*Le programme de reconstitution automatique des familles saguenayennes : données de base et résultats provisoires**

par Gérad BOUCHARD** et Patrick BRARD***

Ce texte a pour but de soumettre à la critique une partie des travaux méthodologiques réalisés dans le cadre du *Projet d'Histoire sociale de la Population du Saguenay*. Ces travaux ont consisté, en particulier, à mettre au point un procédé original de reconstitution automatique des familles et ils ont impliqué un certain nombre de choix techniques ou méthodologiques qui sont ici décrits. Il sera cependant utile de présenter d'abord brièvement notre projet de recherche en résumant ses principaux objectifs et ses réalisations à ce jour.

I

I. — LE PROJET D'HISTOIRE SOCIALE DE LA POPULATION DU SAGUENAY (1842-1941)[1].

Ce projet de recherche, qui est né en 1972, porte sur l'ensemble de la région du Saguenay, laquelle est située à 200 kilomètres au nord de la ville de Québec, dans la province du même nom. Ouverte officiellement au peuplement en 1842, elle compte aujourd'hui 280,000 habitants répartis entre une centaine de paroisses. Notre objectif est de constituer, pour la période 1842-1941, un registre universel, aussi exhaustif que possible et entièrement automatisé, de cette population régionale[2]. Trois sources ont été jusqu'ici mises à contribution. Il s'agit d'abord des registres de baptêmes, mariages et sépultures, dont 200 000 actes (sur un total de 340 000) ont été dépouillés et mis sur bandes magnétiques. Nous avons dépouillé en outre les recensements nominatifs réalisés pour l'ensemble des paroisses de la région par le gouvernement canadien aux années 1852, 1861 et 1871 (les recensements ne sont pas accessibles au-

* Les travaux qui ont conduit à la rédaction de cet article ont été rendus possibles grâce à l'assistance financière du Conseil des Arts du Canada (devenu depuis le Conseil de recherches en sciences humaines) et de l'Université du Québec à Chicoutimi.

** Département des Sciences humaines, Université du Québec à Chicoutimi.

*** Centre d'informatique, Université du Québec à Chicoutimi.

[1] Nous avons œuvré au départ dans le cadre de la période 1842-1911, puis 1842-1931. Des dispositions prises très récemment nous permettent désormais d'embrasser la période de 1842-1941. Rappelons, d'autre part, que l'équipe du *Projet d'Histoire sociale de la Population du Saguenay* est présentement formée de Michel Bergeron, recherchiste, de Patrick Brard et François Martin, analystes en informatique, de Yolande Lavoie et Raymond Roy, démographes, de Christian Pouyez et Gérard Bouchard, historiens.

[2] Pratiquement, ce registre informatisé prendra la forme d'un fichier-réseau (« data base ») en cours de construction.

delà) de même que la moitié environ des recensements nominatifs opérés par les curés lors de leurs visites paroissiales, ce dernier travail devant être terminé à la fin de l'année 1979.

Les données extraites de ces sources servent à créer les fiches individuelles et familiales et à y inclure, en un premier temps, des renseignements principalement démographiques. Dans une seconde étape — qui, à vrai dire, est déjà en cours — notre programme de travail prévoit y intégrer des données à caractère économique, social, culturel, médical, etc. En vue de telle opération, nous disposons déjà de deux fichiers nominatifs extrêmement riches, l'un concernant les élèves ayant fréquenté le séminaire de Chicoutimi entre 1873 et 1948, l'autre les religieux nés au Saguenay entre 1842 et 1945[3].

Parallèlement à ces travaux, nous nous sommes employés à construire à l'aide de l'ordinateur des instruments de traitement automatique des données. Dans cette direction, nos efforts portent sur les trois points suivants:

— création d'un programme de reconstitution automatique des familles;
— jumelage, par ordinateur, des dossiers de familles reconstituées et des données des recensements nominatifs;
— mise au point d'un procédé de construction automatique des arbres généalogiques.

Sur les deuxième et troisième points, les travaux sont très avancés. Sur le premier, ils sont près d'aboutir. Ce sont eux qui font l'objet du présent exposé dont le principal objectif est de présenter le programme MISTASSINI (qui consiste en fait en un ensemble de programmes interdépendants) à l'aide duquel nous nous proposons d'effectuer la reconstitution automatique des familles saguenayennes[4].

II. — LA RECONSTITUTION AUTOMATIQUE DES FAMILLES: QUELQUES DONNÉES GÉNÉRALES.

A. SCHÉMA DE LA RECONSTITUTION DES FAMILLES.

Quels que soient le procédé ou la technique adoptés, la reconstitution des familles comporte toujours deux temps, consistant d'abord dans

[3] Pour un exposé détaillé sur tout ce qui précède, se reporter à Gérard BOUCHARD, «L'histoire de la population et l'étude de la mobilité sociale au Saguenay, XIX^e-XX^e siècles», *Recherches sociographiques*, vol. XVII, n° 3 (sept.-déc. 1976), pp. 353-372; «Introduction à l'étude de la société saguenayenne aux XIX^e et XX^e siècles», *Revue d'Histoire de l'Amérique française*, vol. 31, n° 1 (juin 1977), pp. 3-27; avec Yolande LAVOIE, «Le Projet d'Histoire sociale de la Population du Saguenay: l'appareil méthodologique», à paraître à la fin de 1978 dans la *Revue d'Histoire de l'Amérique française*.

