PROJET BALSAC

Volet : Infrastructures

Un aperçu des méthodes de reconstitution des familles et de jumelage des données du fichier de population BALSAC

par

Mario BOURQUE

avec la participation de
Bernard Casgrain, Michèle Jomphe, Manon Bouchard,
Sophie Claveau, Marc Tremblay

(**Document no : I-C-231**)

(adresse postale)

Projet BALSAC
Université du Québec à Chicoutimi
555 boulevard de l'Université, Chicoutimi (Québec) G7H 2B1
Téléphone: (418) 545-5517
Télécopieur: (418) 545-5518
courriel : balsac@uqac.ca

Le fichier BALSAC est une banque de données informatisées permettant la construction des histoires familiales, individuelles et des généalogies à l'échelle de la population du Québec depuis le 17e siècle. Les données proviennent de l'état civil et se prêtent à des domaines d'exploitation partagés entre plusieurs champs disciplinaires tels que la génétique humaine, la démographie, la géographie, la sociologie et l'histoire. La première phase de construction du fichier fut consacrée à la population de la région du Saguenay-Lac-St-Jean. Environ 660 000 actes de baptême, de mariage et de sépulture enregistrés dans cette région de 1838 à 1971 ont été saisis et jumelés (Bouchard 2008). Le fichier BALSAC fut par la suite progressivement étendu à l'ensemble du territoire québécois, pour les 19e et 20e siècles, en restreignant la saisie et le jumelage des données aux actes de mariage seulement, sauf pour quelques régions pour lesquelles les actes de baptême et de sépulture étaient déjà informatisés (Charlevoix et les Iles-de-la-Madeleine en partie). À ce jour, la saisie et le jumelage de presque tous les actes de mariage de confession catholique du Québec de 1800 à 1940 (soit environ 1,25 million d'actes) est complétée.

Ces travaux ont nécessité l'élaboration de divers programmes informatiques de mises en forme et de procédures de jumelage des données, dont l'aboutissement fut la création d'un système de reconstitution automatique des familles, qui permet de réunir dans un même dossier tous les renseignements relatifs à l'histoire d'une union (Bouchard et al. 1985a). Ce document fournit un aperçu des procédés utilisés et de leur efficacité.

Principes généraux du jumelage des données

Les conditions dans lesquelles s'effectue le processus du jumelage varient selon le type de données utilisées, la dimension de la banque de données soumise à la reconstitution et les objectifs de recherche. Fondamentalement, le travail de jumelage consiste à effectuer des rapprochements entre des dossiers pour créer des liens et ultimement réunir tous les éléments d'un même ensemble. La nature et la qualité des données sont d'une importance capitale pour les choix méthodologiques inhérents au système de jumelage.

Les données de base

La source première du système de reconstitution des familles sont les actes d'état civil du Québec. En plus des noms et prénoms des participants à l'événement, l'acte contient diverses informations comme l'état matrimonial, le lieu de résidence, la profession, l'aptitude à signer, etc. La qualité des registres québécois a fait l'objet d'évaluations (Bouchard et Larose 1976; Roy et Charbonneau 1976; Bourque et al. 1984; Bourque et Bouchard 2003). Ces études montrent que bien qu'il y ait des variations, dans le temps et dans l'espace, de la qualité de l'enregistrement, la tenue des registres en général semble très bonne, sauf en ce qui concerne les actes de confessions autres que catholique. Une analyse portant sur un échantillon d'actes de religions non catholiques pour le 19e siècle montre que non seulement la tenue de ces registres était de qualité inférieure, mais qu'ils étaient somme toute inutilisables à des fins de jumelage (Bourque et Bouchard 2003).

La saisie des données

La saisie des données consiste à faire la lecture et l'informatisation des renseignements contenus dans les actes de l'état civil, en vue d'alimenter la banque de données. L'informatisation suppose

une certaine mise en forme des données, puisque bon nombre d'informations sont codées afin d'en accélérer la saisie et de faciliter la correction. La saisie se fait en lot, les actes étant classés par paroisse et par ordre chronologique, afin d'éviter des erreurs sur des données répétitives. Plusieurs dispositions sont prises pour s'assurer que les informations colligées soient fidèles à la source. Le programme de saisie comprend des instructions permettant de réguler le travail des préposés à la saisie. Un système de contrôle des relevés permet de surveiller la vitesse et la qualité du dépouillement réalisé et d'apporter rapidement les correctifs nécessaires. Les résultats de ces contrôles permettent aussi de calibrer les interventions des préposés à la validation.

