UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Développement d'un module d'assistance au jumelage dans le cadre de la réingéniérie du logiciel de gestion de registres de population : Analypop

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

ETIENNE MORENCY-BACHAND

MAI 2007

You don't have to be a scientist [...]
in order to understand enough science
to overtake your imagined need
and fill that fancied gap.
Science needs to be released from the lab into the culture.

"The Richard Dimbleby Lecture : Science, Delusion and the Appetite for Wonder"
Richard Dawkins

REMERCIEMENTS

Je voudrais tout d'abord remercier ma directrice Anne Bergeron pour ses commentaires pertinents ainsi que pour son support financier.

De plus, un sincère merci à ma co-directrice Francine M. Mayer ainsi qu'à Mireille Boisvert qui ont été présentes tout au long du processus d'écriture de ce mémoire, toujours disponibles pour discuter des concepts nécessaires à l'élaboration de la nouvelle version du logiciel.

Un très gros merci à Jacques, Danielle, Lise, Judith et Marie-Eve!

TABLE DES MATIÈRES

TABLE DES FIGURES

RÉSUMÉ

Le présent document présente d'abord un bref historique ainsi que les concepts de base de la génétique de population pour ensuite présenter les concepts utilisés pour développer le module d'assistance au jumelage du programme Analypop, un programme de gestion et d'analyse de registre de population. Par la suite, nous verrons comment la conception choisie permet aux chercheurs d'effectuer la consultation ainsi que d'enregistrer certains changements dans le registre de population. Finalement, nous verrons deux fonctionnalités qui permettent d'exporter les données pour qu'elles puissent être traitées par d'autres programmes.

Mots-clés : Logiciel, Démographie, Registre de Population, Analypop.

INTRODUCTION

Au mois de juin 2004, dans le cadre d'un stage en bioinformatique au sein du laboratoire de recherche en anthropologie biologique de Francine M. Mayer, une équipe a été formée afin de planifier et réaliser la réingéniérie du logiciel Analypop. Ce logiciel gère des informations sur les individus d'une population ainsi que les évènements qui les unissent. Les informations concernant les individus d'une population sont contenues dans ce qu'on appelle le *registre de population*. Le registre de population est composé des divers actes de l'état civil tels que les actes de naissance, baptême, mariage, décès et sépulture que l'on jumelle. Le *jumelage* est l'action par laquelle on confirme qu'il existe un lien de parenté entre deux individus présents dans le registre de population. On peut déclarer que deux individus ont un lien de parenté ou bien déduire que deux individus inscrits dans le registre de population sont en fait le même individu. Donc, l'élaboration d'un registre de population nécessite de dépouiller l'ensemble des actes d'état civil et de comprendre les liens qui unissent les gens qui y sont mentionnés. Lorsque toutes les mentions concernant les individus dans les différents actes ont été rapprochées et jumelées, le registre de population est considéré complet. En réalité, il n'est jamais complet puisqu'il y a toujours des jumelages à faire ou défaire étant donné qu'on a jamais toute l'information nécessaire, ni tous les individus de la population. Finalement, le registre de population permet de présenter les évènements qui lient les individus entre eux. Ces évènements sont présentés sous la forme de *fiches de familles*. Celles-ci regroupent les informations relatives à un individu et ses parents, son(sa) conjoint(e) et ses parents, ainsi que les enfants de ce couple.

Analypop est donc le logiciel utilisé pour stocker les données du registre de population et gérer les liens de parenté. Il permet de faire des jumelages et il donne la possibilité d'effectuer plusieurs calculs propres au domaine de la génétique de population.

Étant donné que la version d'Analypop utilisée en 2004 nécessitait de nouveaux développements et que l'analyste concepteur d'Analypop avait quitté l'équipe de Francine M. Mayer, la décision a été prise de développer une nouvelle version du logiciel plutôt que d'ajouter des fonctionnalités à l'ancienne version. La nouvelle version devait rendre les chercheurs plus autonomes dans leur utilisation du logiciel ainsi que permettre des développements subséquents. Le but de ce mémoire est de décrire la réingéniérie de ce logiciel ainsi que que présenter le développement du module d'assistance au jumelage.

Afin de présenter le travail effectué dans le cadre de ce mémoire, nous présenterons les différents aspects du logiciel Analypop selon l'ordre logique d'utilisation de ce logiciel. Tout d'abord, en vue de bien comprendre le domaine auquel doit s'appliquer Analypop, nous présenterons au Chapitre 1 l'historique de la génétique suivi de la description de certains concepts de base du domaine de la génétique des populations. Ensuite, étant donné que l'utilisation du logiciel nécessite tout d'abord d'avoir un support logique pour stocker les données du registre de population, le Chapitre 2 sera consacré à présenter le modèle de la base de données qui a été implantée, pour ensuite présenter l'importation de données. Une fois que la base de données contient les données importées du registre de population, nous pouvons nous consacrer à effectuer les opérations de jumelage. Au Chapitre 3, nous aborderons donc la gestion des jumelages et de leur historique. Finalement, puisque le logiciel permet d'exporter certaines données pour être visionnées à l'aide d'autres logiciels, le Chapitre 4 présentera les deux types de fichiers de sorties présentement implantés dans Analypop.

Les aspects originaux de ce travail résident d'abord dans le fait que nous avons développé une infrastructure qui permet l'ajout de nouveaux types de données pouvant se rattacher à un registre de population, quels qu'ils soient (voir Chapitre 2). De plus, l'interface graphique développée afin de consulter le registre de population et de faire la gestion des jumelages fournit un environnement beaucoup plus approprié pour les utilisateurs que l'ancienne interface texte. Enfin, la création d'un module de positionnement de données de type géographique (latitude-longitude) est un ajout important qui permettra aux chercheurs de développer de nouvelles voies de recherche (voir Chapitre 4). D'ailleurs,

ce module fait déjà partie d'une initiative de recherche visant à identifier la transmission des terres de 6 familles fondatrices de La Patrie, dans les Cantons de l'Est, afin de détecter les différentes stratégies utilisées pour l'établissement des enfants.

CHAPITRE I

HISTORIQUE ET CONCEPTS DE BASE DE LA GÉNÉTIQUE DES POPULATIONS

Ce chapitre se divise en trois sections. La première présente un bref historique de la génétique afin de mieux apprécier les progrès qui ont été faits dans ce domaine grâce à la contribution de l'informatique. Jugeant que le contexte socio-scientifique qui a mené à l'élaboration de la discipline de la génétique des populations est intéressant, une importance particulière y a été accordée. La deuxième section, quant à elle, présente les concepts de base de la génétique des populations, concepts qui devraient permettre de mieux comprendre les raisons pour lesquelles un logiciel tel qu'Analypop peut venir en aide aux chercheurs. Par la suite, un bref aperçu de certaines études ayant eu recours à Analypop sera présenté dans la dernière section de ce chapitre.

1.1 Historique de la génétique

1.1.1 Les premiers croisements génétiques

Depuis des siècles, l'être humain tente de comprendre et de maîtriser son environnement. La sédentarisation l'a amené à vouloir mieux contrôler le monde qui l'entoure. Il y a environ 10 000 ans, les humains ont remarqué que certains traits caractéristiques des parents peuvent être transmis à leur descendance. Déjà à cette époque, avec la pratique de l'agriculture, les humains effectuèrent des croisements de plantes afin d'améliorer la production agricole. Entre $-$ 9 000 et $-$ 7 000 ans, les humains domestiquèrent

plusieurs espèces animales telles les moutons, les vaches, les chèvres, les cochons et plus récemment, les chevaux et les lamas. Que ce soit pour améliorer la quantité ou la qualité des récoltes ou pour augmenter leur potentiel alimentaire et leur robustesse ou pour domestiquer les animaux, l'être humain a compris qu'il pouvait sélectionner certains traits en effectuant des croisements entre des représentants de l'espèce ayant les traits recherchés. Déjà, il avait la connaissance nécessaire pour effectuer une certaine sélection que l'on jugeait favorable en fonction des objectifs poursuivis, sans toutefois avoir les connaissances des mécanismes sous-jacents qui permettaient de faire cette sélection. Ce sont les Grecs qui formulèrent les premières hypothèses, en majeure partie basées sur des superstitions, afin d'expliquer les mécanismes reliés à l'hérédité. Toutefois, il faut se référer aux scientifiques de l'époque moderne pré-mendélienne tels que Lazzaro Spallanzani, Anton van Leeuwenhoek ou Georges Buffon pour relever les premières hypothèses scientifiques sur la transmission des traits.

1.1.2 La génétique à l'époque pré-mendelienne

Ces scientifiques se sont posé la question à savoir comment un enfant pouvait avoir des caractéristiques semblables à celles de son père ou de sa mère. Par exemple, Lazzaro Spallanzani, biologiste italien ayant effectué la première insémination artificielle (sur une chienne en 1785), faisait partie de l'école des *ovistes*, qui croyaient que l'ovule de la femme comportait tout le nécessaire pour former un être humain et que les spermatozoïdes servaient à déclencher ce processus. À l'opposé, l'école des *spermatistes*, en grande partie appuyée par les travaux de Anton van Leeuwenhoek qui, vers 1671, grâce à sa propre invention de lentilles de microscopes capables de grossir une image plus que tous les autres microscopes de l'époque, était de l'avis que les spermatozoïdes étaient des êtres humains complets, extrêmement petits, prêts à grandir à l'intérieur de l'ovule. C'est à Georges Buffon, auteur de l'encyclopédie "Histoire naturelle" en 36 volumes, complétée en 1804, que l'on attribue généralement l'idée selon laquelle l'ovule et les spermatozoïdes sont d'égale importance lors de la conception. Il était de l'avis que les liquides séminaux de l'homme et de la femme contenaient des particules qui étaient

envoyées par toutes les parties du corps humain et qui s'agençaient de manière à créer le nouvel individu.

1.1.3 Contexte socio-scientifique

À cette époque, une conception de l'histoire naturelle très répandue était la théorie *catastrophiste* qui stipule que les différentes époques géologiques terrestres ont toutes étés marquées par une extinction massive des espèces vivantes avant que d'autres espèces fussent créées. Ceci impliquait que les différentes espèces présentes à une époque géologique donnée avaient été conçues au début de cette ère géologique et resteraient ainsi jusqu'à la prochaine catastrophe. Charles Lyell, avec la publication des trois volumes des principes de géologies de 1830 à 1833, introduisit une nouvelle manière de voir les transformations géologiques (12) : il croyait que les mécanismes qui oeuvrent aujourd'hui ont toujours été présents et que c'est leur lente et constante addition qui explique ces transformations.

Dans le domaine de l'économie, Thomas Malthus, économiste anglais renommé, publia en 1798 un essai qui soutenait l'idée que les populations humaines s'accroissent plus rapidement que les ressources alimentaires nécessaires à leur survie (21). Ses idées venaient contredire l'impression populaire qui croyait que la croissance de la population devait nécessairement mener vers le progrès économique. En effet, la population anglaise était à cette époque en pleine révolution industrielle (la révolution industrielle anglaise débuta vers les années 1770) et était en pleine croissance économique. D'ailleurs, ses écrits ont été à la base des premières études démographiques encadrées car, en 1801, le gouvernement anglais établit le premier recensement par crainte que la production agricole soit insuffisante à cause d'une population croissante.