[4] Nous désirons à ce stade-ci reconnaître la dette que nous avons contractée à l'endroit du Programme de Recherche en Démographie Historique dirigé par Hubert Charbonneau et Jacques Légaré, professeurs au département de Démographie de l'Université de Montréal. Ces chercheurs et les membres de leur équipe n'ont cessé de nous prodiguer encouragements et conseils depuis les débuts de nos travaux.

la création de la fiche de couple[5] puis du dossier de famille proprement dit. Le passage subséquent du dossier de famille à la fiche individuelle est une opération distincte qui n'est pas abordée ici. Le premier volet, on le sait, est voué au rapprochement de tous les actes ou événements rattachés à un même couple tandis que le deuxième a pour objet de relier les événements se rapportant exclusivement aux enfants. Ces deux temps se marquent, bien que d'une manière implicite souvent, dans toutes les expériences connues de reconstitution, la tentative la plus explicite étant venue des chercheurs de Montréal. Dans un effort de clarification, ceux-ci ont en effet opéré une schématisation de leur démarche qui propose, notamment, une nomenclature des diverses relations à établir[6]. Nous nous en sommes inspirés pour construire l'organigramme présenté ci-contre (Figure 1). Notre schéma a été conçu de manière à illustrer très nettement les deux volets de la reconstitution. Toutes les flèches centre-périphérie désignent les relations à établir dans la construction de la fiche de couple; tandis que les flèches qui courent sur la périphérie désignent les relations à établir pour passer au dossier de famille. Les mentions relatives aux parents sont regroupées dans la partie supérieure du schéma, celles relatives aux enfants dans la partie inférieure. Au centre, dans la perspective de la construction de la fiche de couple, la mention de couple peut être tirée indifféremment d'un acte de baptême, de mariage ou de sépulture, au hasard des tris. Cependant, dans le cours du passage au dossier de famille, elle cède la place à la primimention, c'est-à-dire à la plus vieille mention du couple[7].

Ceci dit, nous nous en tiendrons ici au premier temps de la reconstitution, c'est-à-dire à la création de la fiche de couple. C'est sans aucun doute, dans notre cas tout au moins, l'opération la plus importante et la plus difficile. En regard, le passage au dossier de famille pose peu de problèmes, compte tenu de la qualité des sources sur lesquelles nous avons pu travailler (nous y reviendrons). En outre, nous pourrons bénéficier sur ce point, encore une fois, des réalisations du P.R.D.H. (programmes de jumelage des actes concernant un même individu; lien sépulture-baptême, mariage-baptême...). Nous ne parlerons pas non plus de la fiche de famille comme telle, puisqu'il est peu probable que nous y recourrions. Pour une part en effet, il nous suffit de communiquer à l'ordinateur des instructions précises pour qu'il dispose dans un ordre donné les informations qu'il reçoit et à partir desquelles il devra effectuer lui-même les diverses opérations d'analyse. Et pour le reste, il faut bien dire que cette notion de fiche de famille est rendue pratiquement inopérante du fait des prolongements qui seront donnés au dossier de famille, savoir: passage à la fiche individuelle et intégration de multiples données non démographiques. Il résultera inévitablement de ces opérations une quan-

[5] Terme que nous empruntons à la terminologie du P.R.D.H.

[6] Voir par exemple le rapport annuel du P.R.D.H. (1975-76), p. 7.

[7] Ce concept est aussi emprunté au P.R.D.H. cf. Bertrand DESJARDINS, *Élaboration des fiches de couple en vue de la reconstitution automatique des familles: application du programme Hochelaga à la population canadienne du XVII^e siècle*. Mémoire de maîtrise, département de Démographie de l'Université de Montréal, 1975, p. 37.

tité d'informations débordant les dimensions de la fiche traditionnelle ou de tout autre instrument analogue. Le problème, désormais, consiste à concevoir une structure technique appropriée à l'ensemble des données rassemblées, et ce problème sera surmonté, pour ce qui nous concerne, par le truchement du fichier-réseau évoqué plus haut.