Les informations transcrites à partir des actes servent de matière première pour la construction du fichier. La bonne tenue des registres d'état civil est donc étroitement liée à la réussite des opérations. Si le contenu de l'état civil saguenayen est représentatif de l'ensemble des registres québécois, on peut conclure que l'enregistrement des caractéristiques afférentes aux diverses opérations de la reconstitution des familles est très satisfaisant (Bourque et al. 1984).

La reconstitution automatique

La reconstitution automatique des familles suppose des opérations destinées à lier et traiter les renseignements relatifs à l'histoire d'un couple. Le jumelage est donc fondé presque exclusivement sur la comparaison de données nominatives. Le procédé suit une démarche progressive, des liens les plus faciles aux liens les plus difficiles. Les règles et les conditions de jumelage varient selon la complexité des situations, déterminée par la fréquence et l'interaction des variations nominatives et de l'homonymie. Un même nom ou un même prénom peuvent en effet se présenter sous des graphies différentes, voire complètement distinctes. L'étude des données saguenayennes a révélé l'ampleur du problème que posent les variations nominatives (Bouchard et al. 1985b). Deux types d'intervention ont été préconisés pour surmonter ce problème : un traitement des données brutes en amont du jumelage (opérations préliminaires) et l'utilisation d'instruments de comparaison opérant dans la mécanique du jumelage pour mesurer la similitude entre les données nominatives. Pour ce qui concerne l'homonymie, qui conduit à confondre et jumeler des couples distincts, l'ampleur du problème est liée à la population étudiée, à sa dimension et surtout aux procédés de jumelage. Des règles de jumelage trop permissives peuvent atténuer le pouvoir discriminant des données nominatives et amener à créer de faux liens.

La mention de couple

Le système de jumelage consiste à regrouper dans un même enregistrement toutes les mentions relatives à un même couple, de manière à créer les fiches de couple ou biographies familiales. Chaque mention de couple comprend quatre éléments nominatifs, soit les nom et prénom(s) du conjoint et les nom et prénom(s) de la conjointe. Dans un acte de mariage, par exemple, les possibilités de mentions de couple sont : époux-épouse, père et mère de l'époux, père et mère de l'épouse, époux et ex-conjointe, épouse et ex-conjoint. Le fait de considérer la mention de couple comme unité de comparaison, de préférence à une unité basée sur des informations strictement individuelles, minimise l'impact négatif des variations nominatives sur la réussite du jumelage.

Le système SOREP

Le système SOREP est la pièce maîtresse de ce qu'on appelle aujourd'hui le système de reconstitution des familles BALSAC. À l'origine, ce système a été élaboré en partant du principe que les trois types d'actes (baptême, mariage et sépulture) serviraient de matière première pour la composition du fichier. Le système comporte plusieurs étapes réparties entre deux grands sous-ensembles, soient les opérations préliminaires et le jumelage comme tel.

Les opérations préliminaires

Ces opérations visent à préparer les données brutes en vue du jumelage. Deux programmes interviennent : Fonem et Nettoie. Le programme Fonem (Bouchard et al. 1985a, pp. 118-125) est un code phonétique qui permet de supprimer les variations nominatives sans altérer la structure phonétique. Il opère sur chacun des éléments nominatifs avant que les mentions de couple soient formées. Il aide à éliminer les variations nominatives les plus superficielles dans le but d'améliorer le rendement des opérations de jumelage. Le programme Nettoie (Bouchard et al. 1985a, pp. 97-100) élimine les mentions de couple jugées inadmissibles parce qu'elles ne comportent pas les quatre éléments nominatifs attendus. Ces mentions, peu nombreuses, sont isolées pour être traitées manuellement à la toute fin des travaux de jumelage.

Le jumelage

Le deuxième sous-ensemble d'opérations procède à la construction des fiches de couple. Les jumelages réalisés par les programmes de reconstitution sont basés sur la comparaison des mentions de couple. Le corpus de base comprend toutes les mentions se rapportant à une population visée pour une période donnée. Le but est de regrouper dans un même enregistrement les diverses mentions se rapportant à une même famille : mariage, sépulture et remariage des parents, naissance, mariage et décès d'enfants. La construction des fiches de couple est réalisée en deux étapes qui renvoient à diverses routines et opérations.