1.1.4 La théorie de Darwin : la sélection naturelle

C'est dans ce cadre social et scientifique que Charles Darwin s'engagea à bord du HMS Beagle durant l'année 1831, à l'âge de 22 ans, à titre de naturaliste et assistant per-

sonnel du capitaine Robert FitzRoy. Ce voyage autour du monde lui permit d'observer plusieurs phénomènes naturels ainsi que de faire la collecte de plusieurs spécimens. Ses observations venaient appuyer la théorie de Charles Lyell pour ce qui est des phénomènes géologiques transformant la surface de la planète mais ne venaient pas appuyer l'idée selon laquelle les espèces avaient étés créés immuables. En effet, il avait pu faire l'étude de fossiles partageant plusieurs traits avec des espèces contemporaines en plus de remarquer que certaines espèces adoptaient des comportements différents selon le milieu où elles vivaient. Il est intéressant de relever qu'Alfred Wallace, lui aussi naturaliste et lui aussi ayant fait un voyage d'observation et de collecte de spécimens de plusieurs années, était arrivé à des conclusions similaires et qu'il avait décidé d'envoyer son papier scientifique à Darwin en 1858. Darwin choisit de partager l'annonce de sa théorie avec Wallace en faisant présenter leurs deux écrits au *Linnean Society* à Londres en juin 1858. L'année suivante, Darwin publia sa théorie où il expliquait sa vision de l'origine des espèces comme étant le résultat de la sélection naturelle ; il définissait les espèces comme le résultat de la lutte pour la survie dans un milieu où les ressources naturelles sont limitées, ce qui confère un avantage reproductif aux individus les plus aptes (2). De plus, il soutenait que toutes les espèces comportant des traits similaires de nos jours descendent d'ancêtres communs.

1.1.5 Les expériences de Mendel

La théorie de Darwin, quoique apportant une immense contribution à la connaissance de l'époque, demeurait basée sur des observations qualitatives. C'est en 1866, suite aux fameuses expériences faites par Gregor Mendel sur des plants de petits pois, que la communauté scientifique pu avoir accès à des preuves quantitatives démontrant les mécanismes liés à la transmission des caractères lors de la reproduction sexuée. Grâce à ses expériences sur l'hérédité de sept caractères chez les petits pois, il a pu décrire le modèle de l'héritage génétique. Mendel fit l'hypothèse que chaque plante comporte deux *allèles* (un allèle est une variante d'un gène au sein d'une espèce) pour chacun des caractères de celle-ci ; un trait provenant du plant mâle et l'autre du plant femelle. À son

tour, chaque plant, lorsque croisé avec l'autre plant, ne pourra transmettre qu'un seul facteur pour chaque caractère qu'il possède et ce facteur est sélectionné au hasard. Les travaux effectués par Mendel lui permirent d'énoncer deux lois sur l'hérédité, qui sont encore valides aujourd'hui, quoique enrichie depuis. La première loi de Mendel, appelée *loi de ségrégation*, stipule que :

"Les deux membres d'une paire de gènes se disjoignent (ou ségrégent) lors de la formation des gamètes, de telle manière qu'une moitié des gamètes portent l'un des membres de la paire et la moitié restante de l'autre."(5)

Cette loi permet d'expliquer que, pour un caractère donné, c'est le hasard qui déterminera lequel des deux allèles l'enfant obtiendra de sa mère et lequel de son père. La deuxième loi de Mendel, appelée aussi *loi de la ségrégation indépendante* est la suivante :

"Pendant la formation des gamètes, la ségrégation des allèles d'un gène donné s'opère indépendamment de celle des allèles d'un autre gène."(5)

C'est cette loi qui explique que si l'on considère plusieurs gènes, les différents allèles n'apparaissent pas toujours ensembles. Une mère ayant les cheveux bruns et les yeux bleus peut transmettre ces allèles de façon indépendants et ainsi avoir un enfant aux cheveux bruns et yeux bruns.

1.1.6 Autres mécanismes de l'hérédité

Ce n'est qu'en 1900, soit 34 ans plus tard, que l'importance de ses observations fut révélée au monde scientifique par la "redécouverte" des mécanismes de l'hérédité, par Hugo de Vries, Carl Correns, Erich von Tschermak et quatres autres chercheurs, soit Thomas Morgan et ses étudiants : Alfred Sturtevant, Calvin Bridges et Hermann Muller. Malgré le fait que de Vries soit en accord avec l'explication de Mendel définissant des unités indépendantes responsable de transmettre l'information héréditaire, il croyait

que chaque unité, qu'il appelait "pangène", pouvait contrôler une grande quantité de traits; il a alors introduit la terme mutation pour expliquer le processus par lequel un changement dans une de ces unités pouvait mener à une nouvelle espèce. À l'époque, cette théorie permettait de fournir une alternative à la théorie de la sélection naturelle de Darwin dans laquelle les espèces n'étaient plus le résultat d'un long processus d'accumulation de petits changements mais plutôt le résultat d'un nombre restreint de mutations sur ces "pangènes". Dans les mêmes années, en 1902, Walter Sutton, un américain, et Theodor Boveri, un allemand, constatèrent que le comportement des caractères de Mendel était très semblable au comportement des chromosomes lors de la méiose : les chromosomes ségrégent vers les deux gamètes et chaque chromosome est indépendant l'un de l'autre. Cette ressemblance leur permit de conclure que les gènes étaient localisés sur les chromosomes et ils nommèrent cette théorie, la *théorie chromosomique de l'hérédité.*

1.1.7 Les découvertes de Thomas Morgan

Durant les premières décennies du 20^{e} siècle, un professeur de zoologie expérimentale, Thomas Morgan fit plusieurs expériences sur les mouches drosophiles (*Drosophila melanogaster*) qui permirent de confirmer les écrits de Mendel ainsi que la théorie chromosomique de l'hérédité tout en ajoutant quelques mécanismes tels que le *linkage* et le *crossing-over*. Le *linkage* est le terme utilisé pour décrire le fait que deux ou plusieurs gènes situés sur un même chromosome ont une plus grande probabilité d'être transmis ensembles que s'ils sont sur des chromosomes différents. Lorsqu'un chromosome n'est transmis qu'en partie, l'autre partie est complétée par le deuxième chromosome formant la paire, c'est ce qu'on appelle la *recombinaison.* On parle d'un *descendant recombiné* lorsqu'une recombinaison a affecté un ou plusieurs des gènes de cet individu. L'individu recombiné se retrouve donc avec un nouvel ensemble de gènes qui diffèrent de celui de chacun de ses parents. À partir de ce phénomène, Morgan, proposa sa théorie à l'effet que les gènes sont arrangés les uns à la suite des autres sur les chromosomes. En observant la fréquence à laquelle deux gènes étaient transmis ensembles, Morgan créa

la première carte de liaison chromosomique sur laquelle les gènes sont situés selon leur position relative les uns par rapport aux autres à partir de la *fréquence de recombinaison.*

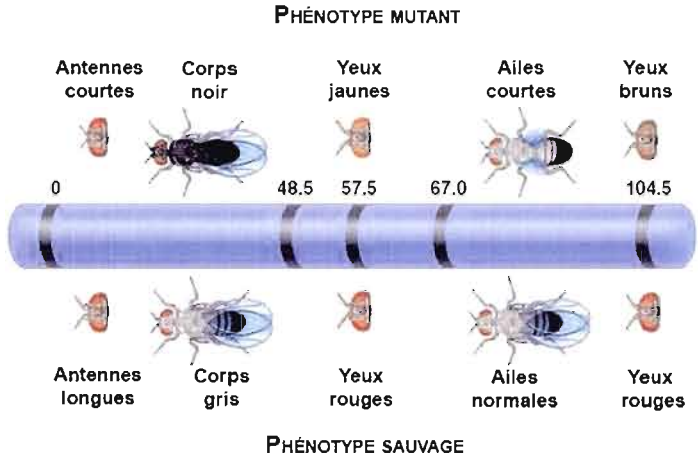

Fig. 1.1 Exemple de carte de liaison de gènes situés sur le chromosome 2 de *Drosophila melanogaster.* Source : http ://bio1151.nicerweb.com

Définition 1.1.1 La *fréquence de recombinaison* se calcule selon la formule suivante (le *centimorgan* est l'unité de mesure de la distance relative des gènes) :

$$\mathrm{F} = \frac{nb\ de\ descendants\ recombinés}{nb\ de\ descendants\ total} \times\ 100$$
$$1\ \%\ de\ recombinaison = 1\ cM\ (centimorgan)$$

La Figure 1.1 présente la carte de liaison chromosomique de la mouche drosophile. Les caractères représentés sur cette carte sont ordonnés selon leur fréquence de recombinaison. Ainsi, on peut déduire qu'il est plus fréquent d'observer chez la mouche une recombinaison entre les caractères antennes courtes et les yeux rouges qu'entre les caractères antennes courtes et corps gris puisque ces caractères sont plus éloignés l'un de l'autre sur le chromosome, donc la probabilité d'une recombinaison est plus grande.

Comme nous l'avons définit plus haut, un *allèle* est chacune des variantes d'un gène. Chez l'être humain, on possède deux lots de chromosomes, les gènes sont donc représentés

par deux allèles. Les caractères *récessifs* sont des caractères dont l'allèle doit être présent en deux copies pour s'exprimer, alors que les caractères *dominants* s'expriment dès qu'un des allèles est présent.

Le terme *homozygote* se dit d'un individu qui possède en double l'allèle d'un caractère donné par opposition au terme *hétérozygote* qui se dit d'un individu qui possèdent deux allèles différents pour un caractère donné.

Par exemple, considérons une population de 100 mouches *Drosophilia melanogaster*. Étudions cette population selon trois traits liés, c'est-à-dire situés sur le même chromosome, soit :

La couleur du corps : noir ou gris
La couleur des yeux : rouges ou jaunes
La longueur des ailes : courtes ou normales

Les caractères suivants sont dominants :

G pour corps gris
R pour yeux rouges
N pour ailes normales

Et les caractères suivants sont récessifs :

g pour corps noir
r pour yeux jaunes
n pour ailes courtes

On décrit le *génotype* d'un individu en faisant la liste des allèles qu'il possède. Par exemple : $GgRrnn$. Le *phénotype* d'un individu concerne son aspect physique. Par exemple, un individu $GgRrnn$ aurait le corps gris, car c'est un caractère dominant, les yeux rouges, qui est également dominant et les ailes courtes puisque le caractère récessif est présent en deux copies.