B. Les mutations nominatives.

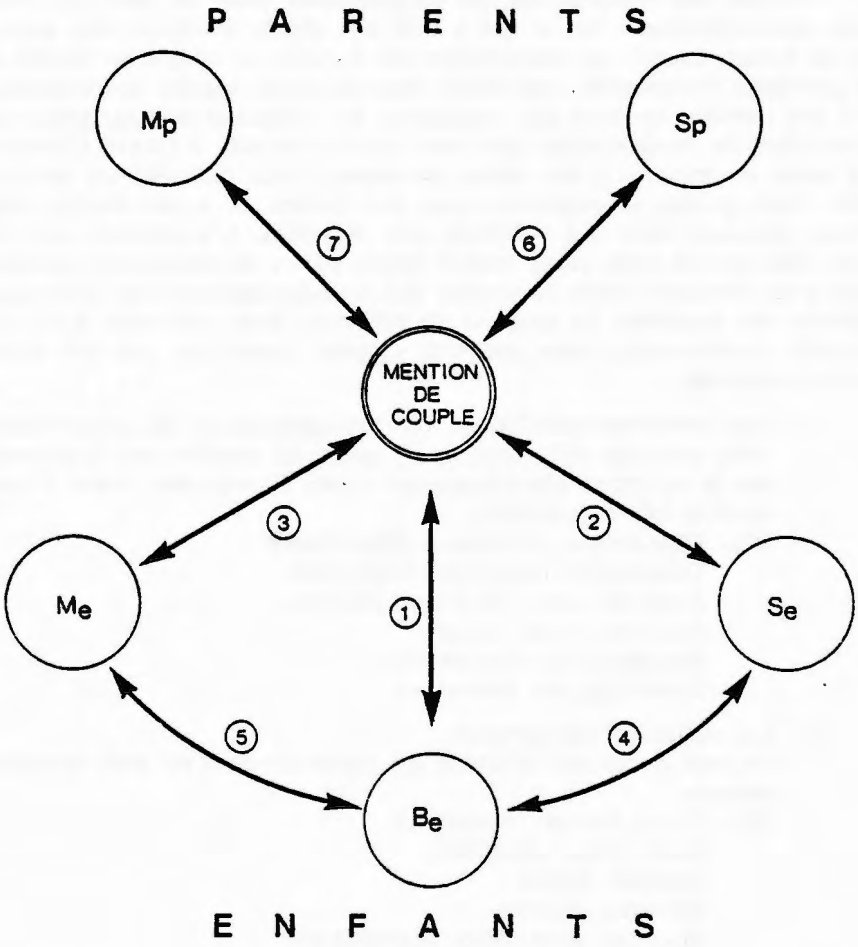
Lorsque les registres ont été correctement tenus et que leur contenu, particulièrement en ce qui a trait aux sujets d'acte et aux dates, est de bonne qualité, la construction de la fiche de couple se heurte à un problème fondamental qui réside dans ce qu'on appelle couramment, et d'une manière quelque peu restrictive, les variations orthographiques, c'est-à-dire les modifications que peut subir d'un acte à l'autre l'énoncé des noms et prénoms d'une même personne. Cette difficulté est universelle, bien qu'elle se manifeste sous des formes et à des degrés fort divers, appelant ainsi des solutions très variables. L'expression est du reste impropre et sans doute vaut-il mieux parler de mutations nominatives pour désigner toute la gamme des transformations dont sont susceptibles les mentions de noms et de prénoms. Sans prétendre à un inventaire systématique, nous pouvons signaler, parmi les cas que nous avons rencontrés:

1 – Les variations graphiques ou orthographiques au sens strict. Elles sont des déformations de mentions usuelles qui n'altèrent pas la structure phonétique des noms et prénoms, sinon d'une manière très superficielle.
EX.: Desmeules, Demeule, Des Meule
Theotiste, Theotite, Theotisse
Rose de Lima, De Lima, Delima
Allard, Alar, Alare
Boudreault, Bouderau
Saint-Gelais, Singelais

2 – Les variations phonétiques.
Ce sont celles qui affectent les phénomènes d'un nom ou d'un prénom.
EX.: Dule, Dulie, Theodule
Line, Lina, Adeline
Phébée, Bébée
Honoré, Henry
Maxime, Maximien, Maximilien
Vénal, Survenal
etc.

3 – Les patronymes composés.
Ce sont des associations de noms qui, en se rompant, font apparaître tantôt un terme, tantôt l'autre.
EX.: Georges Gauthier-Larouche
Georges Gauthier
Georges Larouche

ORGANIGRAMME DE LA RECONSTITUTION DES FAMILLES



LÉGENDE : S_p, M_p : Sépulture, mariage ou remariage des parents.
M_e, B_e, S_e : mariage, baptême, sépulture des enfants.

ou

Onésime AUDET-LAPOINTE
Onésime AUDET
Onésime LAPOINTE

4 – Les prénoms composés.
Analogue des précédents.
EX.: François-Xavier TREMBLAY
François TREMBLAY
Xavier TREMBLAY

5 – Les prénoms usuels multiples.
Ce sont des substitutions aléatoires de prénoms qui ne sont habituellement jamais associés. Nous pensons qu'elles proviennent notamment du fait que, des trois ou quatre prénoms donnés au baptême, plus d'un ont survécu à l'usage, mais en alternance. Ce cas semble donc se distinguer du précédent, lequel implique toujours des associations familières, stéréotypées (François Xavier, Louis-Philippe, Jean-Joseph, etc.). Elles ne sont pas assimilables non plus à des variations orthographiques proprement dites puisqu'il n'y a pas déformation mais substitution de formes usuelles.
EX.: Le couple
Maurice BRASSARD / Geneviève TREMBLAY
devient
David BRASSARD / Geneviève TREMBLAY
puis redevient
Maurice BRASSARD / Geneviève TREMBLAY

Les substitutions les plus inattendues se produisent parfois: Adélard pour Onésine, Charlotte pour Marguerite, etc.

Il est très important de faire remarquer que le fait d'effectuer la reconstitution par ordinateur ou manuellement ne change rien à la nature du problème: il faut, dans l'un et l'autre cas, mettre au point des procédés de réduction, de standardisation ou de « normalisation » des formes nominatives qui permettent de réunir à coup sûr sous un même vocable toutes les mentions relatives à une même personne, et ce au gré des critères et de règles explicites, scientifiquement établies. Sous ce rapport, on dispose d'ores et déjà de plusieurs modèles, plus ou moins formalisés et rigoureux, de standardisation ou de réduction appuyés sur des critères inégalement explicités. Standardisation à priori (Perrot-Daubèze) ou a posteriori, pouvant faire intervenir des éléments soit de codification (Russel, Henry, Blayo), soit de probalités et de pondération (Newcombe, P.R.D.H., Skolnick) ou les deux à la fois, toutes ces formules possèdent leurs avantages et leurs limites et il serait téméraire d'en livrer ici une évaluation tant le contexte de chaque recherche est ici souverain, savoir: la nature et les objectifs de l'enquête, l'état des sources, les ressources financières, etc. Pour cette raison, il est sans doute opportun de commencer l'exposé de notre procédé par une description sommaire des conditions qui ont entouré notre enquête et en particulier des sources que nous avons pu utiliser.