La première étape, appelée les Tris élémentaires, sert à traiter les variations nominatives les plus courantes. Il s'agit de créer des sous-ensembles qui permettent de comparer toutes les mentions candidates à un jumelage. Il y a sept Tris élémentaires. Le premier de ces tris, appelé le « 4x4 », rassemble les mentions de couple dont tous les éléments sont parfaitement identiques. Les six autres tris consistent à assembler des mentions dont deux ou trois paires d'éléments sur quatre sont identiques, suivant toutes les combinaisons d'appariement possibles. Au terme de chaque tri, une comparaison s'effectue à l'aide de programmes conçus pour détecter et mesurer des degrés et des formes de similitude entre deux noms ou deux prénoms (Bouchard et al. 1985a., 4^{e} partie) et une décision de jumelage est prise. Cette décision fait appel à une table consignant divers critères et seuils (Bouchard et al. 1985a, pp. 214-215). Bien qu'il n'y ait aucune intervention manuelle au cours des Tris élémentaires, des tests de cohérence sont effectués après jumelage, qui permettent de déceler des situations d'homonymie et de faire de la récupération manuelle. Des résultats portant sur la reconstitution automatique des familles saguenayennes ont montré que plus de 95% des liens sont créés au terme de cette première étape (Bouchard et al. 1985a, pp. 233-250).

Suite aux Tris élémentaires, les fiches de famille sont en bonne partie constituées et riches de renseignements complémentaires aux données nominatives, comme le profil socioprofessionnel et l'itinéraire résidentiel des couples. On bénéficie ainsi d'une somme d'informations pour appuyer les décisions de jumelage qui seront effectuées lors de la deuxième étape qu'on appelle les Tris complémentaires. Ces tris créent d'autres sous-ensembles, sur la base de critères plus libéraux : il suffit que les mentions de couple aient un seul élément identique, plus les trois premières lettres d'un autre pour y être admises. Une nouvelle table de décision est utilisée et les instruments qui la composent s'appuient en partie sur les données non nominatives pour statuer sur les mentions comparées (Bouchard et al. 1985a, annexe 17). Au terme de cette opération, plus de 98% du fichier est constitué. Une phase de récupération manuelle ciblant surtout les mentions isolées termine le travail de composition des dossiers familiaux.

Après la construction des fiches de couple, il arrive que des fiches distinctes soient fusionnées, à cause du problème de l'homonymie. Des validations sont alors nécessaires pour s'assurer de la bonne composition du fichier. Des tests automatiques permettant de détecter des incohérences sont effectués sur les fiches de couple. Ils sont de plusieurs types : durées inadmissibles entre les événements, séquence chronologique hors norme, déclaration d'âge incompatible avec la participation à un événement, répétitions d'événement à caractère unique. Comme ces tests sont élaborés dans l'intention de détecter un maximum de fiches erronées, ils pèchent parfois par excès de prudence. Par conséquent, toutes les fiches pointées par les tests sont soumises à l'attention d'un préposé au jumelage afin de s'assurer que seuls les cas erronés soient retouchés.

Nouvelle approche

À l'aube d'importants travaux de développement qui ont été amorcés à la fin des années 1990, des modifications ont été apportées au système de reconstitution des familles SOREP (dorénavant système BALSAC). Les principales raisons et motivations se résument en quatre points : l'expérience acquise dans la construction du fichier, la décision de jumeler à partir des actes de mariages seulement, les impératifs de la recherche et les développements technologiques.

Les travaux méthodologiques qui ont précédé la construction du fichier ont montré que plus la somme d'informations est importante dans un dossier, plus la possibilité d'en assurer la validité est forte. Il est donc avantageux de jumeler sur la base des trois types d'actes, pour mieux appuyer les décisions de jumelage plus à risque. Malgré cela, les travaux de validation qui ont suivi la construction du fichier ont montré que les méthodes de reconstitution avaient leurs limites, particulièrement dans les dossiers à nombre de mentions de couple peu élevées où les tests de cohérence avaient peu de prise. Suite à la décision d'étendre les travaux à l'ensemble du territoire québécois mais uniquement à partir des actes de mariage, il a donc fallu adapter le système de saisie et de jumelage.

