valley during a century and a half of French rule. The only period of relatively high immigration was from 1663 to 1673, when the King of France sent some 800 'Filles du Roi' to overcome the shortage of women and to encourage soldiers from the 'Régiment de Carignan' to settle in the colony (Landry, 1992). The vast majority of immigrants came from France, while most of the others originated from countries bordering France. When New France became a British colony in 1759, the population was around 70,000. There were a few thousand Aboriginals and the rest were almost all French Canadians settled in the Laurentian Valley.

Following the British takeover, the French-speaking immigration virtually stopped except for a few thousand Acadians[2] who took refuge in Quebec following deportation by the British authorities (Dickinson, 1994). The majority of newcomers were British immigrants or Loyalists escaping the American War of Independence. It should be noted that the French population being Catholic and the English-speaking immigrants being Protestants, mixed marriages took place but remained infrequent.

In the 19th century, the French-Canadian population progressively overflowed the Laurentian corridor leading to the opening of new regions. Most of the immigration continued to come from the British

1 The research conducted with BALSAC first took place within the Société de recherches sur les populations (SOREP) created at UQAC in 1976 before becoming an interuniversity group in 1982. In 1994, SOREP became IREP, the Interuniversity Institute for Population Research, bringing together researchers from seven Quebec universities. Since 2002, BALSAC users are no longer grouped within a specific entity and the database is under the joint responsibility of UQAC, Laval University, McGill University and the University of Montreal. The BALSAC Project at UQAC manages it.

2 The Acadians are descendants of French immigrants who settled in Eastern Canada in the 17th century. In 1755, the British authorities ordered the deportation of Acadians who were dispersed in France, England and the English colonies of America. It is estimated that between 2000 and 4000 Acadians settled in Quebec.

Isles as thousands of immigrants from England, Scotland and Ireland settled in the province, most of them in the urban areas of Montreal and Quebec City (McInnis, 2000b). By 1851, Quebec had 890,000 inhabitants three quarters of whom were French Canadians.

From the beginning of the 20th century, the origins of the immigrants diversified, with many newcomers arriving from Southern and Eastern Europe (McInnis, 2000a; Piché, 2003). More recently, immigrants from Asia, South America and the Caribbean have outnumbered those from Europe (ISQ, 2019). From these various movements of immigration and settlement, results a population of some 8.5 million inhabitants, within which we find, in addition to Aboriginal communities, a French-speaking majority, an English-speaking minority and a segment made up of recent immigrants. The last two groups are mainly concentrated in the Montreal region although present in variable proportions in all regions of the province. As we will see below, the coverage of each of these groups in BALSAC depends on the characteristics of the settlement history, but also on the quality and availability of vital statistics which vary substantially across groups.

3 THE CONSTRUCTION OF THE BALSAC DATABASE

BALSAC started in the early 1970s as the project of a historian, Gérard Bouchard, who had just completed his doctorate in France where he had used the methodology developed by Louis Henry for the reconstitution of families from parish registers (Fleury & Henry, 1956). As new professor at the Université du Québec à Chicoutimi, he initiated the BALSAC project aimed at reconstructing the Saguenay–Lac-Saint-Jean (see Figure 1) population from 1837, the start of the French-Canadian settlement in the region, until 1971 using the 660,000 birth, marriage and death certificates (or acts) recorded during this period. This work completed in 1986 was the first major achievement in the development of the database.

Figure 1 *Geographical location of the cities and regions referred to in the text*

Source: Centre interuniversitaire d'études québécoises (CIEQ), Université Laval

Subsequently, the database gradually expanded to include all regions of Quebec. The name 'BALSAC' comes from an acronym made up of the first letters of the name of eastern regions of the province which constituted the very first large-scale corpus. During this period, in addition to the work conducted in historical demography and social history, a vast research program on population genetics and hereditary diseases was set up. This led to the decision to prioritize the entry of marriage certificates because of the importance attached to the genealogical approach for the exploitation of the database in the field of human genetics. This

second phase of development, extending up to 2011, added to BALSAC more than two million marriages covering all of Quebec until 1965.

Since 2010, a new phase of development is ongoing led by Hélène Vézina who has succeeded Gérard Bouchard as director. The main objectives of this new stage are 1) to add births and deaths to marriages for complete family reconstitution; 2) to adapt the database structure and linkage procedures to facilitate the connection between civil records and other types of population data; 3) to facilitate access to the database by setting up web portals.

It is in this context that BALSAC piloted, from 2013 to 2017, the creation of the Integrated Infrastructure of Historical Microdata of the Population of Quebec (IMPQ) in partnership with the Programme de recherches en démographie historique (PRDH) and the Centre interuniversitaire d'études québécoises (CIEQ) (Vézina, St-Hilaire, Bournival, & Bellavance, 2018). In the course of this project, Quebec births and deaths for 1800–1849 (more than 1.6 million records) were added to the database and civil records from BALSAC were linked to the Canadian censuses from three regions (Saguenay, Gaspésie and Côte-Nord) and two cities (Quebec and Trois-Rivières) for the period 1851–1911 (see Figure 1). Thanks to this initiative, individual life courses in BALSAC now include appearances to censuses, which contribute to filling the sometimes very long intervals among vital events and to optimize the chances of success in the various linkage operations.

Since 2019, BALSAC has been conducting a new project aimed at the creation of i-BALSAC, an infrastructure to study the Quebec population through a joint genealogical, genomic and geographical approach. The project is set around five components: integration of demographic, genetic and geographic data, development of statistical and mapping tools in order to optimize exploitation of this data and implementation of a web portal for access ('BALSAC', 2020). Through the demographic component of the project, birth and death certificates for the entire Quebec population from 1850 to 1916 (approximately six million records) will be integrated into BALSAC relying on handwritten text recognition (HTR) technology. The goal is to complete families and pedigrees in BALSAC to get as close as possible to a full population coverage and give access to omics-oriented researchers, among others, to the genealogical and familial structure up to the first decades of the 20th century.