Suite au croisement de mâles hétérozygotes $GgRrNn$ (corps gris, yeux rouges et ailes normales) avec des femelles elles aussi hétérozygotes $GgRrNn$, dont nous savons que les

caractères dominants (GRN) sont sur le même chromosome et les caractères récessifs (grn) sont sur le même chromosome, nous nous attendons à avoir les génotypes suivants dans les descendants :

♂ \ ♀	GRN	grn
GRN	$GGRRNN$	$GgRrNn$
grn	$GgRrNn$	$ggrrnn$

Selon ces génotypes, on peut déduire que les phénotypes des mouches seront soit des mouches avec corps gris aux yeux rouges et ailes normales ($GGRRNN$ ou $GgRrNn$) soit des mouches avec corps noir aux yeux jaunes et ailes courtes ($ggrrnn$). Dans le cas où on observerait, par exemple, une mouche au corps noir aux yeux rouges mais avec des ailes courtes (G_R_nn), nous serions en présence d'une mouche mutante suite à une *recombinaison.* Cette opération consiste en l'échange de matériel génétique entre deux chromosomes similaires, portant des allèles possiblement différents. La Figure 1.2 est l'illustration que Morgan utilisa pour expliquer le phénomène qu'est la recombinaison (on y a ajouté les lettres correspondants aux caractères mentionnés çi-haut) :

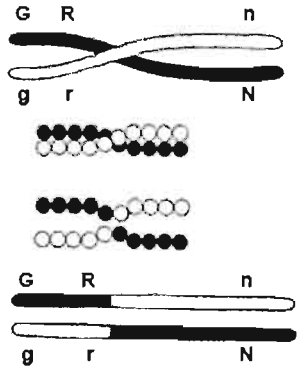

Fig. 1.2 La recombinaison telle que présentée par Thomas Morgan (22).

1.2 Concepts de bases de génétique des populations

1.2.1 Définition

La génétique des populations tente, à l'aide de modèles mathématiques, de comprendre les conséquences des principes de transmission de gènes sur l'état du pool génétique (l'ensemble des gènes présents dans une population) passé, présent et futur d'une population. La section suivante présentera les concepts mathématiques les plus couramment utilisés dans le domaine de la génétique des populations.

1.2.2 Loi de Hardy-Weinberg

En 1908, quelques années après la découverte du modèle de Mendel, un mathématicien anglais, Godfrey Hardy, et un biologiste allemand, Wilhelm Weinberg, ont proposé de manière indépendante une formule permettant de décrire une population à l'état d'équilibre. *L'état d'équilibre* se dit d'une population où la fréquence du caractère étudié reste stable au cours des générations et est déterminé par le seul jeu des lois de l'hérédité. La loi de Hardy-Weinberg est un outil important lorsqu'on désire étudier la distribution des fréquences d'un gène dans une population et identifier les mécanismes évolutifs qui modifient cet équilibre neutre. Tout gène peut avoir plusieurs allèles, qui déterminent souvent l'apparition de caractères héréditaires différents. Chez un individu possédant deux lots de chromosomes, appelé *diploïde*, les gènes sont représentés par deux allèles qui peuvent être différent ou pareils.

Loi 1.2.1 Loi de Hardy-Weinberg
Cette loi stipule que, pour une population à l'état d'équilibre, un caractère avec deux allèles, A représentant l'allèle dominant et a représentant l'allèle récessif, qui ont respectivement les fréquences p et q dans la population étudiée, la fréquence du génotype AA sera égale à p^2, celle du génotype Aa sera égale à $2pq$ et celle du génotype aa sera égale à q^2.

Puisque chaque individu de la population est soit homozygote (AA ou aa) ou hétérozygote (Aa), on peut affirmer que :

$$p^2 + 2pq + q^2 = 1$$

On peut illustrer cette formule à l'aide de *L'échiquier de Punnett*, du nom du généticien anglais Reginald Punnett qui inventa son utilisation. Il est utilisé pour calculer la probabilité que la descendance exprime un certain génotype lorsqu'on connaît la fréquence génotypique des parents. Dans le cas qui nous intéresse, on connaît la fréquence allélique dans la population pour les allèles a et A : respectivement p et q. L'échiquier nous permet de représenter tous les génotypes possibles lorsqu'on croise deux individus en considérant un croisement aléatoire.

♂ \ ♀	A $F = p$	a $F = q$
A $F = p$	AA $F = p^2$	Aa $F = pq$
a $F = q$	aA $F = qp$	aa $F = q^2$

On voit donc que la probabilité du phénotype AA est égale à p^2 ($F(AA) = p^2$) , que la probabilité du phénotype aa est égale à q^2 ($F(aa) = q^2$) ainsi que la probabilité du phénotype Aa (équivalent à aA) est $pq + qp$ donc $2pq$ ($F(Aa) = 2pq$).

Exemple 1.2.1 Si l'on considére une population de 100 individus dans laquelle nous étudions le gène qui détermine la couleur des yeux. Le caractère B est dominant, ce qui confère aux homozygotes BB et aux hétérozygotes Bb, la couleur de yeux bruns et le caractère b est récessif, ce qui confère aux homozygotes bb la couleur de yeux bleus. Si nous savons que dans cette population, les yeux bleus sont présents chez 21 personnes, on peut déterminer la fréquence des caractères B et b (la valeurs p et q) par la loi de Hardy-Weinberg :

$$q^2 = 0{,}21$$
$$q = 0{,}45$$

Comme $p + q = 1$:

$$p = 1 - 0{,}45$$
$$p = 0{,}55$$

Nous savons donc que la fréquence du caractère B est 0,55 et que la fréquence du caractère b est 0,45.

Maintenant que nous connaissons la distribution du caractère responsable de la couleur des yeux dans cette population, nous pouvons déterminer la probabilité des phénotypes et/ou génotypes lors du croisement de deux individus pris au hasard provenant de cette population.

Comme nous pouvons le constater, la formule d'Hardy-Weinberg est très utile pour déterminer la distribution des fréquences d'un caractère dans une population. Par contre, pour appliquer cette formule, nous devons faire certaines hypothèses dont la taille infinie de la population, qu'il n'y ait pas de migrations, qu'il n'y ait pas de mutations, qu'il n'y ait pas de sélection, que le nombre d'hommes soit égal au nombre de femmes et que cette population soit dans un *état panmictique*, c'est-à-dire que les unions se fassent au hasard et où il n'y ait pas de choix préférentiels ou évitements de conjoints. Toute modification aux conditions d'applications de cette hypothèse vient modifier l'état d'équilibre de la population. Nous verrons, dans ce qui suit, les conséquences que peut avoir le non respect de l'hypothèse de base.

1.2.3 Dérive génétique et effet fondateur

Lorsque la population est de petite taille, la fréquence des allèles d'un gène peut varier de façon beaucoup plus importante que dans une population de grande taille. En effet, dans une population de petite taille, par le jeu du tirage au hasard des gamètes reproducteurs, un individu possédant un gène rare et ayant une nombreuse descendance d'une génération à l'autre, aura une grande influence sur la fréquence de ce gène aux générations suivantes. Par contre, ce même individu, dans une population de grande

taille, n'aurait pas pu, à lui seul, avoir un impact significatif sur la fréquence de ce gène aux générations suivantes. C'est ce qu'on appelle la *dérive génétique*. À long terme, la dérive génétique tend à l'homogénéisation de la population car étant donné que l'on suppose que les unions se font au hasard, après un certain nombre de génération, un très grand nombre d'individus se retrouve avec une copie du même allèle (A ou a) et l'autre est perdu après plusieurs générations. La vitesse à laquelle cela se produit dépend de la fréquence initiale de l'allèle ainsi que de la taille de la population.

Exemple 1.2.2 Dans une population de 6 individus (3 hommes et 3 femmes) où on fait l'hypothèse qu'il y a un homme et une femme qui ont les yeux bleus (bb), les autres étants tous homozygotes dominants BB. Avec ces informations, nous pouvons calculer la fréquence des allèles b et B de cette population. Il y a en tout 12 allèles (2 par personnes) et les deux individus aux yeux bleus contribuent pour 4 allèles b. Il y a donc 4 allèles b et 8 allèles B ($F(q) = 0{,}33$ et $F(p) = 0{,}66$).

Si l'on supposes que le couple composé de l'homme aux yeux bleus et de la femme aux yeux bleus ait 6 enfants, nécessairement aux yeux bleus, et que les 2 autres couples ont chacuns 2 enfants qui ont les yeux bruns, puisque tous BB. Si on recalcule la fréquence des allèles à la nouvelle génération, on s'aperçoit que la fréquence des allèles B et b a grandement changé. En effet, à la nouvelle génération, il y a 10 enfants, donc 20 allèles. Il y en a 6 d'entres eux qui ont les yeux bleus, donc bb, qui comptent pour 12 allèles b. Les autre enfants homozygotes aux yeux bruns comptent alors pour 8 allèles B. Donc, à la nouvelle génération, la fréquence de l'allèle b est de 0,6 ($F(q) = 0{,}6$) et la fréquence de l'allèle B est maintenant de 0,4 ($F(p) = 0{,}4$).

Ce changement important dans les fréquences a été rendu possible par le fait que la population soit de petite taille.

L'*effet fondateur*, est impliqué lorsqu'un sous-groupe d'une population donnée forme une nouvelle population, par migration ou suite à un isolement. Lors de l'établissement de cette nouvelle population, étant donné que le nombre d'individus fondateurs est

restreint, il se peut que la distribution des fréquences des caractères présents dans cette population diffère de celle de la population d'origine, ce qui implique que la distribution chez leur descendants sera différente aussi.

Exemple 1.2.3 Reprenons la nouvelle génération de l'exemple 1.2.2 :
Si un sous-groupe quitte cette population et établit une nouvelle population et que cette sous-population est constitué de 5 enfants aux yeux bleus et d'un enfant aux yeux bruns, nous pouvons facilement constater que les fréquences des allèles sera différente de la population d'origine simplement due à l'effet d'échantillonnage ($F(p) = 0{,}17$ et $F(q) = 0{,}83$).

1.2.4 La migration

La dérive génétique qui, à long terme, tend à homogénéiser la population, n'est possible que dans le cas où la population est complètement isolée pour une très longue période. Dans les faits, il suffit qu'il y ait un mince filet d'immigration pour permettre de conserver une certaine diversité au sein de la population isolée. Ceci est dû au fait que les immigrants apportent une certaine quantité de gènes "neufs". De plus, même si le nombre d'immigrants est limité, le comportement démographique de ceux-ci peut faire en sorte que leurs gènes se répandent assez rapidement dans la population. En effet, bien souvent, les immigrants ont tendance à avoir une progéniture plus nombreuse que le reste de la population. Ceci peut s'expliquer par deux facteurs : ils possèdent un certain avantage biologique qui leur a permis de migrer de leur population d'origine (facteur biologique) ou bien ils tentent de recréer un monde qui leur est familier et ont donc un plus grand nombre d'enfants (facteur sociologique) (6).

1.2.5 La sélection naturelle

Un autre processus pouvant modifier l'état d'équilibre et amener l'évolution proprement dite, est la sélection naturelle. Ce terme fut introduit par Darwin afin de décrire le mécanisme par lequel une population peut accumuler les changements génétiques, donc

évoluer et, ultimement, mener à l'établissement d'une nouvelle espèce. On peut résumer la théorie de Darwin avec les trois principes suivants.

• Toute population est composée d'individus qui représentent une certaine variation morphologique, physiologique ou comportementale.

• Les descendants ressemblent davantage à leurs parents qu'à des individus non apparentés donc ces variants sont héréditaires et ils confèrent ou non des avantages à ceux qui les possèdent.

• Étant donné un certain milieu, certains descendants réussiront mieux que d'autres à survivre et à se reproduire et ainsi transmettront leurs caractéristiques à plus de descendants.