III. — LES DONNÉES SAGUENAYENNES.

Notre recherche a bénéficié de ressources financières relativement substantielles (environ $200,000 depuis 5 ans) et d'un accès facile à l'ordinateur. Le fait de travailler en histoire contemporaine nous a aussi valu de précieux atouts, à commencer par des sources d'une très grande qualité.

1. La qualité des données.

Tous les tests que nous avons faits jusqu'ici (sur plusieurs milliers d'actes) pour évaluer sommairement le contenu des registres ont révélé que le % d'omissions portant sur les dates et les mentions nominatives (sujets d'acte, parents et conjoints) était très inférieur à 1 pour les trois types d'actes[8]. Les actes sans mention de couple se trouvent aussi dans une très faible proportion, ce qui fait évidemment espérer un très fort rendement dans la création des fiches de couples. Cette donnée revêt la plus haute importance, eu égard au choix du procédé de reconstitution. Des actes incomplets, pauvres en mentions de couples, obligent à un jumelage ponctuel consistant à rapprocher les actes deux à deux (ex.: jumeler des actes de sépultures à des actes de mariage ou des actes de mariage à des actes de baptême, etc). Dans le cas contraire, qui est le nôtre, les données autorisent un jumelage en grappe, à partir des mentions de couples, autour desquelles les actes viennent s'agglomérer. Cette deuxième voie, lorsqu'elle peut s'appuyer sur des registres d'une excellente qualité, a l'avantage d'être plus précise et plus efficace[9].

2. Les sources d'appoint.

Il faut signaler aussi l'abondance et la richesse des sources d'appoint, c'est-à-dire les recensements systématiques réalisés par les soins du clergé et du gouvernement, les rôles d'évaluation, les listes d'arpentage, de souscription, etc. À quoi il faut ajouter un recours assez inusité mais

[8] C'est un résultat sommaire obtenu par des sondages. L'étude systématique et approfondie des sources et la critique des données sont en cours. Pour un bref aperçu, voir Gérard Bouchard et Michel Bergeron, « Les registres de l'état civil de Notre-Dame de Laterrière (1855-1911) », *Archives 75.3*, VII, 3 (sept.-déc. 1975), pp. 164-173.

[9] C'est cette même voie que le P.R.D.H. a empruntée (voir Pierre Beauchamp, Raymond Roy, Jacques Légaré, « Reconstitution automatique des familles par le programme Hochelaga II », *Population et famille*, no. 33, 1974, pp. 1-40). Mais il est bien évident que le choix des démographes de Montréal, ainsi que le nôtre, n'ont été possibles que grâce à la qualité des registres catholiques québécois. Ailleurs, en Angleterre et en Italie par exemple, la forte proportion des actes sans mention de couple contraint à des procédés très complexes et pose des difficultés parfois insurmontables. Là-dessus, cf. R.-S. Schofield, « La reconstitution de la famille par l'ordinateur », *Annales E.S.C.*, vol. 27, nos. 4-5 (juil.-oct. 1972), pp. 1071-1082; (avec E.A. Wrigley), « Nominative Record Linkage by Computer and the Logic of Family Reconstitution », in Schofield et Wrigley, *Identifying People in the Past*, London, 1973, 159 pages; M.H. Skolnick, *The Construction and Analysis of Genealogies from Parish Registers with a Case Study of Parma Valley*, Italy, Ph. D. dissertation, Stanford University, 1974 (disponible sur micro-film à l'Université du Michigan, Ann Arbor).

dont nous avons vérifié la très grande efficacité, en l'occurrence l'enquête téléphonique. Il est aisé en effet de colliger en les confrontant des témoignages très fiables sur des événements qui se sont produits même au XIX^{e} siècle.

3. Les données nominatives.

Ici encore le fait de travailler sur une période récente nous simplifie quelque peu la tâche. Il a déjà été constaté[10] que, eu égard aux siècles antérieurs, les XIX^{e} et XX^{e} siècles se caractérisent par une relative fixation des données nominatives. Par rapport aux données de la Nouvelle-France traitées par le P.R.D.H. par exemple, nous observons un phénomène analogue, même si les mutations demeurent importantes.

Enfin, le Saguenay a toujours eu la réputation (laquelle s'avère parfaitement fondée, du reste) d'être une région à très forte consanguinité, ceci étant dû à l'isolement où s'est trouvée sa population pendant plus d'un siècle. Cet élément laissait entrevoir, aux débuts de nos travaux, une fréquence élevée d'homonymie parfaite et faisait craindre des embûches de taille pour la reconstitution des familles. Les données dont nous disposons maintenant sur ce sujet établissent que cette crainte était nettement exagérée[11]. Il semble que les vieux noyaux de population sédentaire qui contribuent à perpétuer et à diffuser certains patronymes font en quelque sorte illusion, masquant la grande diversité causée par la mobilité générale de la population.