4 DATA COLLECTION

Since the beginning of the French settlement, Catholic priests kept registers of vital events. From 1679, they were given the mandate to keep these registers in duplicate, one under ecclesiastical jurisdiction and the other, sent to courthouses, by virtue of what was to become Quebec's civil registration system (Bouchard & LaRose, 1976). This method of registration was maintained under the English Regime and continued until 1994 with the reform of the Civil Code of Quebec. Almost all the registers have been very well-preserved enabling their microfilming and more recently their digitization (LaRose, 2015).

Although they were subjected to the same civil regulation, there is a marked difference between records coming from Catholic parishes and those recorded in non-Catholic parishes (essentially Protestant) which are less rich in content making their linkage much more difficult and often impossible. For this reason, the transcription of non-Catholic records in BALSAC has not been systematic and, to date, only a few regions and periods have been processed. As we will see later, ongoing developments could help correct this situation in the coming years.

Most of the data integrated into BALSAC comes from the civil copy of the registers kept in courthouses. In the first phase of development, research assistants had to go to the courthouses in Saguenay–Lac-St-Jean region to transcribe the records on paper files and then come back to the office to transfer the data on punch cards for computer processing. In the second phase, marriage records were first entered using a microfilmed copy of Quebec registers from the Fonds Drouin. Then, with the gradual digitization of civil records, we started working with images obtained from the Directeur de l'état civil and more recently from the Bibliothèque et Archives nationales du Québec (BAnQ) through collaboration agreements. Transcribed records retain a double link with the register since for the majority of them, a link to the image is created, which facilitates consultation of the original document. It is also easy to trace the act in the register as we keep information on the location (type and number of the act in the margin). Finally, during linkage operations and genealogical reconstructions, it is sometimes necessary to consult external sources (genealogy websites, registers outside Quebec, etc.), to obtain, for example, additional information on events that have taken

place outside Quebec or on the origins of immigrants. Except for the data obtained through exchanges with partners, the data entry work has entirely been done manually. However, the use of handwritten text recognition (HTR)[3] now opens the way to automatically transcribing large batches of data at a lower cost and over a shorter period of time.

Although an increasingly powerful tool, the HTR process faces several challenges (Vézina, Kermorvant, Bonhomme, & Bournival, 2019). While the algorithm will 'learn' how to interpret page structure and text, any lack of uniformity constitutes an obstacle that must be addressed in some way. In addition, the larger the dataset, the greater the variability. In the i-BALSAC project, the dataset includes two million images and more than 40,000 registers. The quality of the images varies, some of them being of lower grade. This variability is also present in the languages found in the registers (French and English), in the thousands of handwriting styles, as well as in the structure of the registers that can be observed across Catholic and non-Catholic denominations.

Since each image cannot be assessed manually, the process commands an automatic page recognition prior to text recognition. This step identifies pages containing text and therefore removes blank and cover pages. Then, the process identifies how the acts are structured on the pages before reading them. The ultimate goal is to extract specific entities such as names, dates, occupations, and places. As we have access to the complete transcriptions, other entities such as honorific titles, literacy, attendance, ethnocultural group and many others could be extracted subsequently.

Fortunately, the structure of acts varies little over time, which favors a uniform collection of the information contained in the registers. Excluding godparents and witnesses, the number of participants in each type of act remains constant. A birth certificate normally contains three persons, the subject and his or her parents. There are six persons mentioned for a marriage, the spouses and their respective parents, unless it is a remarriage in which case the ex-spouse will be mentioned in place of the parents (which does not exclude that parents can also be named). Finally, depending on whether the death certificate concerns a married individual or a single individual, the act will contain two (subject and his or her spouse) or three (subject and his or her parents) mentions. Each act also contains information specific to each person such as occupation, place of residence, age (major or minor in the case of a marriage), presence, ability to sign, honorary titles, geographic origin and ethnocultural group. The frequency of these characteristics varies according to the type of event, the period of registration and the person's role. Due to the costs associated with manual data entry, some information, mostly secondary actors like godparents or witnesses, were omitted. However, the use of automatic text recognition will allow the extraction of all relevant information in the register in a near future.

The guiding principle of transcription is that the data should reflect the source as precisely as possible, including the inaccuracies and errors it contains. Names and dates are entered as they appear, despite the fact that inconsistencies in these fields may reduce the chances of a successful match. Limits due to nominative variations are taken care of by the use of dictionaries and standardization steps (see the Data linkage section below).

Any addition of data requires one or more procedures to control the quality of the information, upstream or downstream. Whether during integration or linkage operations or while validating the integrity of the database, automated queries are used to detect potential sources of errors or internal inconsistencies (see the Validation section below).

5 CONTENT OF THE DATABASE

As shown in Table 1, the database contains marriage certificates from the Quebec registers until 1965. Coverage is considered complete for Catholic records (which represent 80 to 100% of the population depending on the region and the period), but it is only partial for the other denominations. For the latter, great variability in the format of the certificates and in the information they contain (or do not contain) complicates or even makes it impossible to attempt linkage. This is why the transcription work was not systematically performed.

3 Work performed in the i-BALSAC project is conducted in partnership with Teklia (https://teklia.com/) and Transkribus (https://transkribus.eu/Transkribus/) is used for the ground truth step.

Table 1 *Spatial and temporal coverage of the three types of events in BALSAC*

	Births		Marriages		Deaths	
	From	To	From	To	From	To
SLSJ Region*	1838	1971	1838	1971	1838	1971
Charlevoix Region	1680	1945	1686	1992	1686	1992
Rest of Quebec**	1800	1849	1621	1965	1800	1849

* SJSJ = Saguenay–Lac-Saint-Jean

** Marriage records prior to 1800 come from the PRDH. They were integrated to BALSAC through a collaboration agreement.

Note: With the ongoing work as part of the i-BALSAC project, births and deaths of the whole of Quebec for the 1850–1916 period will be integrated by the end of 2022.

Birth and death certificates from Catholic parishes for the whole province are now integrated for the period 1800–1849. Only the Saguenay–Lac-Saint-Jean and Charlevoix regions (see Figure 1) have more extensive temporal coverage. Saguenay–Lac-Saint-Jean marriages, births and deaths cover the period 1838–1971, while for Charlevoix coverage extends from the end of the 17th century until the early 1990s, with the exception of births whose transcription stops in the 1940s. As mentioned above, the reconstitution of the Saguenay–Lac-Saint-Jean population represented the first phase of construction of the database. Work on the Charlevoix region, bordering and historically very connected to the Saguenay, was performed at the very beginning of the second phase before the decision to focus on marriage records was taken. Numerous studies were conducted on these two regions both from the perspective of demographic and social history and from that of population genetics (see for instance Bouchard, 1996; Bouchard & Braekeleer, 1991).