Il est facile de concevoir que si un gène avantage un organisme en terme du nombre de descendants qu'il peut produire, ce gène se retrouvera en plus grande quantité dans les générations successives. La sélection naturelle permet d'établir un nouvel équilibre propre au milieu dans lequel se trouve la population. La sélection, par elle-même, ne permet pas l'apparition de nouveaux gènes, mais favorise ceux qui confèrent un avantage reproductif ou adaptatif. Le seul cas, dans la nature, où un nouveau gène "apparaît"est lors d'une mutation.

1.2.6 La mutation

La *mutation* est un évènement rare mais qui demeure la source originelle de la variation. Cet évènement modifie le patrimoine génétique de l'individu de telle sorte qu'il est différent de celui légué par ses parents et a donc un impact sur la descendance de l'individu. Il est important de savoir qu'une mutation, pour être transmise à la génération suivante, doit se produire à l'intérieur des cellules germinales. Les *cellules germinales* sont les cellules qui sont destinées à la reproduction de l'organisme, elles se développent pour former des ovules ou des spermatozoïdes. Ainsi, une mutation dans les cellules de la peau qui causerait un cancer, par exemple, ne sera pas transmise aux descendants de cet

individu, par contre, si cette mutation se produit à l'intérieur d'une cellule germinale, il se peut qu'elle soit transmise à la génération subséquente lors de la reproduction.

1.3 Études réalisées avec le support d'Analypop

Depuis sa création, en 1987, Analypop a été un outil informatique important utilisé pour alimenter plus d'une quinzaine de recherches dans les laboratoires de Francine M. Mayer à l'UQÀM, de Gilles Boetsch du CNRS et de Manuela Lima de l'Université des Açores. Mentionnons quelques unes des utilisations qui en ont été faite.

- Au Portugal, dans l'archipel des Açores, Manuela Lima, professeur à l'Université des Açores, et son équipe ont constitué un registre à l'aide du logiciel Analypop pour ensuite effectuer des analyses d'épidémiologie génétique en remontant les ascendances d'individus atteints de la maladie de Machado-Joseph, une maladie due à une mutation qui ajoute une répétition dans le code génétique. La reconstitution des ascendances a permis de remonter à deux groupes d'individus fondateurs provenant de la côte du Portugal ayants introduit la mutation (11).
- Une autre équipe franco-italienne, l'équipe de Emma Rabino-Massa, de l'Université de Turin, et Gilles Boetsch, de l'Université de la Méditerranée, s'est intéressée à la vallée de Vallouise en Briançon, dans les Hautes-Alpes, et a reconstitué la structure généalogique de cette population à l'aide du module de reconstitution de familles d'Analypop (17). Cette reconstitution a ensuite permis d'étudier les stratégies de reproduction de deux classes d'individus : les décideurs, qui étaient historiquement choisis parmi les riches, et les fermiers. En étudiant le coefficient de consanguinité à l'intérieur de ces deux classes, ils ont pu remarquer que l'apparentement des individus de la classe des fermiers était légèrement plus grand que celle des décideurs. Dans ce cas-ci, Analypop servi a gérer le registre de population ainsi qu'à reconstituer la structure généalogique de la population.
- Toujours pour la même population, l'équipe de Gilles Boetcsh a étudié le nombre d'enfants naissants et d'enfants utiles (enfants qui auront à leur tour une descendance) (1). Ils ont utilisé le logiciel Analypop afin d'identifier, à partir du registre de

population, les enfants utiles. Cette étude leur permit de conclure que le nombre d'enfants utiles est relativement peu élevé si on le compare au total du nombre d'enfants naissants.

- Analypop a été développé pour répondre aux besoins des recherches du laboratoire de Francine M. Mayer portant sur la dynamique bioculturelle de petites populations humaines isolées et l'épidémiologie génétique de certaines pathologies à composantes héréditaire dans la Caraïbe et au Québec. Présentement en cours, l'étude du laboratoire de Francine M. Mayer concernant les contributions des immigrants franco-américains au patrimoine génétique de la population des Cantons de l'Est permettra d'ajouter au registre de population de La Patrie l'information concernant l'origine de ses fondateurs. Une fois que le jumelage sera effectué, en remontant les ascendances au États-Unis et au Québec, ils pourront réexaminer les contributions génétiques des fondateurs de La Patrie. Dans ce cas-ci, Analypop sert à jumeler les individus ainsi qu'à gérer le registre de population.

1.4 Détail du processus de réingéniérie d'Analypop

Tout d'abord, il serait important de définir le sens accordé au terme réingéniérie dans le présent document. Ici, la réingéniérie est le processus par lequel nous avons analysé l'ancienne version d'Analypop, réévalué les besoins des chercheurs pour finalement développer une nouvelle version du logiciel Analypop.

La réingéniérie de cette application s'est effectuée en plusieurs étapes. Chacunes de ces étapes permettant de bien préparer la prochaine pour enfin arriver au résultat final avec un produit qui répond bien aux besoins et exigences des chercheurs.

La première étape fût d'étudier la faisabilité de cette réingéniérie. Cette étape a été en partie effectuée par Joseph Kamwena, étudiant à l'École Polytechnique de Montréal dans le cadre d'un projet de fin d'études.

Ensuite, un comité, formé de Jean-Charles Bernard, de l'École Polytechnique de Montréal, Joseph Kamwena, Francine M. Mayer ainsi que Mireille Boisvert de l'UQÀM, et Etienne

Morency-Bachand, étudiant en bioinformatique à l'UQÀM, s'est penché sur les besoins et les exigences pour la nouvelle version du logiciel. Il était clair, dès cette étape, que la réingéniérie concernait une réécriture complète du logiciel et non pas une amélioration de l'ancienne version. De plus, nous avons pu cerner les besoins des chercheurs ainsi que de définir un langage commun qui nous a permis, par la suite, de mieux se comprendre. Ces rencontres qui eurent lieu durant l'été 2004 ont permi d'identifier les technologies qui allaient être utilisées pour l'implantation de la nouvelle version. Le point culminant de ces rencontres menèrent d'ailleur à présenter la venue du nouveau logiciel aux entretiens Jacques-Cartier à Montréal, en septembre 2004 (13).

Durant ce temps, plusieurs prototypes ont été développés afin de préciser les fonctionnalités du futur logiciel et de permettre d'éclaircir les processus sous-tendants le travail d'un chercheur en démographie historique.

Étant donné la grande disponibilité des principaux chercheurs, Mireille Boisvert et Francine M. Mayer, l'étape suivante fut de faire l'ébauche du programme en terme d'écrans de saisies et de quelques fonctionnalités, ceci excluant le jumelage qui était jugée critique mais nécessitant encore une meilleure compréhension entre les membres en informatique et en biologie afin de produire un résultat correct.

Durant les deux années qui suivirent, les efforts se sont concentrés sur la rédaction du présent mémoire ainsi que sur le développement et l'amélioration du logiciel Analypop. Vers l'hiver 2006, l'application étant jugée stable et possédant les fonctionnalités de base nécessaires, une présentation a été organisée avec les membres du comité initial afin de présenter le travail effectué et permettre certains ajustements. Suite à cette rencontre, la manière dont était géré l'historique a été modifiée afin de le rendre encore plus flexible.

L'implantation finale pour permttre d'être testée officiellement était prévue à l'automne 2007, mais étant donné des contraintes d'horaire et des besoins plus urgents de la part de l'équipe de Francine M. Mayer, elle fût reportée. Il reste donc à implanter le logiciel au laboratoire de Francine M. Mayer pour qu'elle puisse valider la nouvelle version.

Étant donné mon implication dans l'équipe de recherche de Francine M. Mayer, je demeure présent pour supporter les futurs développements ou pour encadrer de futurs étudiants qui viendraient faire un stage dans le cadre du programme en bioinformatique.

CHAPITRE II

BASE DE DONNÉES ET IMPORTATION DES DONNÉES

L'outil de base en génétique des population humaines demeure le registre de population. C'est grâce à ce registre que l'on peut suivre l'évolution biologique en reconstruisant l'histoire reproductive des individus qui la composent. Ce chapitre présente tout d'abord l'historique des méthodes pour constituer un tel registre, pour ensuite s'intéresser aux concepts relatifs aux bases de données. Ceci nous permettra d'établir les bases conceptuelles nécessaires afin de pouvoir présenter la structure du modèle de données qui a été développée pour le logiciel Analypop.

2.1 Base de données

2.1.1 Historique

La première incursion de l'informatique dans le domaine de l'anthropologie biologique et de la démographie historique a été de remplacer le mode d'archivage des fiches de familles qui se faisait sur papier afin d'en faire la sauvegarde sur ordinateur. Les *fiches de familles* regroupent les informations relatives à un individu et à ses parents, son(sa) conjoint(e) et ses parents ainsi que les enfants de ce couple : elles sont à la base des analyses en démographie historique. Comme nous l'avons mentionné plus haut, le *registre de population* est l'ensemble des informations jumelées provenant de l'état civil ; il permet la consultation des fiches de familles et, entre autres, la reconstitution de généalogies *ascendantes* (on part d'un individu et on remonte vers tous ses ancêtres)

ou *descendantes* (on part d'un ancêtre et on descend vers tous ses descendants). Les Figures 2.1 et 2.2 présentent les deux types de généalogies rendues possibles grâce au registre, respectivement une généalogie ascendante et descendante.

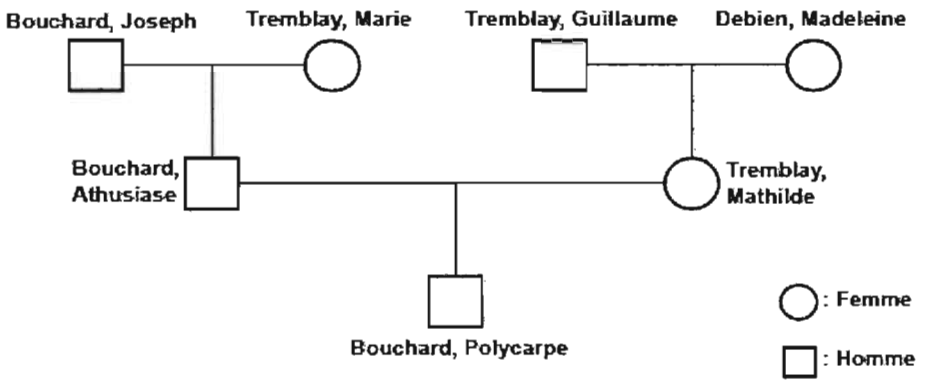

Fig. 2.1 Généalogie ascendante de l'individu Polycarpe Bouchard. Un trait horizontal entre deux individus définit une union et un trait vertical définit un enfant.