Par ailleurs, la demande croissante de la recherche, surtout en matière de construction généalogique, a aussi nécessité des changements au processus de jumelage. La construction de corpus en lots délimités régionalement suppose qu'on doive d'abord effectuer toutes les opérations de saisie avant d'entamer les travaux de jumelage. La démarche est lourde et les données jumelées ne sont pas disponibles pour la recherche tant que le travail n'est pas complété. De plus, les reconstitutions généalogiques automatisées ne peuvent être réalisées que sur des

portions du territoire québécois et doivent être complétées en grande partie manuellement. Pour remédier à ce problème, un fichier entièrement dédié aux travaux généalogiques (RETRO) a été construit afin de répondre aux besoins pressants de la recherche (Jomphe et Casgrain 2000). Cependant, la démarche utilisée pour la construction du fichier BALSAC a aussi été modifiée.

Enfin, lorsque le système de reconstitution automatique a été structuré (au milieu des années 1980), les équipements informatiques et les outils de développement étaient beaucoup moins performants et surtout moins souples qu'aujourd'hui, ce qui a amené à construire des systèmes indépendants les uns des autres. La puissance et la capacité de stockage croissantes des ordinateurs ont ouvert de nouveaux horizons et ont permis de refondre les divers systèmes pour faire place à un modèle plus convivial.

La saisie interactive

Une nouvelle configuration a été mise en place, de type client-serveur. Un parc de postes de travail est relié au serveur principal de données et à un serveur de gestion. Le module de saisie de données fut le premier élément du système à être modifié. Les changements apportés se résument en cinq points :

- Les données sont intégrées immédiatement dans le fichier BALSAC
- Les dictionnaires sont incrémentés à l'entrée de nouvelles valeurs
- Les éléments nominatifs sont soumis au programme de phonétisation
- L'acte intégré est fractionné en mentions de couple
- Une première phase de jumelage est opérée

Ces changements ont pour but de structurer les données dès leur intégration dans le modèle relationnel BALSAC (Casgrain et al. 1991) et d'amorcer les travaux de jumelage en réalisant la seule portion entièrement automatisée du système. Ce jumelage intervient dès le moment où l'acte est saisi, après la phonétisation des éléments nominatifs. Les mentions comprises dans le nouvel acte saisi sont automatiquement jumelées à des mentions déjà existantes dans la base de données BALSAC, si tous les éléments nominatifs sont identiques deux à deux (jumelage 4x4). Comme la saisie s'opère selon une démarche chronologique, les fiches de couple s'ouvrent par la mention de mariage des parents (sauf si le couple s'est marié hors Québec), appelée aussi mention fondatrice. Ensuite, viendront se greffer les autres mentions se rapportant au couple (mariages d'enfants et remariage de l'un ou l'autre des conjoints). Seules les mentions remplissant le critère d'identité parfaite sont acceptées, s'il n'y a qu'un seul candidat possible. Sinon la mention à jumeler est relayée à l'étape suivante appelée le « jumelage assisté par ordinateur ».

Le jumelage assisté par ordinateur

Pour comparer les mentions candidates à un jumelage, la mécanique de tris expliquée précédemment s'avère encore un instrument performant, surtout quand vient le temps de traiter des mentions touchées par les variations nominatives. Cependant, dès l'étape des six Tris élémentaires traitant les mentions affectées par des variations nominatives que le code fonem n'a pu normaliser, le préposé au jumelage peut prendre une décision à l'aide du procédé de jumelage assisté par ordinateur. La date de la mention la plus ancienne contenue dans une fiche de couple est le repère chronologique qui définit l'ordre dans lequel les fiches seront soumises au préposé.

À ce stade du jumelage, près des deux tiers des fiches s'ouvrent par la mention fondatrice (soit le mariage du couple parent). Bien que ces fiches ne soient pas nécessairement complètes, toutes les mentions susceptibles de s'y greffer font partie d'autres fiches qui ne contiennent pas la mention fondatrice et dont la date d'ouverture est forcément postérieure à la leur. Sont donc soumises à l'attention des préposés uniquement les fiches qui ne s'ouvrent pas par la mention fondatrice. Le but est d'arrimer ces fiches de couple à des fiches contenant la mention fondatrice ou de les documenter quand le mariage des parents a été célébré hors du Québec ou est inexistant.

L'environnement informatique a été conçu pour favoriser le traitement rapide des situations de jumelage soumises aux préposés. Plusieurs outils sont disponibles, dont des interfaces facilitant la configuration des données et leur consultation sous diverses formes. Des passerelles permettent de naviguer entre les fenêtres de représentation des données. Le préposé peut aussi accéder à des sources de données sur Internet et à des corpus formatés et intégrés à l'environnement BALSAC.