Les caractéristiques qui viennent d'être évoquées paraissaient donc autoriser la construction d'un procédé original et simple de reconstitution automatique, axé sur la recherche d'une très grande efficacité et d'une précision supérieure.

IV. — LE PROCÉDÉ DE RECONSTITUTION.

En fait, ce dernier est constitué de quelques programmes essentiels qui servent à entrer et valider les données de même qu'à extraire, trier et comparer les mentions de couples, plus quelques programmes annexes de listage et de statistiques diverses. Ils interviennent à chacune des étapes de la reconstitution, dont voici les principales[12].

[10] Yves Blayo, «Name variations in a village in Brie, 1750-1860», in E.-A. Wrigley, *Identifying People in the Past*, 1973, p. 59.

[11] Nous en ferons état dans une publication ultérieure. Mentionnons seulement que le prénom le plus répandu (Louis) ne représente pas plus de 3% des occurrences. Le patronyme le plus fréquent (Tremblay), guère plus de 10%.

[12] La présentation qui suit s'en tient forcément aux grandes lignes. Le lecteur désireux d'obtenir plus de détails est prié de communiquer avec les auteurs.

A. La structure du fichier.

Toutes les données sont d'abord portées sur disques puis, après correction et validation automatique[13], transportées sur bandes. Elles sont cependant ramenées sur disques pour les opérations de traitement, lesquelles sont exécutées sur l'ordinateur de l'Université du Québec à Chicoutimi (Honeywell Xérox 560).

La première opération consiste à préparer les actes de la banque en vue du jumelage. Chacune des mentions de couple qui y sont contenues est prélevée et se voit accoler un numéro qui l'identifie. En même temps, la première mention de couple prélevée dans chaque acte donne son numéro à cet acte. Les autres mentions tirées du même acte porteront à la fois leur propre numéro et celui de l'acte. De cette façon, il est possible de repérer aisément les mentions de couples (leur numérotation est séquentielle) et les actes d'où elles sont extraites. En outre, chaque mention de couple reçoit un troisième numéro, de 1 à 5, qui renvoie à une échelle de classement telle que:

— dans les actes de baptême
  1: renvoie au couple père-mère
— dans les actes de sépulture
  1: renvoie au couple père-mère (cas de sépulture d'un enfant ou d'un célibataire)
  2: renvoie au couple formé par le sujet de l'acte et son conjoint vivant ou défunt
— dans les actes de mariage
  1: renvoie au couple époux-épouse
  2: renvoie aux père et mère de l'époux
  3: renvoie aux père et mère de l'épouse
  4: renvoie au couple formé par l'époux et son ex-conjointe
  5: renvoie au couple formé par l'épouse et son ex-conjoint.

Cette dernière numérotation permet une meilleure connaissance du fichier et elle a prouvé son utilité au moment de dresser la statistique des mentions de couple, d'orienter les tris requis par le jumelage ou de reformuler les règles commandant la formation des paires. Au premier jet, le fichier prend donc la forme d'une liste de mentions alignées comme suit:

[13] Toutes les étapes préalables (dépouillement, vérification, entrée des données, etc.) ont déjà fait l'objet d'une présentation; voir Gérard Bouchard et Yolande Lavoie, «Le Projet d'Histoire sociale de la Population du Saguenay: l'appareil méthodologique», à paraître durant l'année 1978 dans la *Revue d'Histoire de l'Amérique française*. Pour un exposé plus détaillé de certaines questions à caractère technique, cf. Gérard Bouchard et Yolande Lavoie, *Description des opérations relatives à la saisie et à l'élaboration des données*, Document de travail n° 27, Projet d'Histoire sociale de la Population du Saguenay, 18 pages.

Numéro de la mention	Numéro de l'acte	Classement	Mention
442	442	1	Étienne GIRARD / Marie TREMBLAY
443	443	1	Philippe JEAN / Hélène SAVARD
444	443	2	Jacques JEAN / Lise BOULIANE
445	443	3	Sylvain SAVARD / Line GIRARD
446	446	2	Gilles RINFRET / Claudette MORIN
Etc.			

Au cours de cette étape, les actes sans mention de couple, c'est-à-dire ceux qui ne contiennent pas au moins 3 éléments nominatifs sur 4 (un prénom et deux noms, deux prénoms et un nom) sont rejetés. Dans la plupart des cas, ce sont des actes de baptême où le père est dit inconnu ou des actes de sépulture de veufs (ves) dont l'ex-conjoint (e) n'est pas mentionné (e). Ces actes seront récupérés et traités au dernier stade de la reconstitution. Mentionnons d'autre part que ce fichier est directement accessible en partant des numéros de mention; il permet donc constamment des ajouts et des retraits à volonté.

B. Le problème du jumelage.

L'opération subséquente consiste à rapprocher, au moyen de différents tris, toutes les mentions qui concernent un même couple. L'ordinateur est en mesure, du moins en principe, de former toutes les paires concevables, certaines devant évidemment être refusées et d'autres acceptées. Les paires parfaitement identiques sont aussitôt retenues, cela va de soi. Mais toutes les autres, celles qui diffèrent sur un élément ou davantage, posent un problème considérable qui s'énonce très simplement. Soit deux mentions de couple non identiques: comment déterminer avec certitude si la différence est imputable à une mutation nominative — auquel cas les deux mentions doivent être fondues — ou au fait qu'il s'agit bel et bien de deux couples différents?