Since the population database is constructed from civil records, all unregistered events are, by definition, unknown to us. In the absence of a clear denominator, it is therefore difficult to accurately measure the completeness of BALSAC coverage. Concerning the Catholic population, the rigor observed by the priests for the keeping of registers and the strict rules of transcription during the data entry process suggest that under-registration is minimal.[4] However, some groups are clearly less well represented. We are thinking in particular of the Aboriginal groups, which have largely escaped religious registration. Although it is possible to trace individuals in the registers through mixed marriages or declarations of ethnicity or even origin, the coverage is far from optimal.[5]

This situation is, however, bound to improve. The use of HTR for the transcription of births and deaths for the period 1850–1916 will allow us to process, in addition to records from Catholic parishes, those from Protestant, Jewish and Orthodox parishes, as well as from several Aboriginal communities. Thanks to the HTR, it will also be possible to integrate the non-Catholic records of the previous periods left aside in the previous phases of development. At the end of this integration planned for 2022, the overall coverage of BALSAC and the representation of the Quebec population will thus be significantly enhanced. More than six million documents will have been processed in three years, equivalent to twice what was compiled during the first 50 years of BALSAC's existence.

The nature of the data lends itself to three levels of observation: individuals, couples (or unions) and events. The frequency of each of these units of observation and of each type of event is presented in Table 2. The number of individuals and couples are the numbers after linkage meaning that individuals and couple are counted only once independently of the number of events where they appear. It can be noticed that the number of unions exceeds the number of marriages. This is because some couples are mentioned as parents in their children's records but we do not have their own marriage certificate. Other unions are known but do not come from formal readings of Quebec register. In the context of

4 Missing data and under-registration have been extensively investigated by the PRDH researchers for the period of the French regime (Dillon et al., 2018). During the first phase of development of the database, a study was conducted at BALSAC on the parish of Laterrière in Saguenay (Bouchard & Bergeron, 1975). We also discuss this topic in a paper currently under review (Bournival, St-Hilaire, & Vézina, 2020).

5 As part of the IMPQ construction project, the censuses from the Côte-Nord, a region that contains a large Aboriginal population, display the lowest linkage rates to BALSAC, and this is particularly pronounced for the 1851 and 1861 censuses.

research projects using genealogical reconstructions, interruptions in genealogical lines are investigated in external sources and when the union is found it is integrated in the same way as a marriage. This distance taken from the registers is beneficial in that it makes it possible to add generations to lines affected by emigration or to document the reason for the interruption of a genealogical line. Over the years, tens of thousands of marriage certificates, mainly outside Quebec, were entered in BALSAC contributing to a finer understanding of migratory movements and family history of migrants.

Table 2 *Frequency of observation units in BALSAC*

Unit of observation	Number
Individuals	6,351,130
Unions (couples)	2,660,521
Events	4,327,002
Births	1,445,224
Marriages	2,303,306
Deaths	578,472

Note: With the ongoing work as part of the i-BALSAC project, the births and deaths of Quebec for the period 1850–1916 will be integrated by the end of 2022.

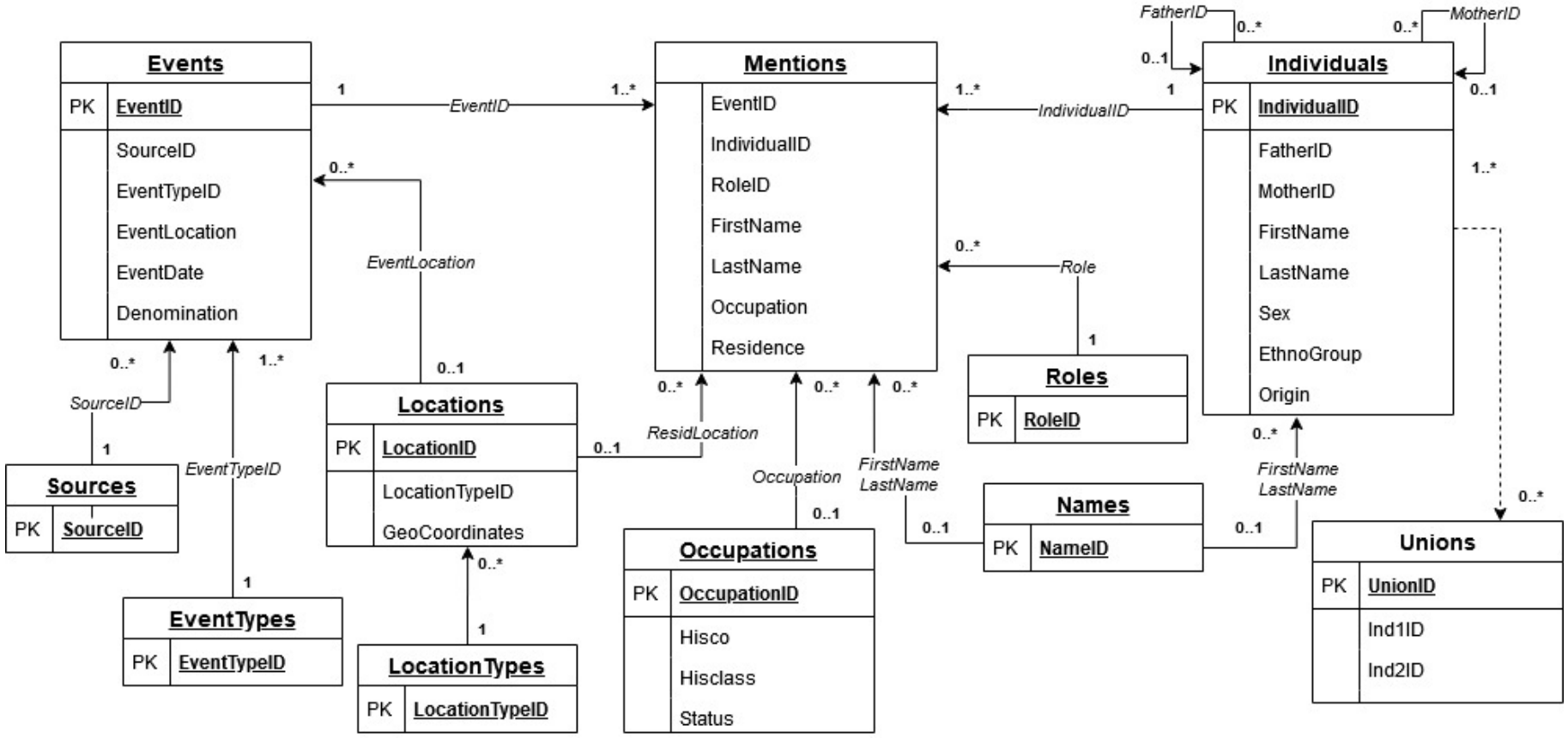
These levels or units of observation are obviously interdependent, but they can be used to answer different research questions. The observation of an individual begins with the first recorded event where he or she appears, whether as a subject or as a parent, and ends in the same way. The life cycle of a family starts when a couple gets married and ends when both members are dead or one is dead and the other remarries (unless it is lost to observation obviously).