Le travail du préposé

Le traitement d'un cas s'amorce par une recherche de candidats potentiels pour la mention fondatrice. L'appel des candidats se fait selon les principes et conditions des six derniers Tris élémentaires. Le programme génère une ou des fiches candidates à un jumelage si les critères de tris et d'appariements sont satisfaits (au moins deux éléments nominatifs sur quatre sont identiques et les deux autres répondent à des critères de ressemblance déterminés à l'aide de programmes mesurant la similitude entre les données nominatives). Pour les fiches comprenant plus d'une mention, la comparaison se fera à partir d'une mention clé qui est un résumé des mentions nominatives de la fiche tenant compte de la diversité des éléments qui la composent. Le but est de présenter au préposé toutes les situations possibles et plausibles de jumelage, puisqu'il est le seul à ce stade à pouvoir prendre les décisions.

S'il y a des candidats, le préposé valide chacune des propositions à l'aide d'un programme qui jumelle virtuellement les fiches. La création de fiches virtuelles permet de placer les mentions à jumeler dans un ordre séquentiel, favorisant ainsi l'examen de la fiche. La fiche créée active un programme de détection d'incohérences qui signale les risques potentiels de procéder à un jumelage. Ce programme détecte des durées d'observation improbables (par exemple une durée de 75 ans et plus entre la première et la dernière mention d'une fiche), un nombre de mentions jugé suspect (plus de 15 sur une même fiche), des contradictions entre l'âge déclaré au mariage et l'écart entre le mariage d'un enfant et celui des parents, et la répétition d'un prénom d'enfant dans la même fiche. Comme les programmes de cohérence sont limités, le préposé dispose d'une série d'instructions avant d'accepter qu'une fusion virtuelle se concrétise en fusion réelle. Une attention particulière est portée aux itinéraires résidentiels et professionnels irréguliers ainsi qu'aux éléments nominatifs moins discriminants qui peuvent engendrer des jumelages de couples homonymes. Le but ultime est de greffer la mention fondatrice à la fiche traitée. Cependant des mentions « frères et sœurs » peuvent être jumelées si elles font partie des fiches candidates même si elles sont datées pour être traitées ultérieurement. Il est en effet utile de fusionner le plus de mentions possible au dossier en traitement; cela permet d'ajouter un complément d'informations qui facilite la prise de décision lors de l'appariement de la mention fondatrice.

Si la mention fondatrice ne fait pas partie des candidates, le préposé a accès à un annuaire électronique indexé, catalogué par mention clé et appelé « Têtes de famille ». Cet annuaire peut être interrogé selon sept ordres alphabétiques de tris; pour chaque tri, la mention clé du dossier en traitement est accolée à toutes les mentions clés triées sur les mêmes éléments nominatifs. Ainsi, des mentions clés qui n'ont pu être appelées comme candidates par le programme de Tris élémentaires parce que ne répondant pas aux règles de tris ou aux critères d'appariement, sont ici rapprochées des fiches en traitement. Chaque mention clé est associée à une date d'entrée en observation, à un lieu de mariage si disponible et à un indicateur de la présence de la mention fondatrice dans la fiche. L'usage des Têtes de famille à des fins de jumelage permet d'identifier des fiches candidates affectées par des variations nominatives sévères allant jusqu'à la substitution (changement complet de nom ou de prénom), fiches qui étaient dans l'ancien système appelées par les Tris complémentaires.

D'autres sources documentaires sont disponibles, dont un accès privilégié à des données de recensements canadiens qui permet de faire des recoupements et de suivre une nouvelle piste de recherche. Les préposés peuvent aussi consulter dans la salle de documentation du Projet BALSAC un large éventail de recueils généalogiques couvrant la plupart des régions du Québec, toutes les sources dispensées par le fonds généalogique Drouin et quelques sites Internet versés dans la généalogie (Rootsweb, BMS 2000, Ancestry, Mes Aieux). Ces sources additionnelles servent pour appuyer les décisions ou orienter la recherche de candidats.

En toute circonstance, le préposé peut défaire des liens erronés qui auraient été créés lors d'une phase de jumelage précédente.