Les situations les plus diverses peuvent se présenter. De ce point de vue, le fait de travailler sur un passé récent et d'avoir une très bonne connaissance de la population étudiée (ce qui est le cas pour quelques membres de notre équipe) constitue un avantage énorme qui permet souvent d'éviter de graves confusions. Ainsi, Trembley et Tremble sont très évidemment des déformations de Tremblay. De même Godreau et Godereau sont à coup sûr des variations de Gaudreault, et Singelais un dérivé de Saint-Gelais. Mais on aurait certainement tort d'assimiler Duchene à Deschenes, Taillon et Tallon, Guy et Guay, Morin et Marin. D'autre part, le vocable Line peut servir de diminutif à plus d'un prénom (Adeline, Obeline, Évangeline...) et donc désigner des personnes différentes. Meunier et Minier sont synonymes en certaines occasions seulement; de même Laure et Lord, Mayrand et Myrand, Morais et Morin, etc.

Le procédé que nous avons élaboré consiste en une série d'opérations au cours desquelles les mutations nominatives sont graduellement dépistées et assimilées à des mentions standardisées. D'entrée de jeu en effet, nous avons postulé que la quasi-totalité des variations pouvaient

être reconnues et éliminées, le jumelage devant alors consister à ne regrouper que des paires parfaitement identiques, c'est-à-dire des paires de type S1 (cf. Tableau 1).

Tableau 1[14]

NOMENCLATURE PROVISOIRE DES PAIRES DE MENTIONS DE COUPLE.
prénom(s) *identique*(s)

	2	1	0
2	Situation 1 (S1)	Situation 2 (S2)	Situation 3 (S3)
1	Situation 4 (S4)	Situation 5 (S5)	Situation 6 (S6)
0	Situation 7 (S7)	Situation 8 (S8)	Situation 9 (S9)

NOM(S) IDENTIQUE(S)

Afin d'établir les chances réelles d'une telle démarche, un premier essai a été fait à partir d'un échantillon de 1682 actes tirés du registre de Laterrière et traités manuellement (il sera plus longuement question de ce test un peu plus loin). Sans effectuer aucun travail sur les variations nominatives et en ne jumelant que les paires S1 à l'état brut, l'ordinateur a pu réaliser les deux tiers des jumelages. À ce stade et étant donné les conditions de l'expérience, ce résultat a été jugé satisfaisant.

La recherche a alors porté sur les 33% de paires non jumelées, celles-ci comprenant une proportion indéterminée de paires non jumelables. En d'autres mots, ces deux sous-ensembles confondus constituaient une liste de paires S2, S3, S4, etc. dont certaines pouvaient être transformées en S1 par standardisation, et d'autres non. Il nous a semblé que ce travail devait être accompli par étapes, au moyen de tables d'équivalences.

C. La table d'équivalences universelles.

Nous appuyant sur la connaissance que nous avons des prénoms et patronymes usuels saguenayens et sur les relevés des variations orthographiques effectués par les releveurs (chacun, au cours du dépouillement, devait dresser un répertoire au jour le jour), une première table d'équivalences a été construite, dont l'objet était d'éliminer *les variations les plus mineures et les plus évidentes*, c'est-à-dire celles qui n'altèrent pas la structure phonétique des noms et prénoms, sinon d'une manière très superficielle.

[14] Ce tableau reproduit les différentes situations qui peuvent se produire lorsque deux mentions de couples sont rapprochées. Il va de soi que les situations 1, 2, 3, 4, 5 sont plus susceptibles de donner lieu à un jumelage que les situations 6, 7, 8, ou 9. D'autre part, signalons que cette nomenclature est donnée ici à titre provisoire. Elle est appelée à être transformée pour faire place à la variable époux-épouse, dans la perspective, notamment, du jumelage des dossiers de familles avec les ménages des recensements nominatifs.

Ex.: ACHILLE: Achil, Achile, Achilles, Archil
BELZEMIRE: Belzimire, Balsamir, Belsimyre
LECLERC: Leclair, Leclere, Leclaire
GUILBEAU: Guilbault, Guilbaut, Guillebeau, Guilbot
GILBERT: Gilber, Gillebert, Gilberre
ROCHEFORT: Rechefort, Rocheford, Rochford
Etc.

Cependant, pour éviter les dangers inhérents à cette première étape au cours de laquelle certaines formes nominatives sont standardisées à priori, nous avons procédé d'une manière extrêmement conservatrice, nous assurant en tout temps de ne pas poser d'équivalence qui risquent de fusionner des prénoms ou des noms distincts.

Cette première table permet de transformer en paires S1 une proportion importante des paires non jumelées au premier essai. Sa construction s'effectue au moment de la validation automatique des données nominatives et elle comprend aujourd'hui plusieurs centaines de noms et de prénoms avec leurs variations correspondantes. Enfin comme il est toujours possible, malgré les précautions prises, que cette table contienne de fausses équivalences, des tests de cohérence très serrés sont appliqués par ordinateur sur tous les dossiers de famille créés au cours de cette étape[15].