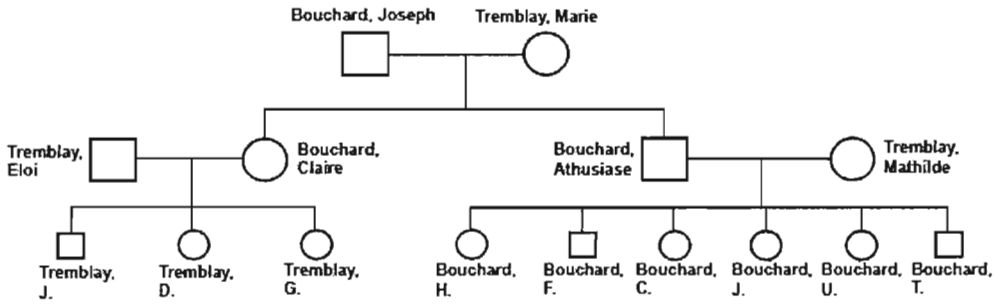

Fig. 2.2 Généalogie descendante de l'individu Joseph Bouchard.

En examinant la Figure 2.2, on remarque que les trois mariages concernent chacun un individu Bouchard et un individu Tremblay. Ceci n'est pas une erreur mais représente as-

sez bien la situation qui prévalait au Québec il y a quelques siècles. En effet, étant donné le type de population, les effectifs et la fécondité différentielle de certains immigrants ou de certains descendants, il n'est pas surprenant de trouver de telles situations. Ceci rend encore plus complexe le travail des démographes lorsqu'ils tentent de jumeler l'information à partir des actes dépouillés. En effet, la méthode qu'utilisent les démographes pour jumeler consiste à rapprocher toutes les mentions de couples dans tous les actes d'état civil. À partir des actes de mariage et à partir des couples mentionnées comme les parents du marié et de la mariée ils tentent de retrouver l'acte de mariage des parents et ainsi remonter les générations. On peut comprendre que plus l'information nominative présentée sur ces actes est unique, plus il est facile par la suite de faire le jumelage des individus. Dans le cas où l'on rencontre des mentions de couples homonymes, le démographe doit confronter d'autres données dépouillées et faire la comparaison des dates de naissance, des dates de décès, calculer l'intervalle entre les naissances etc. afin de distinguer quelle information doit être attribuée à quel couple.

Il faut comprendre qu'à l'époque des premières tentatives de reconstitution de registres de population informatisés, dans les années 1970, la tâche consistait, à partir des actes de baptêmes, de mariages et de sépultures, à créer sur papier des fiches de dépouillement pour chaque individu mentionné pour ensuite pouvoir les mettre en relation et permettre d'avoir une vision de l'état de la population à une certaine époque ou pour suivre l'évolution de cette population. La mise en relation des fiches de dépouillement permettaient de créer les fiches de famille et attribuer des matricules unique aux individus. Cette mise en relation des fiches était autrefois fait à la main, travail qui pouvait devenir extrêmement laborieux dès que la population étudiée comportaient plus que quelques centaines d'individus. L'informatisation de ces fiches de famille a tout d'abord permis d'offrir un support pour la sauvegarde des données, et ensuite, un support pour faciliter la consultation et la mise en relation des individus pour constituer un registre de population totalement informatisé. À l'aide de ce registre, les démographes et les historiens peuvent ensuite faire des analyses qui leur permettent de formuler certaines observations sur la nature et la dynamique des populations.

À cet effet, il est intéressant de mentionner que plusieurs considèrent la première base de données comme étant le recensement américain de 1890. Afin de pouvoir traiter l'information d'une population en pleine croissance, Herman Hollerith, un statisticien américain, inventa une machine à perforer des cartes qui permit de traiter les données du recensement en seulement 18 mois (20). Le recensement précédent de 1880 aurait été traité en 7 ans et, due à l'immigration de plusieurs millions de personnes entre les années 1880 et 1890, on estimait que le recensement de 1890 prendrait plus de 13 ans à traiter, ce qui aurait rendu presque inutile le fait de faire un recensement. Les cartes perforées utilisées pour le recensement de 1890 furent inspirées de la manière dont les contrôleurs de train poinçonnaient les titres de transport pour encoder l'apparence physique de la personne qui les possédaient (grandeur, sexe, types de cheveux...), ainsi que sa destination.

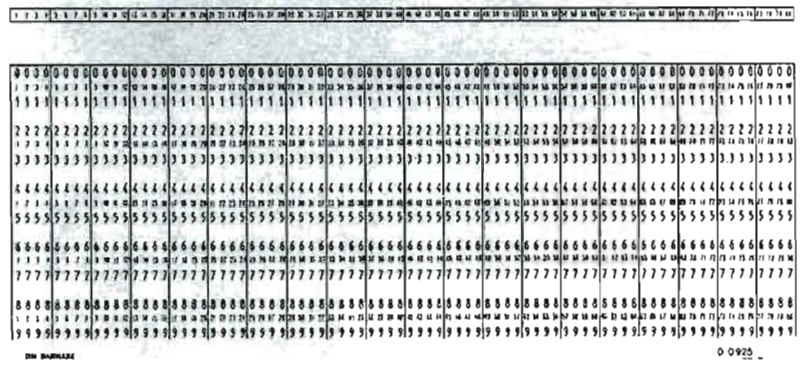

Fig. 2.3 Exemple d'une carte perforée.

Les premières tentatives d'informatisation des fiches de dépouillements ont été effectuées effectivement en sauvegardant ces fiches sous forme de cartes perforées (Fig. 2.3). Par la suite, lorsque les scientifiques voulurent utiliser ces données afin de les analyser, il fallu choisir un type de représentation de toutes ces données. Le type de base de données choisie fût la base de données simple, appelée aussi "*flat file database*".

Une *base de données simple* consiste à sauvegarder les informations les unes à la suite des autres dans un fichier texte, un enregistrement par ligne. Un *enregistrement* est

une unité d'information qui contient l'information nécessaire pour représenter un objet quelconque où chaque partie cohérente de cet enregistrement est appelé un *champ*.

Exemple 2.1.1 Exemple d'un enregistrement composé de champs pour définir l'adresse civique d'une personne :

Nom	Prénom	No. civique	Rue	Ville	Province	Pays	Code postal
Huet	Antoine	4572	Laforge	Rimouski	Québec	Canada	G5L 3A1

Lorsque les champs sont écrits dans une base de données simple, on doit pouvoir être en mesure de retrouver l'information par la suite, il est donc important de bien la structurer. Deux méthodes sont utilisés pour écrire les données dans une base de données simple de façon à ce que les champs puissent être retrouvés dans le même état qu'ils ont été enregistrés. La première consiste à séparer les champs avec un *délimiteur*, c'est-à-dire, un caractère qui signifie la fin du champ courant et le début du champ suivant.

Exemple 2.1.2 Enregistrement avec délimiteur de champs :

```
John$Coltrane$1926-23-10$
Ella$Fitzgerald$1917-25-04$
```

La deuxième méthode consiste à utiliser des champs de longueurs fixe et de compléter chaque champ avec un caractère neutre.

Exemple 2.1.3 Enregistrement avec champs de taille fixe :

```
John      Coltrane      1926-23-10
Ella      Fitzgerald    1917-25-04
```

Étant donné ce contexte informatique, les démographes, lorsqu'ils voulurent construire les premiers registres de population informatisés, ont utilisé une structure de base de données simple constituée de fichiers principaux : un fichier INDIVIDU et un fichier UNION. Le fichier INDIVIDU contient les informations relatives aux individus tels que

leur nom, prénom, date de naissance etc., tandis que le fichier UNION permet de mettre en relation deux individus, de conserver l'information concernant le couple (date et lieu de mariage lorsque disponible) et de leur associer leurs enfants.

Exemple 2.1.4 Voici un exemple d'une base de données simple avec un fichier INDIVIDU et UNION :

INDIVIDU					UNION		
Individu	Prénom	Nom	Date de naissance	Issu de	Union	Homme	Femme
1	John	Coltrane	1926-23-10	100	100	3	4
2	Ella	Fitzgerald	1917-25-04	101	101	5	6
3	Lester	Coltrane	1902-13-05	102	...		
4	Emma	Young	1904-22-11	105			
5	Coleman	Fitzgerald	1897-10-05	124			
6	Lucy	Hawkins	1899-04-12	132			

On peut voir que le fichier UNION permet de définir les parents des personnes présentes dans le fichier INDIVIDU. Par exemple, la première ligne du fichier UNION associe les individus Lester Coltrane (individu numéro 3) et Emma Young (individu numéro 4) avec leur enfant John Coltrane (individu 1 dont le champ Issu de est 100).

Les bases de données simples se révèlent rapidement inadéquates lorsque l'on traite un grand volume de données. Leurs deux principaux inconvénients sont qu'étant donné que les données sont sauvegardées dans des fichiers texte, il est fréquent que des erreurs de corruption surviennent. De plus, ce type de fichier n'est pas vraiment adapté à gérer les relations entre les données puisque la recherche de l'information à l'intérieur de ces fichiers doit se faire de façon séquentielle, ce qui ralentit la mise en relation d'informations.

De nos jours, le modèle de base de données le plus couramment utilisé est le modèle *relationnel*. Une récente étude nous indique qu'il y a environ 70% des bases de données scientifiques qui sont de ce type (8). Ce modèle permet de présenter les données sous forme de tables, chacune regroupant un sous-ensemble cohérent des données. Les différents

champs d'information déterminent les colonnes de la table et chaque ligne représente un enregistrement. Les bases de données relationnelles permettent une plus grande indépendance ainsi qu'une meilleure gestion des données. De plus, ce modèle permet facilement d'ajouter de nouveaux champs ou de modifier la taille de ceux-çi alors que le modèle de base de données simple avec champs de taille fixe, ne le permet pas. Même si, maintenant, la plupart des bases de données de registre de population sont relationnelles, les termes fichier INDIVIDU et fichier UNION sont encore couramment utilisés.

2.1.2 Choix du système de gestion de base de données

Un système de gestion de base de données est un logiciel ou un groupe de locigiels permettant de créer, de gérer et d'interroger efficacement une base de données. Les systèmes de gestion de base de données, en général, offrent plusieurs fonctionnalités dont entre autres : gérer les utilisateurs accédant à la base de données, assurer de conserver la cohérence des données telles que définies par la base de données et permettre la recherche rapide d'informations. Pour ce projet, nous avons choisi le système de gestion de base de données MySQL 5.0. En plus d'être un logiciel ayant faits ses preuves pour plusieurs projets d'envergures (Ensembl Genome Browser, Los Alamos National Laboratory, Google, Bureau de recensement américain etc.), il a comme avantage d'être sous licence GPL, ce qui signifie notamment qu'il peut être utilisé à des fins académiques sans frais (10). Les premiers développements de MySQL remontent au début des années 1980 et depuis, plusieurs versions ont été proposées au grand public. MySQL a su s'imposer comme étant une alternative viable aux logiciels de gestion de base de données commerciaux tels que Oracle ou Microsoft SQL Server (14). Sans faire une comparaison détaillée de MySQL avec les autres systèmes de gestion de base de données, nous allons tout de même présenter les principales fonctionnalités ayant permis de choisir ce logiciel plutôt qu'un autre.

Accessibilité

Le fait que MySQL puisse être accessible à travers une *interface de programmation*, c'est-à-dire qu'il permette d'utiliser certaines de ses fonctions à l'aide d'un langage

de programmation (dans notre cas Java) nous permet d'intégrer un grand nombre de fonctionnalités de manipulation de données dans Analypop sans avoir à recourir à des programmes externes en particulier la recherche d'information et l'ajout de nouvelles données. Sans cette interface, il faudrait utiliser un programme spécifique pour communiquer avec la base de données.