In Quebec, data from birth, marriage and death certificates become public after 100 years. It is therefore not possible to disseminate information from vital events registered less than 100 years ago that could enable the identification of a person. The BALSAC Researchers' Support Service receives requests for access and ensures that they comply with the terms of the BALSAC Data Access Policy. It produces for researchers datasets that respect the rules of confidentiality and protection of personal information. Data recorded more than 100 years ago, which are public, are also available for consultation for research purposes on the IMPQ portal.

6 DATABASE ARCHITECTURE AND VARIABLES

BALSAC is structured as a set of relational tables where primary keys are the identifiers of events and individuals. Most of the variables come directly from the source, but some of them were created to specify or add information. For example, a date marker is used to document the accuracy of dates. In Figure 2, we focus on the content and architecture of the database according to the main source of data, namely the civil registers. But the architecture is flexible and with the integration of new data additional tables and variables can be created. The database includes three main tables *Events*, *Mentions* and *Individuals*. This is the result of recent changes made to the original structure of BALSAC in order to offer enhanced reliability as well as maximum flexibility to query the data and extract sets designed to answer the increasingly diversified needs of researchers. In fact, this modification has several advantages as it makes it possible to eliminate redundancy of information, to reduce internal consistency errors due to asynchronous information in linked tables and to develop a 'universal' linkage tool that is no longer limited to couples and can take into account other types of relationships between individuals (as described below) for the selection of candidates. The descriptive tables are devised to contain text values of entities found in each act (names, roles, occupations, locations, among others) as well as coded values that make the junction with the main tables. Another important change is that the table *Unions* no longer exists as such but rather takes the form of a dynamic view where information on couples is obtained from records in real time.

Figure 2 *Architecture of the BALSAC database*



Note: For the sake of concision, only the main variables of each table are presented.

The *Events* table was designed to identify and classify the different acts contained in the registers. Each event is given a unique number and characteristics such as the type of event, the place and date of registration, the date of the event as well as the source it was retrieved from are compiled. The religious denomination of the church or parish where the event was recorded is also indicated. Information on the individuals mentioned in the records are listed in two tables: *Individuals* and *Mentions*. The *Mentions* table contains all the entries likely to vary for a given person from one act to another (such as occupation, residence, role in the event, presence), while the *Individuals* table comprises the fixed characteristics such as sex, dates of birth and death, birth status (legitimate or not), geographic origin and ethnocultural group as well as the parents ID. The occurrence of these variables can vary across records, but they are always transcribed when they appear. The name of an individual can change so it is recorded in the *Mentions* table and its most frequent version appears in the *Individuals* table.

These variables are of great interest for carrying out comparative studies on various topics related to fertility, mortality and migration. It is also possible to track the social mobility of individuals and families across the occupations recorded in the events. Used in conjunction with honorary titles, occupations can serve as indicators of socioeconomic conditions. To facilitate national and international comparisons, a large part of occupations and honorary titles have been coded according to the HISCO (van Leeuwen, Maas, & Miles, 2002) and HISCLASS (van Leeuwen & Maas, 2011) classifications. Of the entire dictionary of occupations, almost 63% of the entries were coded. In terms of frequencies, the coded occupations cover 98% of all the events contained in the database. As many records contain more than one occupation, the count was done on the basis of the presence of at least one occupation coded in HISCO in the record. However, this concerns mostly men as women are clearly underrepresented in terms of the declaration of occupation (Bourque, Markowski, & Roy, 1984).

Regarding geographic variables, events generally contain several entries. First, the place of registration is always on a parish level. This is probably the most consistent information in all types of events. Then, in each type of record, certain persons (mainly subjects, parents and spouses) declare places of residence: even if the parish is often mentioned, no geographical level is prescribed so there is a great variability — from the road to the continent — which poses a problem for the standardization of the information collected. In addition to a simple ID, locations are also characterized by a specific level (*LocationTypeID*). The use of a geographic dictionary allows the different mentions to be grouped in different levels or scale units. Municipalities and parishes of Quebec are geocoded and our ongoing projects will lead to a better cartographic representation of these various levels, including census districts and sub-districts as well as some cities outside Quebec (mainly in the Canadian provinces and in the United States).

It is also possible to retrieve information on the religion, origin and ethnocultural group of individuals. The religion is not an individual characteristic per se as the information is derived from the denomination of the church where the event was recorded. As mentioned earlier, the BALSAC database covers the entire Catholic population, but only partially the non-Catholic population. The scope of this variable therefore does not depend on information contained in the registers, but rather on the source itself. Concerning origin and ethnocultural group, the information may be transcribed from the acts, but it is also searched in complementary sources when needed in research projects or genealogical work. Finally, when reconstructing genealogies, we assign a migratory status to ancestors according to whether they are native, immigrants or have never come to Quebec (most of the time the latter are parents of immigrants married in Quebec who appear in their child's marriage certificate). These variables have proven very useful in documenting the origins of the Quebec population.