D. LA TABLE D'ÉQUIVALENCES AD HOC.

Pour ce qui concerne les variations plus prononcées dont nous avons donné quelques exemples plus haut (variations phonétiques), il n'était certes plus possible de procéder a priori. D'autre part, en vertu de ce qui a été dit sur la relative fixation des données nominatives saguenayennes et sur l'excellente tenue des registres, il paraissait vraisemblable de croire qu'en travaillant exclusivement sur les paires S2 et S4, on avait des chances de dépister la plupart des variations restantes[16]. Partant d'une liste de toutes les mentions de couple (dont certaines étaient des têtes de famille) comprises dans l'échantillon, nous avons donc isolé toutes les paires S2 et S4, c'est-à-dire celles pour lesquelles un seul élément nominatif sur quatre faisait défaut. Puis, pour chacune d'entre elles, nous avons examiné les deux mentions divergentes (noms ou prénoms).

Ex.: CÔTÉ Joseph / TREMBLAY Zébie
CÔTÉ Joseph / TREMBLAY Eusébie

SIMARD Claude / BOIVIN Phrasie
SIMARD Claude / BOIVIN Euphrasine

[15] Appliqués par ordinateur sur l'ensemble des dossiers de familles, ces tests visent à détecter des contradictions ou des invraisemblances créées par de faux jumelages (ex.: actes de naissance antérieurs à l'acte de mariage des parents, deux naissances en moins de 210 jours, deux décès d'une même personne, etc.).

[16] Il était peu probable en effet qu'après l'application de la première table, il subsistât dans le fichier une quantité de mutations capables de transformer plusieurs paires S1, en paires S5, S6 ou S7, par exemple. Nous reviendrons sur ce point.

MEUNIER Charles / LAVOIE Émilienne
MINIER Charles / LAVOIE Émilienne

Nous avons donné ces mentions comme équivalentes à chaque fois qu'elles montraient une affinité orthographique ou phonétique accusée, à l'aide d'un code à trois valeurs[17]. Ces équivalences ne sont cependant pas universelles. Elles ne s'appliquent que lorsque les trois autres membres de la paire sont parfaitement identiques; c'est pourquoi elles sont dites ad hoc.

Il a fallu nous prémunir ici contre deux excès opposés, l'un consistant à jumeler trop et l'autre trop peu. Nous examinons tout de suite le premier. Il est possible en effet qu'ici encore, nous ayons établi de fausses équivalences, ce qui aurait pour effet de fusionner des familles distinctes. Mentionnons tout d'abord que, pour cette raison, tous les jumelages obtenus par application de la table ad hoc sont soumis aux tests de cohérence. En outre, nous avons cru éviter de faire de la table ad hoc un code trop rigide en affectant chaque équivalence d'un numéro qui désigne un degré de fiabilité ou de certitude. Par exemple, il est prévu que les tests de cohérence sont d'abord appliqués aux dossiers contenant un ou des jumelages où sont intervenues des équivalences de type 3, c'est-à-dire celles qui faisaient naître un doute. Toute erreur systématique peut être rattrapée de cette façon. Enfin, divers facteurs font que la construction de la table ad hoc est une opération relativement sûre. Par exemple, l'essai qui a été fait sur les 20,000 premières mentions de couple de la banque a montré que, dans la grande majorité des cas, ce sont les mêmes équivalences qui reviennent, ce qui accélère d'autant le travail de lecture et en même temps permet de confirmer les décisions antérieures. Aussi, la fréquence d'apparition des mentions impliquées pèse dans les décisions: par exemple, il y a présomption d'équivalence lorsque, de deux formes parentes, l'une compte 5 ou 10 occurrences et l'autre 1 seulement[18].

E. LE TEST SUR LES REGISTRES DE LATERRIÈRE.

Les jumelages effectués au cours de ces deux premières étapes sont considérés comme définitifs, bien que, par mesure de précaution, nous tenions à les identifier dans la banque. Il est, par suite extrêmement important de connaître la proportion des mentions de couples qui ont été jumelées à ce stade. Toute notre méthode repose en effet sur le parti qu'après application des deux premières tables, un très fort pourcentage des mentions jumelables ont effectivement été réunies; on espère alors qu'un résidu négligeable sera traité par ordinateur s'il est possible, et manuellement dans le cas contraire. L'expérience qui a été faite à partir d'un échantillon d'actes de la paroisse de Laterrière avait précisément

[17] Chaque équivalence ad hoc se voyait ainsi affectée d'un coefficient indiquant un degré de certitude, lui-même fondé sur le degré de ressemblance entre les mentions.

[18] Les listes font toujours apparaître, au bout de chaque mention, le nombre d'actes qui y sont rattachés.

pour but de mesurer le rendement des tables d'équivalences universelles et ad hoc ainsi que d'évaluer le nombre de paires résiduelles devant être analysées en troisième étape.

Cet échantillon, composé de 1682 actes de baptêmes, mariages et sépultures, a livré 1980 mentions de couple. Le rendement de la reconstitution automatique a pu être établi par comparaison avec un jumelage manuel qui avait été préalablement effectué. À la suite de trois opérations successives, nous avons mesuré la proportion des paires jumelables qui avaient bel et bien été rassemblées. Au cours de la première opération, l'ordinateur eut comme instruction de ne jumeler que les paires S1 «naturelles», l'ensemble du fichier n'étant d'aucune manière pré-traité. Une deuxième reconstitution fut réalisée après application de la table universelle et une troisième après celle de la table ad hoc. Les résultats furent les suivants:

opérations	*Paires jumelées (%)*
1	66
2	88
3	97.5

Les opérations 1 et 2 ont été aussi réalisées sur un échantillon de 20,000 mentions de couple prélevées au hasard; les résultats ont été de 68% et 85% respectivement[19]. L'application de la table ad hoc devrait hausser ce rendement à un niveau sensiblement égal à celui qui était atteint dans le test sur Laterrière. Nous avons donc retenu de tout cela que le procédé était bon et qu'il était utile de le pousser plus avant.