Portabilité

La très grande portabilité de MySQL répondait au besoin que l'on avait identifié lorsque nous avions choisi le langage de programmation Java. Un des objectifs visés par la réingéniérie d'Analypop était de le rendre plus *portable*, c'est-à-dire qu'il puisse être installé sur plus d'un type de système d'exploitation. Dans cette optique, il aurait été trop restrictif de choisir un système de gestion de base de données qui ne fonctionne que sur un seul système d'exploitation. Cet aspect a fait en sorte de privilégier MySQL, puisqu'il peut être installé sur plus d'une trentaine de systèmes d'exploitation différents, ce qui ne limitera pas les futurs utilisateurs dans leurs choix.

Sécurité

MySQL permet une gestion sécuritaire des accès à la base de données en obligeant les utilisateurs à avoir un identifiant d'utilisateur et un mot de passe. Puisque les utilisateur doivent s'identifier, il est possible de restreindre l'accès ou les privilèges à certaines catégories d'utilisateurs. Les *privilèges* que peuvent avoir les utilisateurs concernent principalement la consultation, l'ajout, la modification ou la suppression de données dans la base de données. Ainsi, en gérant différents types d'utilisateurs, il est possible de permettre à une personne de consulter, d'ajouter, de modifier et de supprimer des données alors qu'un autre ne pourrait qu'avoir le droit de consulter les données. Dans le cas d'Analypop, nous pouvons nous assurer ainsi qu'un seul utilisateur puisse être responsable des nouvelles données ajoutées tout en permettant à d'autres utilisateurs de consulter les données du registre de population sans toutefois pouvoir y faire de modifications.

Connectivité

Le fait que l'on puisse se connecter au serveur de base de données MySQL à travers un réseau permet d'installer le registre de population sur un ordinateur dédié et ainsi s'assurer de la centralisation des données. Dans le cas du registre de population, le fait que la base de données soit centralisée garantit que tous les utilisateurs travaillent sur le même registre et profitent des mêmes modifications.

Gratuité

MySQL rend l'utilisation de son logiciel libre de tous frais si cette utilisation se fait dans un cadre académique. Ceci est extrêmement avantageux, sachant que certains logiciels de gestion de base de données peuvent coûter jusqu'à plusieurs milliers de dollars.

2.1.3 Architecture de la base de données

L'achitecture de la nouvelle base de données a été en partie influencée par la structure de l'ancienne version d'Analypop, étant donné qu'un des objectifs visés de la réingéniérie était de pouvoir conserver les données du registre de population. Par contre, nous verrons vers la fin de ce chapitre que nous avons utilisé une technique qui permet d'adapter Analypop à diverses bases de données. Le modèle de données présentement proposé pour le logiciel Analypop est présenté à la Figure 2.4.

La structure de base du modèle de données est constitué de six tables. On remarque que l'on retrouve les tables INDIVIDU et CONJOINT, qui correspondent aux tables INDIVIDU et UNION des démographes. À propos du nom de la table CONJOINT, il était prévu de nommer cette table UNION mais étant donné que ce mot est un mot réservé de MySQL, nous avons choisi de changer le nom de cette table pour CONJOINT.

Une des premières choses que l'on peut remarquer à l'examen de ce modèle est la presque absence de clés assurant l'intégrité référentielle. L'*intégrité référentielle* est le mécanisme par lequel on peut garantir une certaine cohérence des données. Si la suppression ou la modification de données peut enclencher des problèmes sur une table mise en relation, le Système de Gestion de Base de Données l'empêche ou applique la modification ou la

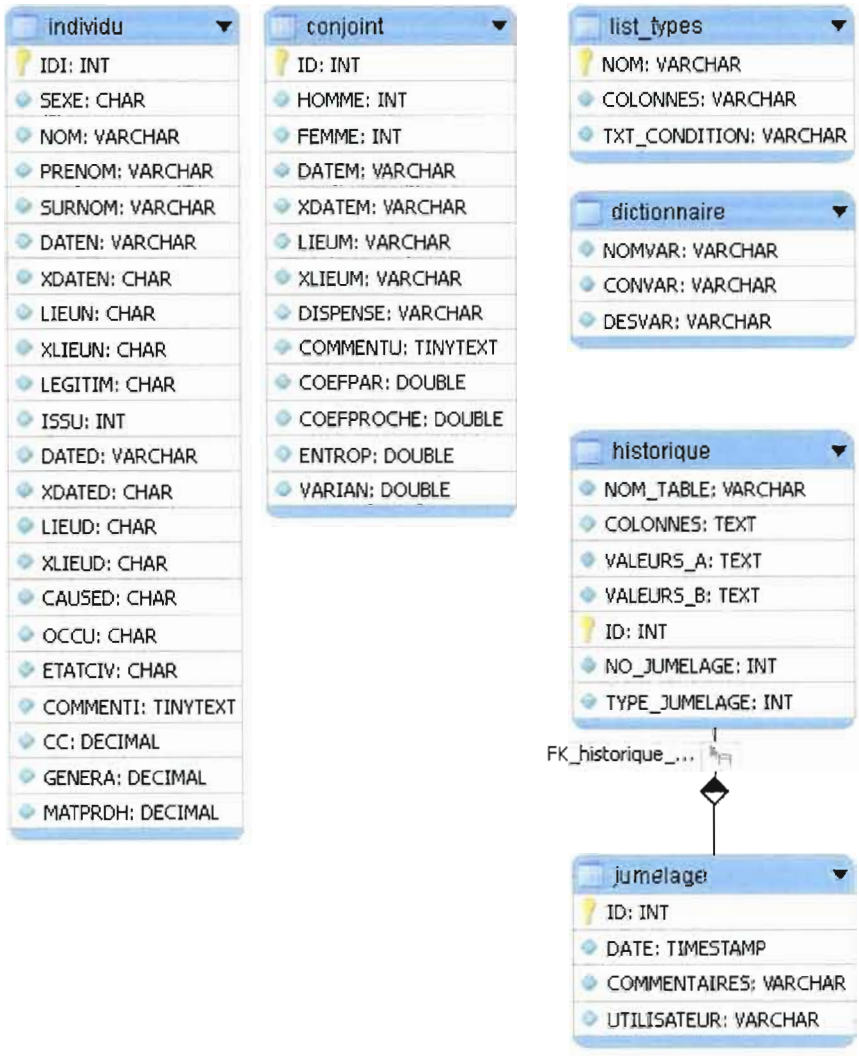

Fig. 2.4 Modèle de la base de données d'Analypop.

suppression à la table liée. Par exemple, il serait plus prudent de définir une contrainte d'intégrité référentielle entre le numéro d'union duquel l'individu est issu (ISSU) de la table INDIVIDU et le numéro d'union (ID) dans la table CONJOINT. Ainsi, si par inadvertance on effacait une union pour laquelle un ou plusieurs individus y font référence, le système de gestion de base de données empêcherait cette opération puisque la référence

entre la table INDIVIDU et CONJOINT serait alors corrompue.

Dans notre cas, puisque certaines données ne satisfaisaient pas à cette contrainte, nous ne pouvions pas implémenter cette contrainte d'intégrité référentielle. Par exemple, nous ne pouvions pas obliger le SGBD à vérifier que le matricule de l'homme et celui de la femme composant une union soient définis dans la table INDIVIDU puisque certaines données déjà présentes ne respectaient pas cette règle.

2.1.4 Description de la table INDIVIDU

Cette table regroupe les informations qui concernent les individus du registre de population. Dans cette table, la seule information qui est requise est le numéro d'individu : IDI. Ce numéro est primordial afin de lier l'individu aux autres tables qui contiennent de l'information le concernant.

À l'intérieur même de la table INDIVIDU, les individus sont regroupés en sous-groupes selon leur numéro de matricule IDI. En effet, lors du dépouillement, les chercheurs ont réservé certaines plages de nombres afin d'identifier la provenance de l'information. Ainsi, on peut déterminer rapidement, selon le numéro de matricule des individus, d'où proviennent les informations les concernants. Chaque projet peut définir ses propres intervalles de matricules et leur donner la signification qu'il veut.

INDIVIDU

CHAMP	DÉFINITION
IDI	Matricule (entre 0 et 4294967295)
SEXE	Sexe (M ou F)
NOM	Nom
PRENOM	Prénom
SURNOM	Autre nom sous lequel on retrouve parfois l'individu
	(par exemple, Marguerite Duperron est parfois mentionnée comme Marguerite Dugrenier.
	Le prénom est donc Marguerite, le nom Duperron et le surnom Dugrenier)
DATEN	Date de naissance
XDATEN	Code indiquant la provenance de l'information de la date de naissance
	(par exemple, le code AC signifie que la date provient du contrat de mariage)
LIEUN	Identifiant du lieu de naissance
XLIEUN	Code indiquant la provenance de l'information du lieu de naissance
	(par exemple, le code W signifie que le lieu a été trouvé par Stephen White,
	un généalogiste de l'Université de Moncton)
LEGETIM	Code de la légitimité (indique la provenance de l'individu,
	par exemple, le code 4 signifie que l'individu est adopté,
	le code 5 signifie que l'individu est d'origine amérindienne etc.)
ISSU	Numéro d'union du couple ayant eu cet individu
DATED	Date de décès
XDATED	Code indiquant la provenance de l'information de la date de décès
LIEUD	Identifiant du lieu de décès
XLIEUD	Code indiquant la provenance de l'information du lieu de décès
CAUSED	Cause de décès
OCCU	Code d'occupation ou de profession
ETATCIV	État civil qui peut être soit marié(e), divorcé(e), veuf(ve), célibataire.
COMMENTI	Commentaires
CC	Valeur calculée : le coefficient de consanguinité
GENERA	Valeur calculée : numéro de génération
MATPRDH	Valeur calculée : matricule attribué par le PRDH
	(le Programme de Recherche en Démographie Historique de l'Université de Montréal)

2.1.5 Description de la table CONJOINT

Cette table permet d'associer deux individus pour définir un évènement d'union. De plus, c'est grâce à cette table que l'on peut définir une famille. Lorsqu'on inscrit un couple, en spécifiant un matricule pour le champ HOMME et un autre pour le champ FEMME, on définit en fait une union. Si on assigne le même numéro que le champ ID dans le champ ISSU de la table INDIVIDU, on définit un enfant de ce couple. La recherche de tous les enfants de ce couple se fait donc en parcourant la table INDIVIDU et en sélectionnant tous les individus ayant le champ ISSU égal au numéro du couple concerné.