7 LINKAGE METHODOLOGY

The ability to process and integrate a large number of individual data is essential when the objective is to cover an entire population. The strength of such a corpus, however, is that all of this data is linked together. All the records entered in BALSAC are subject to a linkage process which, ultimately, allows the automatic reconstitution of genealogical links and family relationships in the Quebec population.

The basis of the program has remained the same since its development in the 1970s, namely to match individuals based on the nominative information contained in the sources. The changes made since

2018 and the resulting gains are mostly related to IT advances which now offer more flexibility and increased performance. Modifications to the database architecture as well as to the name processing algorithm have also improved the efficiency of the linkage program. To put these recent developments in perspective, we first review the general principles that have helped make BALSAC what it is today.

7.1 THE GENERAL PRINCIPLES OF FAMILY RECONSTRUCTION

The family reconstitution system developed in the initial phase of the construction of the database is based on the nominative information contained in the registers and aims to group in the same file all the mentions referring to the same couple (Bouchard, Roy, & Casgrain, 1985). Thus, the basic information unit for creating the link is the 'couple mention' in a record which contains four nominative elements, namely the names and surnames of the husband and the wife as parents and as bride and groom.

The linkage program is based on two distinct and interdependent steps: the search for candidates and the linkage process itself. The mentions of candidate couples are created on the basis of at least two nominative elements common to the couple to be matched. These mentions are compared using the three modules of the FONEM phonetics program designed to detect and measure the degrees and forms of similarity between two surnames or two first names (Bouchard et al., 1985). ISG (Similarity index) calculates a score based on the degree of similarity between the nominative pairs of elements according to the position of the same letters in the names. INC (Inclusion) deals with truncated names by detecting the suffixes and prefixes in the names and deciding whether one can be treated as being included in the other. ELM (Multiple elements) deals with the situation of first and last names comprising more than one element and decides whether two entities can be treated as equivalent or not.

The linkage decision-making process is based almost exclusively on the nominative information contained in the files, but the consistency of the dates in the sequence of the family history is also taken into account (Bouchard et al., 1985). The linkage operations lead to the construction of family files which include all the events relating to a single couple (their own marriage and death, the remarriage of the surviving spouse, the births and marriages of children and the deaths of single children) as well as pedigrees with the connection of successive generations. All links created through automatic linkage and those made during the computer-assisted manual linkage stage are immediately validated by automated consistency routines (for example: acceptable interval between two events, logical sequence of events, correspondance between reported and calculated age, detection of duplicate events) (Bouchard, Casgrain, & Roy, 1981).

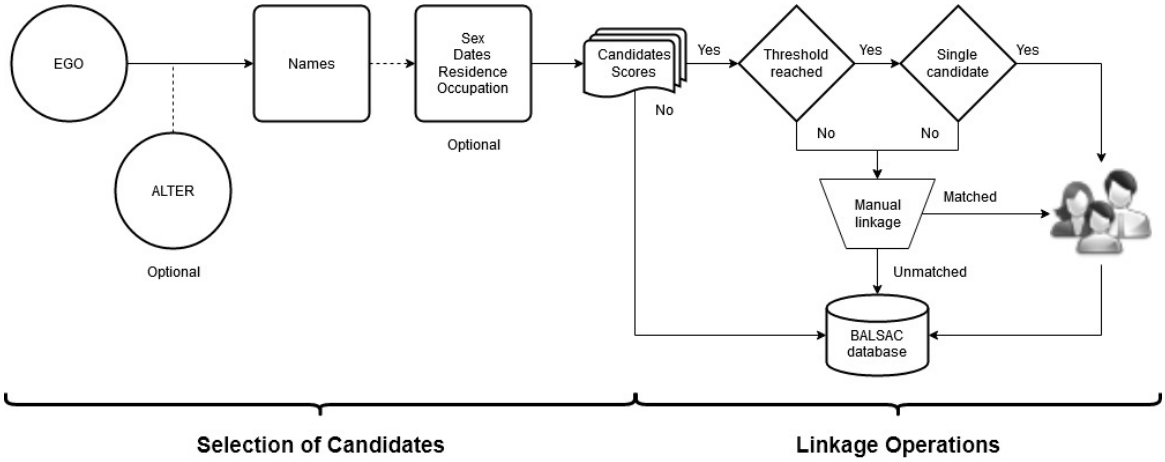
7.2 TOWARDS A 'UNIVERSAL' LINKAGE MODULE

BALSAC's current expansion goals require adapting the linkage method. As mentioned earlier, the underlying mechanics remains the same. However, we have relaxed the rules governing the linkage of civil records and introduced the flexibility necessary to open up to other types of data, while optimizing the analysis and comparison of names which are instrumental for successful linkage. The new version of the program also takes advantage of the richness of the data contained in the database such as the temporal coverage, the multiple individual occurrences, the geographic dictionary or the availability of more than one source for the same event (for instance, having both the religious and the civil copy of the register). Figure 3 presents an overview of the linkage process with its two main steps: 1) the selection of candidates which includes the calculation of a score for each proposed candidate and 2) the linkage operations where decisions are made based on a chosen threshold.

7.2.1 SELECTION OF CANDIDATES

The first adaptation aimed at improving the selection of candidates. The old version worked exclusively on the basis of couples (as parents or as bride and groom) so the selection was limited to the conjugal structure. The objective was to create a universal linkage module making it possible to search for candidates relying on various family structures by exploiting the wealth of genealogical lines. The new version of the linkage program offers the possibility of including other individuals (ALTER in Figure 3) who have a relationship with the subject (EGO in Figure 3). In the case of civil records, this mainly concerns parents, but it applies to any type of family relationship, as well as godparents and witnesses. For example, it is possible to search for a candidate named X whose father is named Y and an uncle is named Z.

Figure 3 *Overview of the linkage process*



Note: EGO is the subject to be linked. ALTER represents individuals related to EGO who can be used to improve candidate selection.