Les statuts de jumelage

Si la mention fondatrice est trouvée et intégrée à la fiche, le cas en traitement est considéré comme réglé. Dans le cas contraire, trois cas de figure peuvent se présenter :

- Le mariage des parents a eu lieu à l'extérieur du Québec
- Le mariage des parents a eu lieu au Québec mais ne fait pas partie de la copie des registres de l'état civil utilisée à BALSAC
- L'information contenue dans l'acte est incomplète et celui-ci ne peut être utilisé à des fins de jumelage

Le préposé doit alors marquer sa recherche en attribuant un statut de jumelage à la fiche de couple qui a fait l'objet d'un traitement. Les deux principaux statuts accordés sont : réglé sans mariage (jumelage impossible) et sous enquête. Le statut « réglé sans mariage » renvoie à une situation où la recherche du préposé a fourni des informations amenant à conclure que le mariage des parents a été célébré à l'extérieur du Québec. Ces informations peuvent provenir des annotations d'origine ou de commentaires inscrits dans les actes d'état civil, ou d'autres sources comme les recensements, les sites Internet, etc. Les informations concernant le lieu de mariage du couple sont codées et des renseignements supplémentaires sur les parents ou la famille peuvent être ajoutés. Un dossier est catégorisé « sous enquête » lorsque le préposé n'a pu trouver la mention fondatrice ni aucune indication valable lui permettant de conclure à une situation de mariage hors Québec. Les cas inscrits dans cette catégorie sont documentés en prévision de la reprise du dossier par un préposé plus expérimenté, un superviseur ou un consultant externe. Un

document de BALSAC fournit un exemple de cas complexe résolu par un préposé (Bourque et Simard 2004); y sont exposés la démarche utilisée et l'ensemble des pièces justificatives.

Le temps de traitement

Un préposé traite en moyenne 15 cas dans une heure. Cependant, il existe un noyau dur qui demande une investigation plus appuyée. Bien que les documents de la salle d'archives et de consultation ainsi que les autres outils disponibles permettent de résoudre la plupart des cas, une documentation plus étendue apporterait un nouvel éclairage à des situations de jumelages très complexes. Faute de pouvoir offrir tous les outils nécessaires et pour éviter que le préposé se perde dans des enquêtes mal ciblées, des paramètres délimitant l'aire de recherche ont été établis. Ainsi, le temps maximum accordé à un traitement de première ligne a été fixé à 30 minutes. Au-delà de cette durée, l'équipe de supervision pourra convenir de l'utilité de poursuivre l'enquête.

Résultats du jumelage (période 1800-1940)

Quelque 1,25 million d'actes de mariage de la période 1800-1940 ont été intégrés et jumelés au fichier BALSAC. De plus, l'arrimage du fichier avec les unions du Registre de la population du Québec ancien (Desjardins 1998), qui couvre la population du Québec des origines jusqu'à 1800, est maintenant terminé.

Sur les 2 126 279 mentions de couple soumises au jumelage, 61,5 % ont été traitées à l'étape du jumelage automatique. C'est un résultat très satisfaisant compte tenu de la sévérité des règles et des critères régissant cette opération. Un peu moins de 38,5 % des mentions (817 653) ont donc été soumises au jumelage assisté par ordinateur, ce qui donne une idée de l'importance du phénomène des variations orthographiques et justifie pleinement les efforts consentis pour les traiter et le pari d'investir davantage dans les opérations manuelles.

Représentativité des données du fichier

Environ 90% des mentions soumises au jumelage ont un lien généalogique confirmé sur le territoire québécois, ce qui laisse au maximum 10% de mentions orphelines (sans mention fondatrice) puisqu'un certain nombre (2%) est en attente d'un complément d'enquête. En outre, la plupart des mentions dites « orphelines » sont documentées, au sens où des informations permettant de les relier à leur lieu et famille d'origine sont inscrites. Sur le plan généalogique, le fichier est donc d'une grande richesse, puisque l'on peut maintenant tracer les lignées d'une grande partie de la population québécoise (environ 87%) pour les générations comprises entre 1800 et 1940. Le 13% manquant renvoie à des sous-populations mal représentées dans le fichier parce que provenant de confessions religieuses autres que catholique. Les registres de la majorité de ces confessions, particulièrement ceux du XIXème siècle, sont d'une tenue inadéquate et fortement carencés sur le plan nominatif. Pour ces raisons l'intégration de ces sous-populations, même si elle ne pourra qu'être partielle, a été reportée à plus tard.