CONJOINT

CHAMPS	DÉFINITION
ID	Matricule associée à l'union
HOMME	Matricule de l'homme
FEMME	Matricule de la femme
DATEM	Date de mariage
XDATEM	Code indiquant la provenance de l'information de la date de mariage
LIEUM	Lieu de mariage
XLIEUM	Code indiquant la provenance de l'information du lieu de mariage
DISPENSE	Code de la dispense s'il y a lieu (une dispense devait être accordée si les époux étaient apparentés ou si l'un deux avait déjà été marié)
DATEFU	Date de fin d'union
XDATEFU	Code de date de fin d'union
LIEUFU	Lieu de la fin d'union (lieu de décès)
CAUSEFU	Fin union divorce, séparation, décès
COMMENTU	Commentaires
COEFFPAR	Valeur calculée : coefficient de parenté
COEFFPROCHE	Valeur calculée : coefficient de consanguinité proche
ENTROP	Valeur calculée : entropie
VARIAN	Valeur calculée : variance

2.1.6 Description de la table LIST_TYPES

Cette table permet de sauvegarder les listes que peuvent définir les utilisateurs. Les chercheurs ont souvent besoin d'examiner des listes, triées selon différents critères afin de pouvoir plus facilement consulter les données ou encore pour structurer l'information en vue d'analyses. Analypop offre donc la possibilité de définir les colonnes qu'on veut afficher à l'écran et de sauvegarder cette configuration de liste afin de permettre aux utilisateurs de réutiliser des types de listes déjà définies.

LIST_TYPES

CHAMPS	DÉFINITION
NOM	Nom que l'on donne au type de liste définit
COLONNES	Nom des colonnes qui devront apparaître dans la liste
TXT_CONDITION	Condition permettant de restreindre le nombre de lignes dans la liste ou permettant de faire le lien entre plusieurs tables de la base de données.

2.1.7 Description de la table DICTIONNAIRE

La table DICTIONNAIRE agit à titre de dictionnaire. Cette table contient toute l'information concernant les codes utilisés dans les autres tables.

DICTIONNAIRE

CHAMPS	DÉFINITION
NOMVAR	Nom de la colonne qui utilise le code
CONVAR	Code
DESVAR	Définition du code

2.1.8 Description des tables JUMELAGE et HISTORIQUE

Le *jumelage* est l'action par laquelle on détermine un lien entre deux individus. On peut définir que deux individus ont un lien de parenté entre eux ou bien définir que deux individus présents dans le registre de population en formation sont en fait le même individu.

Il a été très clair, lors des rencontres préliminaires, que toutes les actions de jumelage faites avec le logiciel Analypop devaient être réversibles. Pour ce faire, à chaque fois que l'on modifie une table suite à un jumelage, on sauvegarde les données qui ont été modifiées dans une table à part. De plus, en sauvegardant l'information sur la date à laquelle le jumelage a été fait ainsi qu'un commentaire justifiant ce jumelage, il est plus facile par la suite de se rappeler le contexte qui a mené à prendre cette décision. Les tables JUMELAGE et HISTORIQUE servent à faire cette gestion.

La table JUMELAGE sert à sauvegarder les informations concernant le contexte d'une décision de jumelage. On y retrouve la date, le nom de l'utilisateur ayant effectué le jumelage ainsi qu'un commentaire expliquant la raison qui a permis de faire ce jumelage. Grâce à ces informations, lorsque l'utilisateur consulte l'historique des jumelages dans le but d'annuler un de ceux-ci, il lui est plus facile de reconnaître spécifiquement en quoi consistait chaque jumelage.

JUMELAGE

CHAMPS	DÉFINITION
ID	Numéro unique qui identifie le jumelage
DATE	Date à laquelle le jumelage a été fait
COMMENTAIRES	Commentaires entrés par les utilisateurs concernant le jumelage
UTILISATEUR	Nom de l'utilisateur qui a effectué le jumelage
	(basé sur les utilisateurs tels que définis avec MySQL)

La table HISTORIQUE pour sa part, conserve les données qui permettent de revenir à l'état précédent un jumelage. On sauvegarde dans cette table le nom des colonnes qui ont été modifiées ainsi que leurs valeurs avant le jumelage. Nous avons défini trois types

de jumelage, soit : "égalité", "conjoint" ou "issu de". Le jumelage "égalité" permet de décider que deux individus présents dans le registre de population sont en fait le même individu. Le jumelage "conjoint " permet d'associer un conjoint à un individu et ainsi définir un couple dans la table CONJOINT. Finalement, le jumelage de type "issu de" permet d'associer un individu à ses parents ou de définir des liens de types : frère-frère, frère-soeur ou soeur-soeur lorsqu'on donne à deux individus le même numéro de couple parents.

HISTORIQUE

CHAMPS	DÉFINITION
NOM_TABLE	Nom des tables affectées par le jumelage
COLONNES	Colonnes affectées
VALEURS_A	Valeurs qu'avaient l'individu A
VALEURS_B	Valeurs qu'avaient l'individu B
ID	Identifiant unique
NO_JUMELAGE	Réfère au jumelage de la table jumelage
TYPE_JUMELAGE	Type du jumelage (égalité, conjoint ou issu de)

2.1.9 Fichiers de propriétés

Le modèle de la base de données présenté ci-haut est un modèle minimal, c'est-à-dire que l'on a prévu la possibilité d'ajouter de nouvelles tables pour répondre plus spécifiquement au domaine d'études de chacun des utilisateurs. Cette adaptabilité d'Analypop est assurée par un fichier de propriétés. Un *fichier de propriété* est un fichier de type texte dans lequel on associe un mot-clé à une valeur à l'aide du caractère "=". Ces fichiers sont généralement utilisés pour définir des options de configuration propre à un poste de travail ou à un projet. On peut donc se servir de ces fichiers pour que le logiciel s'adapte à la langue de l'utilisateur, en définissant chacun des mots ou phrases que le logiciel utilise, ou pour définir le type d'environnement sur lequel le logiciel doit s'exécuter. Dans le cas présent, nous avons utilisé ce fichier (analypop.properties) dans l'optique de rendre le logiciel Analypop le plus flexible possible.

Exemple 2.1.5 Format d'un fichier de propriétés :

```
# Nom du serveur de base de données
server = localhost
# Nom d'utilisateur
user_name = analypop

# Nom de la table INDIVIDU
indiv = individu
# Nom de la table UNION
union = conjoint

# Délimiteur de champs
delimiteur = $

# Nom de la vue à utiliser
principale.1.vue = v_individu
principale.2.vue = v_conjoint

# Noms des sections
principale.titre_sections = Individu$Union$
```

Notons tout d'abord que le caractère "#" permet d'insérer des commentaires dans le fichier. On peut remarquer que la façon d'utiliser un fichier de propriétés est d'assigner du texte à différentes variables. Par la suite, à l'intérieur du programme Java, une commande nous permet de charger ces variables dans un objet de type *Properties* à partir duquel on peut obtenir la valeur associée à une variable en faisant un appel de la méthode *get(nom de la variable)*.

D'après le fichier de l'exemple 2.1.9, on peut remarquer que, dans le cas d'Analypop, le fichier sert à configurer trois aspects du logiciel.

Base de données

Le fichier permet de décrire la base de données. On y spécifie l'adresse du serveur de base de données, l'utilisateur que nous utilisons, son mot de passe et même le nom que portent les tables INDIVIDU et CONJOINT. Ainsi, Analypop pourrait utiliser une autre base de données que celle développée dans le cadre du projet actuel, en spécifiant les bons paramètres dans le fichier de propriétés.

Sources de données

Le fait que les interfaces soient basées sur les vues spécifiées dans le fichier de propriétés (principale.1.vue et principale.2.vue), nous permet facilement d'ajouter de nouvelles sources de données sans avoir à modifier le code d'Analypop. Une *vue* est un sous-groupe de colonnes provenant du résultat d'une requête. Donc, si l'on désirait ajouter de nouvelles sources de données, par exemple : provenant d'actes notariés, des actes de divorce ou des recensements, on n'aurait qu'à ajouter une ligne dans ce fichier qui indiquerait le nom de la table ou de la vue où prendre l'information dans la base de données et le logiciel permettrait de consulter cette nouvelle source de données.

Interface graphique

On peut configurer une partie de l'interface graphique en inscrivant le nom que l'on désire voir apparaître au haut de chacune des sections (principale.titre_sections). C'est de cette manière que l'on pourrait éventuellement traduire l'interface d'Analypop afin de rendre le logiciel accessible dans d'autres langues.

2.2 Importation des données

2.2.1 Saisie de données

Maintenant que nous avons abordé l'aspect de la structure de la base de données ou du registre de population, il sera intéressant d'aborder l'aspect de l'importation des données, ou comment on doit procéder pour remplir les tables du registre de population. Pour effectuer la saisie de données, nous avions identifié deux possibilités. La première

étant de créer un module de saisie à même Analypop et la deuxième étant de conserver la méthode actuelle de saisie qui se fait à l'aide d'un tableur mais de développer plutôt un module d'importation.

La première solution aurait permis de centraliser les opérations à l'intérieur d'un même logiciel, c'est-à-dire fournir à l'utilisateur un environnement de travail complet pour effectuer la saisie et l'analyse de données. Par contre, elle nécessitait un effort de programmation considérable pour obtenir les fonctionnalités normalement attendues d'un programme de saisie. Sachant que les chercheurs du laboratoire de Francine M. Mayer utilisent présentement le tableur Excel pour faire la saisie des données provenant des recensements, il n'aurait pas été acceptable de perdre les options que permet Excel. Il aurait donc fallu développer toutes les fonctionnalités de manipulation de données telles que les fonctions de copie, collage, de répétition d'une même donnée sur une colonne entière ou de saisie automatique afin de ne pas dévaluer le processus de saisie actuel.

La solution choisie a été de développer un module d'importation à même Analypop, qui permet d'importer dans une table de la base de données un fichier préparé à l'aide d'un tableur. Le principal avantage de cette méthode, outre le fait que l'effort de programmation soit moins grand, est que l'on tire profit des fonctionnalités déjà offertes par le tableur Excel et que l'on reste très près du mode de saisie présentement utilisé.

2.2.2 Format CSV

Le format *CSV* pour "*Comma Separated Values*" est un format de fichier texte utilisé pour représenter des données que l'on représenterait normalement sous un mode lignes-colonnes. Ce format consiste simplement à écrire les données d'une ligne les unes à la suite des autres séparées par un point-virgule. Ainsi, le programme qui lit un fichier de ce type peut facilement retrouver le format d'origine.

Exemple 2.2.1 Fichier CSV contenant les données de la table INDIVIDU présenté à l'exemple 2.1.4 :

INDIVIDU

Individu	Prénom	Nom	Date de naissance	Issu de
1	John	Coltrane	1926-23-10	100
2	Ella	Fitzgerald	1917-25-04	101
3	Lester	Coltrane	1902-13-05	102
4	Emma	Young	1904-22-11	105
5	Coleman	Fitzgerald	1897-10-05	124
6	Lucy	Hawkins	1899-04-12	132

```
1 ;John ;Coltrane ;1926-23-10 ;100
2 ;Ella ;Fitzgerald ;1917-25-04 ;101
3 ;Lester ;Coltrane ;1902-13-05 ;102
4 ;Emma ;Young ;1904-22-11 ;105
5 ;Coleman ;Fitzgerald ;1897-10-05 ;124
6 ;Lucy ;Hawkins ;1899-04-12 ;132
```

2.2.3 Importation à l'aide d'Analypop

Comme mentionné à la Section 2.2.1, la première étape pour importer des données dans la base de données d'Analypop consiste à entrer les données dans un tableur tel qu'Excel et à exporter le fichier dans le format CSV. Ensuite, si ces données doivent être contenues dans une nouvelle table, nous devons créer cette table avant de pouvoir procéder à l'étape suivante. Il existe plusieurs logiciels qui permettent de se connecter à une base de données et de créer des tables ; celui que nous avons utilisé est le logiciel proposé par MySQL : MySQL Admininistrator. Ce logiciel est gratuit pour les usages académiques et non commerciaux, ce qui n'implique pas des frais supplémentaires pour les utilisateurs d'Analypop. Par contre, dans le cas où les données que l'on veut importer servent à enrichir les données d'une table déjà existantes, il n'est pas nécessaire de créer de nouvelle table, les données seront ajoutées à une table déjà existante de la base de

données.