As they are the main keys to candidate search and matching, names remain the most important element in all operations. Previously the selection of candidates and score calculation relied exclusively on the nominative information found in marriage records. Now the program considers all occurrences or mentions of an individual to calculate a score (that is all his or her participations as subject, parent or godparent in vital events or as an enumerated individual in a census or listed in some other source). It is very common to find orthographic variations or even complete nominative variations (different names) in the list of mentions to which an individual is tied. While minor spelling variations are generally not a barrier to the selection of candidates, excessively large variations can lead to the creation of false positives or the omission of potential candidates. The program now takes into account all these variations, hence the interest in including various sources which enrich the individual data and maximize the chances of success in linkage operations.

Moreover, as the database already contains a large number of names and since all the mentions of an individual are now considered, it has become easier to link using names as they appear in registers (or any other source) without having to systematically use a standardized form like it used to be, at least for the first pass. Candidates' selection proceeds by iterations using tools called in reinforcement for the management of more complex cases when the names do not generate candidates, or not the right ones. The performance of these tools has been enhanced by the use of more sophisticated dictionaries containing orthographic variations of the same names, standardized names, patronymic equivalences, or even linguistic equivalents (names that have been translated or modified in the context of emigration, such as 'Boisvert' becoming 'Greenwood' in American censuses).

The program also offers the possibility of searching candidates according to additional variables such as dates, places or occupations. The use of these criteria is optional as they can be selected and weighted as needed. Time-changing variables are more likely to create biases in candidate selection. For residences, the calculation of a score based on the similarity of residence over the life course certainly favors sedentary individuals (same parish, same city, or even same region). This is where weights come in, making it possible to assign different weights to optional variables. For example, it is possible to decrease the weight of the declared residence in order to extend the selection and perhaps bring in candidates who would otherwise be rejected. The professions have a more limited usefulness since a large proportion of individuals will be farmers at one time or another in their life (at least up to recent times). However, a more ad hoc use targeted at professions whose occurrence is more moderate, or even low, can be an asset for the selection of candidates while risking creating a bias in favor of individuals who have a more stable professional career. It is important to consider the biases that might be introduced while using these criteria however in some contexts it might be useful and relevant to take advantage of all the information available on individuals.

All the chosen variables for candidate selection are processed simultaneously and act as blockers to discard the candidates who do not meet the search criteria. Sex is, of course, critical in this type of research since it excludes almost half of the database from the start. If a date or an interval is

mentioned, the program only retains events that occurred during this period. A slightly different treatment is assigned to the variables likely to change across records, namely names, residences and occupations. The score of these variables is an average value calculated over all the events in an individual's biography. If we consider for instance surnames, in the case of significant variations, the score will be lower since the candidate does not perfectly meet the specified criteria (it is not similar to the source in all instances). Conversely, this method ensures that no candidate who has reported the searched name at least once is eliminated.

The possibility of choosing the selection criteria and adapting the weighting system makes the program very flexible and facilitates the linkage between various primary sources and vital records. Creating models or templates makes it easier to adjust specific parameters for each data source. Also, in the context of population reconstruction, all of the available information that helps to choose between competing individuals can be mobilized during the linkage process. Thus, locations, professions and even the names of children can play a role in the calculation of scores and often lead to targeting unique candidates, which increases the possibilities of automatic linkage.

In the end, it is the summation of values for each variable that generates a final score for a specific candidate and it is this score that will determine whether there is a match or not. Obviously, the more information (in the source material and in BALSAC), the more effective the selection of candidates.

7.2.2 LINKAGE OPERATIONS

With regard to linkage, the operation is split into an automatic and a computer-assisted manual component. In both cases, the selection of candidates is involved; however, in the second, the conditions for automatic linkage have not been met and the linkage must therefore be submitted to the human eye for decision-making. For the automatic component, we run a first pass using tight criteria: a subject is linked only if there is only one potential candidate for the link and if the nominative information is perfectly identical (high score). If several individuals are proposed as candidates, a minimum threshold eliminates those whose scores are too low to justify an automatic match and linkage takes place if only if an unequivocal choice can be made (only one candidate with score above the threshold). The results of automatic linkage on the civil registers have historically hovered around 80%. As part of the IMPQ project, the linkage of censuses with BALSAC has also shown interesting results. Although the chosen method at the time was based on a human decision, we estimated that the selection of candidates would have made it possible to automatically link to the right candidate in about 75% of the cases.

A new addition to the program rests on the implementation of a second pass to increase the rate of automatic linkage. When the program has not been able to make the appropriate link with the information available in the first pass, the use of the different thesauri allows a certain level of tolerance in the face of nominal variations thus extending the selection of potential candidates. The linkage process is performed as in the first pass. Multiple candidates with scores above the chosen threshold, which are explained most of the time by homonymy (notably with Josephs and Maries) or incomplete information, and other ambiguities are referred to manual processing in proportions which depend to a large extent on source-related factors. Obviously, a certain proportion of the subjects to be linked do not exist in the database and therefore no candidates are proposed. They are then simply added to the database.

Manual linkage is a less standardized operation: it can be limited to the checking (and correcting if necessary) of the transcription by returning to the source, but it can also consist of a manual search for candidates in the database or require the use of external sources such as genealogical indexes or websites to find or clarify information and support decision making. Finally, after all these steps, there are still a number of unresolved links. As with the automatic linkage stage, these 'floating' events are kept in the database and may be the subject of further investigation. Since this can be a long and costly operation, the use of this type of inquiry is generally performed in the context of specific research needs. However, a 'silent' algorithm constantly scans the database for potential matches. Any addition of data therefore enriches the corpus and makes it possible to update certain previously unsuccessful linkages.

Finally, notwithstanding their overall quality and completeness, the registers present certain variations across time and space and, consequently, the proportion of records automatically linked will not be constant. In order to control for this and after performing various tests, we have come to the conclusion that it is preferable to 'simulate' the linkage process and to analyse the predicted results before launching

the actual automatic operations and integrate the results in the database. This process validates the consistency of the results by identifying the potential biases inferred by the data. Automatic linkage rates that appear to be too high or too low provide information on the quality of the registers or on the importance of homonymy. Different strategies can then be applied such as adjusting the parameters of the program in order to restrict or extend the selection of candidates or processing the records in selected sets in order to eliminate certain sources of noise. For instance, linkage attempted on only one parish at a time and over a defined period will allow the least mobile individuals and families to be effectively matched. It can sometimes be very wise to quickly solve these most obvious cases from the start in order to optimize automatic operations and promote gradual consolidation of the database. Subsequent operations on more difficult cases will then be facilitated.