Validation et taux d'erreurs

Puisque le système de jumelage est fondé sur la similitude observée entre noms et prénoms, il existe un risque d'effectuer des jumelages de couples distincts (surjumelage dû à l'homonymie) ou de ne pas assembler des paires de mentions renvoyant au même couple (sous-jumelage dû aux variations et substitutions nominatives). Les problèmes de sous-jumelage sont négligeables puisque toutes les mentions de couple doivent se retrouver sur une fiche avec la mention fondatrice (soit le mariage du couple parent), ou sans la mention fondatrice mais identifiée à un couple immigrant. En fait, les cas de sous-jumelage se cachent dans les fiches contenant du surjumelage, puisque celles-ci contiennent des mentions appartenant à une autre famille, ou font partie des cas qui ont été soumis à une enquête. Les travaux ont donc été orientés vers la détection des familles susceptibles d'être touchées par le surjumelage. Plusieurs séries de validations ont été effectuées : validation de familles à risque pour calibrer les tests et requêtes de cohérence, validation du travail des préposés au jumelage et estimation du nombre d'erreurs sur la base d'échantillons aléatoires.

Au total, 20 000 fiches de familles susceptibles de contenir des mauvais liens ont été validées par des superviseurs d'expérience, qui avaient le loisir de procéder à des enquêtes approfondies. Près de 1 400 erreurs de jumelage ont été détectées et corrigées. Ce nombre peut paraître relativement élevé, mais il faut rappeler que les validations touchent les familles les plus à risque de contenir des mauvais liens (1,5% de l'ensemble des familles de la base de données). Réalisé au cours des travaux de jumelage sur des familles déjà composées, l'examen de ces corpus a donc permis de détecter des surjumelages, mais surtout d'ajuster les paramètres des tests et d'instaurer de nouvelles mesures de contrôle. Les tests de cohérence ont aussi été resserrés.

Le travail des préposés a été scruté pour s'assurer que les consignes de jumelage et les instructions de notification étaient respectées et que de faux liens n'étaient pas créés. Des procédures de vérifications mensuelles ont été instituées visant à vérifier l'ensemble du travail des préposés pendant les deux premiers mois de leur formation et environ 3% des cas traités par la suite. Très peu d'erreurs ont été détectées au cours de ces validations (0,008%) et la plupart d'entre elles ont été commises au stade de formation.

Enfin, des échantillons aléatoires ont été produits afin de mesurer le taux d'erreur à diverses étapes du jumelage et selon la marche de développement du fichier. Il s'agissait surtout de vérifier si tous les efforts de validation donnaient des résultats tangibles. Plus de 60 000 fiches de familles ont été validées et 47 erreurs (0,08%) ont été décelées. Cependant, au terme des travaux de jumelage sur la période 1800-1940, le taux d'erreur subsistant n'est plus que de 0,05% (une erreur sur 2000 fiches). D'autres validations sont en cours, dont des comparaisons périodiques avec le fichier généalogique RETRO.

Les contrôles et opérations de validation, bien que très efficaces, ont leurs limites. Malgré l'excellente qualité des registres de l'état civil québécois, de fausses informations, difficilement quantifiables et identifiables, peuvent exister. On peut penser par exemple aux adoptions cachées ou aux fausses déclarations parentales. Avec l'aide de la génétique, il est possible d'estimer l'ampleur de ces phénomènes et d'évaluer ainsi les erreurs de jumelage qui pourraient subsister. Par exemple, des résultats d'analyse d'ADN mitochondrial et du chromosome Y pourraient être

utilisés afin de valider les lignées généalogiques maternelles et paternelles des individus. Si des individus partagent les mêmes haplotypes transmis par les mères (via l'ADN mitochondrial) ou par les pères (via le chromosome Y), il est presque certain qu'ils partagent aussi l'ancêtre à l'origine de ces lignées. Une estimation préliminaire effectuée à partir de données sur des lignées paternelles saguenayennes a montré un taux d'erreur probable de moins de 1% dans l'ensemble des liens généalogiques identifiés (Heyer et al. 1997). D'autres études plus approfondies devront toutefois être effectuées afin d'avoir une idée plus précise de ce phénomène.

Perspectives

Grâce au développement soutenu du fichier BALSAC, les possibilités d'utilisation des données du fichier dans les champs de la génétique et des sciences sociales ont augmenté de façon significative. Le traitement des requêtes de recherche est beaucoup plus efficace, surtout en ce qui a trait aux reconstitutions généalogiques puisque la partie automatisée des constructions est de plus en plus performante.

Depuis avril 2007, les travaux de développement portent principalement sur la saisie des données de la période 1940-1964, effectuée à partir d'images d'actes numérisés grâce à une entente intervenue entre le Projet BALSAC et le Directeur de l'état civil du Québec, seul dépositaire des actes postérieurs à 1940. Environ 900 000 nouveaux actes de mariage seront intégrés au fichier d'ici mars 2011.