F. Les tâches prochaines.

1. *Le traitement du résidu.*

Nous évoquions plus haut deux excès dont nous devons nous garder. Le premier consistait à adopter des critères trop lâches et à opérer de faux jumelages; le deuxième, au contraire, consiste à resserrer tellement les règles qu'une proportion de mentions relatives à un même couple ne peuvent être jumelées. C'est ce dernier problème que nous abordons maintenant. Il comporte deux aspects. D'une part, les deux premières tables sont construites d'une manière très conservatrice et un certain nombre de variations risquent de leur échapper. D'autre part, il ne faut pas exclure que, l'instabilité des données nominatives saguenayennes pouvant avoir été sous-estimée, les variations gonflent artificiellement non seulement le nombre de paires S2 et S4 mais aussi celui des paires S3, S5 ou S7. Il faudrait, le cas échéant, libéraliser considérablement le mode de construction de la table ad hoc, en particulier. Des essais, en rapport avec ces deux points, sont en cours et il n'est pas possible pour l'instant d'avancer quoi que ce soit de définitif. Nous nous en tiendrons ici à des hypothèses.

[19] Il va de soi que le contrôle du rendement n'a pas pu s'appuyer dans ce cas sur un fichier manuellement reconstitué. Nous avons donc considéré uniquement la proportion des mentions jumelées, supposant que le % des mentions jumelables était sensiblement le même que dans le premier échantillon.

Sur le premier point, les perspectives sont les suivantes. Il est d'abord réconfortant de constater que les limites de notre procédé ne menacent guère d'entraîner des dégâts considérables, étant donné les dimensions restreintes du résidu des paires S2 et S4 — entre 2 et 3%. Sur l'ensemble de la banque, qui comptera au total quelque 400,000 mentions de couple, il faut donc prévoir environ 10,000 mentions à traiter, par ordinateur ou manuellement, au cours de cette troisième étape. D'autre part, nous avons la conviction qu'il est facile d'abaisser encore sensiblement ce chiffre par le biais d'opérations spécifiques portant chaque fois sur telle catégorie de mentions ou tel cas particulier. Par exemple, il est possible de:

a) traiter séparément les paires impliquant des patronymes ou des prénoms composés;
b) orienter la recherche de certaines mentions à partir d'une liste d'actes manquants dans les dossiers de famille (ex.: cas de parents dont on a l'acte de baptême mais non l'acte de mariage, cas d'un fils dont on a l'acte de mariage main non l'acte de baptême, cas d'un enfant décédé dont on n'a pas l'acte de baptême, cas d'un acte de baptême dans lequel le père ou la mère est dit (e) veuf (ve) alors qu'on n'a pas l'acte de sépulture du défunt, etc.;
c) écarter du résidu certaines mentions qui ne peuvent être qu'isolées et par conséquent non jumelables (cas de certaines mentions de conjoints dérivées d'un acte dont ils ne sont pas les sujets).

Notons que ces opérations ne risquent guère d'introduire un enchaînement de faux jumelages dans le fichier puisqu'elles porteront à chaque fois sur une partie seulement du résidu et qu'elles seront aussitôt suivies d'un ensemble de validations prenant diverses formes, telles:

a) application des tests de cohérence;
b) contrôle des nouveaux jumelages par le biais des recensements nominatifs;
c) enquêtes téléphoniques;
d) croisement avec le fonds Larouche-Simard (il s'agit d'un fonds généalogique qui a été mis à notre disposition et qui couvre la période 1842-1870).

Parallèlement, ce travail permettra de mettre à jour une série d'équivalences demeurées jusque là inaperçues. Elles seront colligées dans une troisième table, a postériori, laquelle pourrait s'avérer d'une grande utilité au cours du jumelage des dossiers de familles avec les fiches de ménages tirés des recensements.

Nous en arriverons donc, par approches successives, à constituer un mince noyau de mentions irréductibles (représentant peut-être 1% de l'ensemble?) qui seraient simplement exclues de la banque, en fin de parcours.

2. *La construction automatique des tables d'équivalences.*

Les tables d'équivalences universelles et ad hoc décrites dans les pages précédentes ont été mises à l'essai sur deux échantillons et ont

donné des résultats plus que satisfaisants. Elles ont cependant été construites selon un procédé semi-automatique qui, tout en étant soumis à des normes strictes, n'en laisse pas moins place à certains choix où peut s'introduire une part d'arbitraire. C'est en outre un procédé relativement long. Nous avons cru bon de supprimer ces deux inconvénients en construisant un programme qui permettra à l'ordinateur de construire lui-même ces deux tables au fur et à mesure de la reconstitution. Ce travail arrive à sa phase finale et il donnera lieu prochainement à une publication.

Nous approchons donc de l'objectif visé, savoir la création d'un procédé entièrement automatique de reconstitution, tel que celle-ci impliquerait un minimum de travail manuel préalable et reposerait sur des méthodes aisément transposables à d'autres banques de données et à d'autres contextes. À ce point, la voie sera ouverte à la fois vers l'élaboration d'un registre ou fichier individuel intégré à caractère universel, et vers la construction automatique des généalogies à l'usage de la génétique.