7.3 VALIDATION PROCESS

In the development of a population database, erroneous links distort genealogical lines and may bias research results. Validation therefore represents an important step since it ensures the maintenance of the integrity of the database. A three-part process was put in place to perform data verification at crucial stages.

First, the accuracy of the transcriptions from the registers is verified. In the case of a manual entry, a research assistant validates data entry from a randomly selected sample of records. In the context of the i-BALSAC project for which the reading of millions of pages of registers has been entrusted to handwriting recognition algorithms, quality metrics assess the reading of records against information already contained in BALSAC. Recurring mentions such as surnames, first names, residences and professions can thus be standardized, despite a non-optimal reading by the algorithm. For instance, using the name dictionary, it is possible to measure the difference between a name read by HTR and the 'closest' name in BALSAC. Using these metrics, we can accept, based on a certain threshold, what seems most likely.

At the linkage stage, whether manual or automatic, integration into a family file is marked out by strict rules which ensure minimal consistency. Thus, in the call for candidates, an acceptable duration between two events is considered, in particular for intergenerational differences. In general, it is at this stage that we detect false positives and in particular problems related to homonymy. These tests aim to expose contradictions or inconsistencies in family files suspected of containing mentions referring to two distinct couples.

Even if the consistency tests can uncover the vast majority of homonymous couples wrongly linked, the nature of the data makes it inevitable that a small fraction of these cases escape these controls. In addition, since all the links specific to an individual or to a family file are not necessarily integrated in the database at the same time (for births at the end of a given period it is more than likely that death is missing), validation during data entry or linkage will not detect all internal inconsistencies. For this reason, post linkage validation of the data is also performed to identify family files that contain features that could hint to incorrect links. These checks take the form of automatic requests which can identify files containing potential duplicates, an abnormally high number of records, suspicious observation periods, sequences of improbable events (for example births too close together), inconsistencies between age declarations and actual values obtained from known dates.

8 GENEALOGIES AND KINSHIP RELATIONSHIPS

The linkage process described above has shown how generations and families are reconstituted, one record at a time. Once completed, the population database enables the reconstruction of ascending genealogies of all individuals and the exact measurement of kinship relationships between these individuals. The temporal extent of BALSAC makes it possible to trace lineages over an average of 10 generations and can extend up to 18 generations in some cases. The situation is slightly different for descending genealogies. The absence of births and deaths for certain periods does not yet allow the complete reconstitution of families, but it is possible to look at the married descendants in both extant and extinct lineages.

Using the data in BALSAC, researchers have access to individual biographies and family histories. Basic queries allow the extraction of all the kinship relations of one or more individuals up to the desired generation with the possibility of adding specific criteria (for example, relatives alive or not on a given date). From intergenerational and kinship links, it is possible to carry out several analyses which are used specifically to study genealogical datasets. A large number of queries are grouped together in a single procedure which produces, for a given corpus, the ascending genealogy of each subject, a set of descriptive measurements such as generational completeness and average depth, the portrait of the paternal and maternal lines, the list of immigrant founders, and measures of the frequency and genetic contribution of ancestors according to various characteristics (origin, sex, period of arrival, etc.). It is possible to extend these analyses in the R environment with the GENLIB package (https://cran.r-project.org/package=GENLIB), an open access genealogical analysis module which offers a range of functions to manage, describe and compute various measures for population genetics and genetic epidemiology (Gauvin et al., 2015)

Finally, the recent addition of godparents and witnesses from birth and marriage certificates opens the way for the analysis of extra-family networks. These persons who, mostly due to costs, had never been systematically entered into BALSAC, provide highly valuable information on social and support networks as well as their dispersion in time and space.

9 FUTURE PERSPECTIVES FOR THE DEVELOPMENT OF BALSAC

The data in BALSAC concerns mostly individuals from the past but the infrastructure that hosts and maintains it must remain anchored in the present. This is the case from a technical standpoint to keep pace with IT advances, but also from a research point of view where the needs and requests of users evolve with theoretical and methodological developments in their disciplines. One common feature is that scientists work with ever larger datasets to perform more and more complex and sophisticated analyses. The nature of the data as well as the spatial and temporal coverage makes BALSAC relevant not only for social scientists, but also for geneticists and researchers in the biomedical field. To fulfill its mission and open up a maximum of opportunities, BALSAC's offer must therefore take into account disciplinary trends and specific needs in terms of data, methods and performance.

We wish to adjust to this demand, by diversifying the sources of population data entered in the database and, as a corollary, by optimizing data processing. Various orientations have already been targeted and work is underway for some of them. We are currently working on the connection of genealogical lines interrupted by emigration. It is not uncommon for lineages to be broken off by the marriage of a couple in a locality outside Quebec, especially towards the end of the 19th century when Quebec as a whole was affected by a major wave of emigration to the United States and to other Canadian provinces. In many cases, children or grandchildren of these emigrants return to marry or settle in Quebec but there remains a gap left by the marriages outside Quebec, limiting genealogical reconstruction. Throughout the years, part of these marriages were entered in the course of research projects involving genealogical reconstruction but we now intend to proceed to their systematic integration. This represents tens of thousands of marriages recorded outside Quebec mainly in the second half of the 19th century and at the beginning of the 20th century. They come for the most part from two Canadian provinces, Ontario and New Brunswick both bordering Quebec and New England in the United States. This coverage is not exhaustive but represents a good starting point for the study of migrations, especially for the period 1840–1930.

Other information found in the registers constitutes interesting subsets to be systematically integrated to enrich life courses. We are thinking in particular of marriage licenses and documented adoptions, but also other kinds of events. Lastly, in the years 1980–1990, complementary sources such as lists of students, workers or even religious were digitized and grouped under the name 'sectoral files' without being really integrated into BALSAC. In order to build a constellation of data whose center consists of a population (and its structure) over nearly 400 years, these peripheral files are now being integrated as events in individual biographies and family histories, just like the records from the registers.