D'autre part, on a constaté lors du traitement des données de la période 1800-1940, qu'une large part des décisions de jumelage prises par les préposés s'est concrétisée rapidement et sans recours à d'autres sources que la boîte de candidats fournie par le programme de jumelage. Aussi, dans le but de rendre les données disponibles plus rapidement à la recherche et de réduire les coûts de traitement, nous avons transféré les cas de jumelage les plus faciles à traiter au stade du jumelage assisté par ordinateur à l'étape de la saisie de données. Les premières constatations concernant cette nouvelle approche nous montrent que les résultats sont très prometteurs. En effet, on envisage qu'à terme, près de 90% des jumelages potentiels seront réalisés avec sureté dès la saisie de donnée complétée, sans que cela ne ralentissent de manière notable le rendement de nos préposés.

Enfin, les bouleversements importants qui touchent les pratiques matrimoniales au Québec depuis une trentaine d'années auront une incidence directe sur les travaux de développement du fichier BALSAC. On assiste en effet à une augmentation croissante des unions de fait et des familles recomposées. Phénomène marginal avant les années 1970, les unions de fait concernent 35% des couples québécois recensés en 2006 (Girard 2008). Comme la construction du fichier de population repose sur les actes de mariage, l'existence des unions de fait laisse présager d'importants ajustements, puisque ces unions ne donnent lieu à aucun enregistrement (l'absence d'enregistrement empêche de faire le lien entre un individu vivant en situation d'union de fait et ses parents). Pour le moment, une seule solution est envisageable : l'accès aux informations contenues dans les enregistrements de naissance, disponibles à la Direction de l'état civil. Ces informations aideraient à trouver les chaînons manquants dans les segments généalogiques. Les renseignements contenus dans la déclaration de naissance d'un enfant issu d'un couple vivant en union de fait (nom, prénom et date de naissance du père et de la mère) permettent de jumeler la

naissance de l'enfant à celles des parents et ainsi obtenir les noms des grands-parents. Une union virtuelle remplacerait l'acte de mariage pour le couple vivant en union de fait. Il sera important d'obtenir les enregistrements de naissance à partir de 1940, afin d'être en mesure de repérer la naissance de la plupart des parents concernés. On observe aussi, depuis les années 1970, une forte augmentation du nombre de séparations et de divorces et, conséquemment, un nombre croissant de familles recomposées. Bien que cela ne corrompe pas, du moins théoriquement, les chaînes généalogiques, une certaine confusion est possible dans l'identification des parents biologiques. On devra à tout le moins en évaluer le risque.

Références bibliographiques

Bouchard G (2008). *Projet BALSAC – Rapport annuel 2007-2008*. Projet BALSAC, Chicoutimi.

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

Bouchard G, Roy R, Casgrain B (1985a). *Reconstitution automatique des familles. Le système SOREP*. Dossier no. 2, 2 vol., Université du Québec à Chicoutimi, 745 p.

Bouchard G, Roy R, Otis Y (1985b). Registre de population et substitutions nominatives. *Population*, 3:473-490.

Bourque M, Bouchard M (2003). *Exhaustivité de l'enregistrement des actes des non-catholiques et évaluation de leur contenu nominatif à des fins de jumelage (Québec, 19^{e} siècle).* Document I-C-218, Projet BALSAC, Chicoutimi, 6 p.

Bourque M, Simard P (2004). *Exemple d'un cas de jumelage au Projet BALSAC : démarche et justification.* Document I-C-219, Projet BALSAC, Chicoutimi, 24 p.

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

Casgrain B, Hubert M, Bouchard G, Roy R (1991). Structure de gestion et d'exploitation du fichier-réseau BALSAC. Dans G Bouchard, M DeBraekeleer et al. : *Histoire d'un génome. Population et génétique dans l'est du Québec*. Presses de l'Université du Québec, pp. 47-71.

Desjardins, B (1998). Le Registre de la population du Québec ancien. *Annales de démographie historique*, 1998-2:215-226.

Girard C (2008). *Le bilan démographique du Québec*. Institut de la statistique du Québec, 79 p.

Heyer É, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human Molecular Genetics*, 6-5:799-803.

Jomphe M, Casgrain B (1997). *Base de données généalogiques RETRO: structure des données*. Document III-C-97, IREP, Chicoutimi.

Roy R, Charbonneau H (1976). Le contenu des registres paroissiaux canadiens du XVIIe siècle. *Revue d'histoire de l'Amérique française*, 30-1:85-97.