By the richness and the spatiotemporal coverage of the data it contains, BALSAC now constitutes a base on which to rely to pursue and diversify its development while continuing the integration of

data from Quebec civil records in order to get ever closer to the current period and to cover the entire population (Catholics and non-Catholics). This enrichment can be done through partnerships to bring together data of various natures such as those from historical censuses in the IMPQ or those coming from genetic research projects in i-BALSAC. Other data may be directly entered into BALSAC, for example data from the registers of French-speaking parishes in other Canadian provinces or the United States.

Other datasets of various scope and size could be connected and thus be documented by life courses and family histories. Some appear more relevant than others in the medium term, notably the causes of death, but also other documents (cadasters, directories, judicial records, hospital list, among others). These additions will enrich the family and genealogical corpus with complementary information and, conversely, provide contextual depth to the peripheral data. In the era of 'big data', BALSAC is gradually moving from a historical population database to a multifaceted infrastructure for interdisciplinary research on the Quebec population.

ACKNOWLEDGEMENTS

We express our gratitude to the Canada Foundation for Innovation, the Université du Québec à Chicoutimi and its foundation, the Université de Montréal, Université Laval and McGill University for their financial support. We also thank the research assistants, clerks and technical staff who have contributed to the development of BALSAC. We are grateful to Laurent Richard from the Centre interuniversitaire d'études québécoises at Université Laval who produced the map presented in Figure 1.

REFERENCES

BALSAC. (2020). [Web portal]. Retrieved from http://balsac.uqac.ca/

Bouchard, G. (1996). *Quelques arpents d'Amérique: Population, économie, famille au Saguenay 1838–1971*. Montréal: Boréal.

Bouchard, G., & Bergeron, M. (1975). Les registres de l'état civil de Notre-Dame de Laterrière (1855–1911). *Archives 75.3*, *3*(3), 164–173.

Bouchard, G., & de Braekeleer, M. (1991). *Histoire d'un génome: Population et génétique dans l'est du Québec*. Sillery: Presses de l'Université du Québec.

Bouchard, G., Casgrain, B., & Roy, R. (1981). *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. doi: 10.7202/303510ar

Bouchard, G., Roy, R., & Casgrain, B. (1985). *Reconstitution automatique des familles: Le système SOREP*. Chicoutimi: SOREP.

Bournival, J.-S., St-Hilaire, M., & Vézina, H. (2020). A critical assessment of historical Canadian censuses and Quebec civil registers: How linked datasets can serve as a tool to compare population microdata. *Histoire Sociale/Social History*. Manuscript under review.

Bourque, M., Markowski, F., & Roy, R. (1984). Évaluation du contenu des registres de l'état civil saguenayen, 1842–1951. *Archives*, *16*(3), 16–39.

Charbonneau, H., Desjardins, B., Guillemette, A., Landry, Y., Légaré, J., & Nault, F. (1993). *The first French Canadians: Pioneers in the St. Lawrence Valley*. Newark, London: University of Delaware Press/Associated University Presses.

Charbonneau, H., Desjardins, B., Légaré, J., & Denis, H. (2000). The population of the St-Lawrence Valley, 1608–1760. In M. R. Haines & R. H. Steckel (Eds.), *A Population History of North America* (pp. 99–142). Cambridge: Cambridge University Press.

Dickinson, J. A. (1994). Les réfugiés acadiens au Québec, 1755–1775. *Études Canadiennes/Canadian Studies*, *37*, 51–61.

Dillon, L., Amorevieta-Gentil, M., Caron, M., Lewis, C., Guay-Giroux, A., Desjardins, B., & Gagnon, A. (2018). The programme de recherche en démographie historique: Past, present and future developments in family reconstitution. *History of the Family*, *23*(1), 20–53. doi: 10.1080/1081602X.2016.1222501

Fleury, M., & Henry, L. (1956). *Des registres paroissiaux à l'histoire de la population: Manuel de dépouillement et d'exploitation de l'état civil ancien*. Paris: I.N.E.D.

Gauvin, H., Lefebvre, J. F., Moreau, C., Lavoie, E. M., Labuda, D., Vézina, H., & Roy-Gagnon, M. H. (2015). GENLIB: An R package for the analysis of genealogical data. *BMC Bioinformatics*, *16*(1), 160. doi: 10.1186/s12859-015-0581-5

ISQ, Institut de la statistique du Québec (2019). *Le bilan démographique du Québec. Édition 2019*. Québec. Retrieved from www.stat.gouv.qc.ca/statistiques/population-demographie/ bilan2019.pdf

Landry, Y. (1992). *Les filles du roi au XVIIe siècle: Orphelines en France, pionnières au Canada; Suivi d'un répertoire biographique des filles du roi*. Montréal: Leméac.

LaRose, A. (2015). Le microfilmage et la numérisation des registres paroissiaux du Québec. *L'Ancêtre*, *41*(310), 170–173.

McInnis, M. (2000a). Canada's population in the twentieth century. In M. R. Haines & R. H. Steckler (Eds.), *A population history of North America* (pp. 529–600). Cambridge: Cambridge University Press.

McInnis, M. (2000b). The population of Canada in the nineteenth century. In M. R. Haines & R. H. Steckel (Eds.), *A population history of North America* (pp. 371–432). Cambridge: Cambridge University Press.

Piché, V. (2003). Un siècle d'immigration au Québec: De la peur à l'ouverture. In C. Le Bourdais & V. Piché (Eds.), *La démographie québécoise: Enjeux du XXIe siècle* (pp. 225–263). Montréal: Presses de l'Université de Montréal.

van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS: A historical international social class scheme*. Leuven: Leuven University Press.

van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO: Historical international standard classification of occupations*. Leuven University Press.

Vézina, H., Kermorvant, C., Bonhomme, M., & Bournival, J.-S. (2019). i-BALSAC: Completing families with the help of automatic text recognition. In paper presented at the *Social Science History Association* (pp. 1–25), Chicago, USA.

Vézina, H., St-Hilaire, M., Bournival, J.-S., & Bellavance, C. (2018). The linkage of microcensus data and vital records: An assessment of results on Quebec historical population data (1852–1911). *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *51*(4), 230–245. doi: 10.1080/01615440.2018.1507771