Donc, une fois que le fichier est créé et que la table existe, nous pouvons utiliser l'interface d'Analypop prévue à cet effet. La Figure 2.5 montre l'interface qui sert à faire l'importation. La procédure pour importer des données consiste tout d'abord à sélectionner le fichier CSV et sélectionner la table dans laquelle on veut envoyer les données. Une fois cette sélection faite, c'est en choisissant dans chacune des listes déroulantes dans le haut des colonnes que l'on indique à Analypop dans quelles colonnes de la table insérer les données (Fig. 2.6).

Afin de distinguer les différentes lignes d'information d'une table de la base de données, on utilise une clé primaire. La *clé primaire* est généralement un nombre unique qui doit permettre d'identifier toutes les lignes d'une ou plusieurs table qui ont de l'information complémentaire. Par exemple, en associant un numéro de matricule unique (IDI) à un individu on peut faire le lien entre la table INDIVIDU et CONJOINT lorsqu'on utilise ce même matricule pour inscrire le numéro de l'homme qui compose une union. Si on n'avait pas ce matricule unique, il faudrait, à chaque fois que l'on veut spécifier à qui se rattache l'information, répéter le nom, prénom, date de naissance, bref, toutes les informations qui permettraient d'identifier un seul individu. De plus, la clé primaire permet une plus grande efficacité lors des recherches puisqu'on n'a qu'à faire une seule comparaison pour comparer deux éléments de la base de données.

Lors de l'importation, il faut donc tenir compte de la clé primaire. Dans le cas où elle est incluse dans le fichier CSV, Analypop détecte sa présence grâce à la liste déroulante en en-tête qui indique le nom de la clé primaire et, dans le cas où il n'y a pas de clé primaire (ou de matricule), comme à la Figure 2.6, Analypop génère un nombre qui n'est pas encore utilisé. C'est pour cette raison que l'utilisateur peut, au bas de l'écran, spécifier un intervalle de valeurs à utiliser lors de l'importation. Ainsi, comme nous l'avons mentionné à la Section 2.1.4, l'utilisateur peut créer des catégories d'individus dépendant de la provenance ou de la signification qu'il veut donner aux données importées.

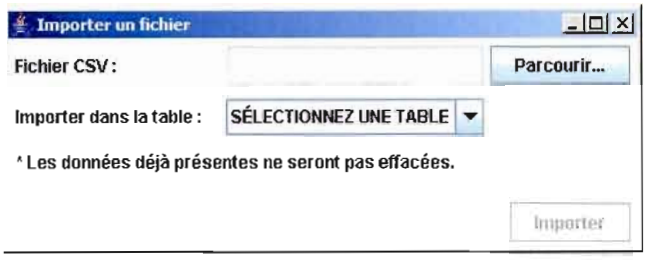

Fig. 2.5 Sélection du fichier de données à importer et choix de la table.

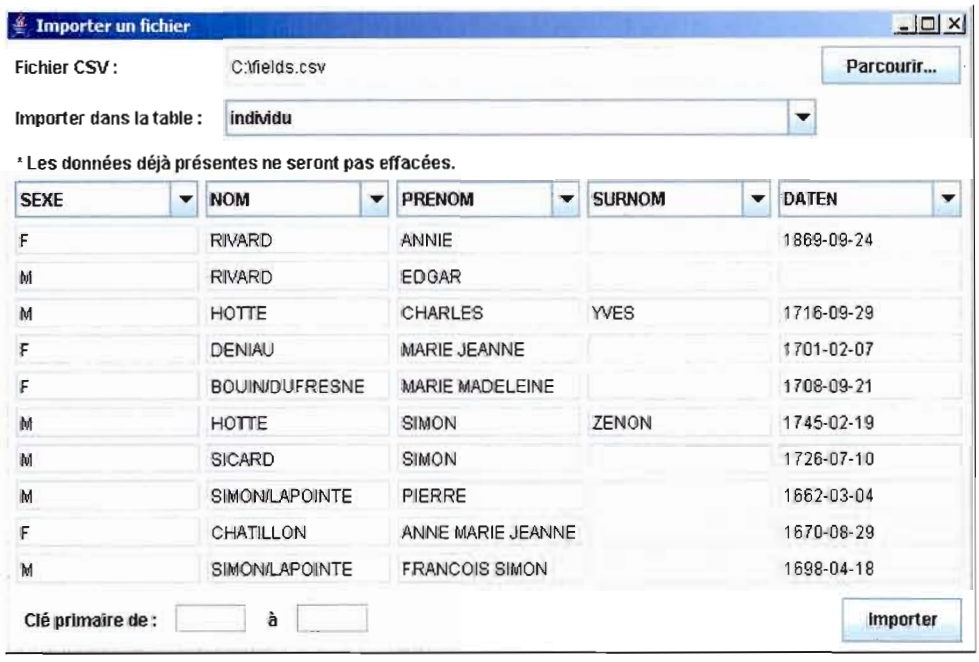

Fig. 2.6 Sélection des colonnes de la table où stocker les données.

CHAPITRE III

GESTION DES JUMELAGES ET DE L'HISTORIQUE DES DÉCISIONS

Un des principaux objectifs de la réingéniérie du logiciel Analypop était de développer un module d'assistance au jumelage afin d'en faciliter la gestion. Il était donc important de pouvoir conserver les décisions de jumelage et de pouvoir, dans le cas où l'on observerait une erreur, revenir à l'état antérieur au jumelage mis en doute. Ce chapitre présente tout d'abord la méthode utilisée pour définir un jumelage et ensuite présente la méthode qui a été développée afin de jumeler l'information et en garder un historique.

3.1 Problèmes liés au jumelage

Le jumelage est un processus qui fait intervenir l'expertise des chercheurs qui se basent sur les divers actes d'état civil pour construire le registre de population. Il n'existe pas de méthode simple qui puisse être appliquée pour assurer un succès à tout coup. Les chercheurs doivent donc employer des astuces particulières pour élucider les évènement qui unissent les individus nommés dans ces actes. Les raisons pouvant causer problèmes lors d'un jumelage peuvent être abondantes (9), nous dresserons néanmoins une liste de celles les plus couramment rencontrées.

- Les registres religieux ou laïcs contenant les divers actes d'état civil sont perdus, endommagés ou incomplets. Donc une partie de l'information nécessaire pour faire le jumelage est manquante.

- Il existe des couples homonymes, c'est-à-dire des couples dont le nom des deux hommes est le même et le nom des deux femmes est le même. Ceci s'explique par la structure isolée d'une population et l'endogamie (où les mariages se font à l'intérieur de la même population).
- Selon le type de peuplement et les effectifs impliqués, les fondateurs peuvent constituer un groupe relativement restreint de sorte qu'il n'y a pas une grande diversité de patronymes.
- Comme le dépouillement se concentre sur une population d'un lieu défini (on ne dépouille pas les registres à l'extérieur du lieu concerné), les actes de batpême, de mariage et/ou de décès des immigrants sont souvents absents puisque l'évènement s'est produit à l'extérieur de la population étudiée.
- À l'époque, il n'était pas rare qu'une personne change de nom durant sa vie, d'où le présence de surnoms. Il faut donc connaître les équivalences entre les noms et les surnoms.

Certains programmes de gestion de registre de population, dont le logiciel du PRDH, utilisent certaines *heuristiques*, c'est-à-dire un ensemble de règles qui donnent généralement la bonne solution, combinées à la décomposition des noms en code phonétique pour tenter d'automatiser le processus du jumelage (16). Cette méthode implique le calcul d'un score de similitude entre deux individus. Ce score se base généralement sur une transformation phonétique des noms ainsi qu'une table d'équivalence de ceux-ci (23), (18), (19). Lorsque le score dépasse un certain seuil, le programme prend la décision de jumeler les deux individus automatiquement. Cette approche semble être efficace pour traiter un grand nombre d'individus si on accepte qu'elle produise un certain nombre de jumelages fictifs. Dans notre cas, étant donné la relative petite taille des populations qui sont étudiées, l'automatisation des jumelages aurait allourdi le processus sans augmenter significativement la performance du jumelage. Il a donc été décidé, dès le départ, que le module de jumelage ne serait pas un module de jumelage automatisé mais plutôt un outil qui permettrait de gérer l'information et de prendre des décisions, d'où le choix du terme : module d'assistance au jumelage.

3.2 Méthode de jumelage

La source d'information qui est privilégiée pour effectuer les jumelages est l'acte de mariage puisque sur celui-ci on y retrouve normalement la mention des deux individus qui se marient ainsi que celle de leurs parents respectifs. Grâce à cette seule source, on arrive à mettre en relation six individus (trois couples). Le lien générationnel est donc saisi et permet de remonter les ascendances. Étant donné qu'un registre de population est constitué dans le but de suivre l'évolution des structures démographiques et généalogiques, il se doit d'inclure l'information sur les d'individus qu'on n'identifie pas avec les actes de mariages mais qui se retrouvent dans les autres actes. Par exemple, il se peut très bien qu'un individu né dans la population étudiée ait émigré avant son mariage ou il se peut qu'un couple ait immigré dans la population, après s'être mariés à l'extérieur, et ait eu une descendance, qui soit restée sur place. Dans ces deux cas, on ne retrouvera pas l'individu si on concentre le dépouillement seulement sur les actes de mariage. C'est pourquoi, une fois le dépouillement de ceux-ci effectué, on doit dépouiller les actes de baptême et les actes de décès qui donnent le nom des parents ou di conjoint. Le rapprochement de toutes les mentions similaires de coupl dans tous les actes dépouillés permet le jumelage des individus.

Donc, une fois le dépouillement effectué, on doit être en mesure de mettre en relation les individus nommés dans les actes et susceptibles d'être jumelés. La méthode actuelle pour faire les jumelages consiste à comparer différentes listes classées selon un critère pré-définit. Par exemple, si le chercheur dépouille un acte de baptême pour l'individu Richard Gagnon né en 1843 dont les parents mentionnés sont Guy Gagnon et Berthe Piché, il imprimera la liste des unions dont le nom de l'homme est Gagnon et la femme est Piché. Ensuite, en comparant les prénoms et les dates de mariage, il arrivera à déterminer un sous-ensemble de couples mariés dans la population ayant pu donner naissance à un enfant né en 1843. Il existe certaines heuristiques que l'on applique afin de déterminer s'il est possible que le couple mentionné dans l'acte de mariage soit le même que celui qui a engendré cet enfant. Les critères à vérifier sont :

- L'âge de la mère doit être supérieur à 15 ans et inférieur à 50 alors que l'âge du père peut être compris entre 15 et 80 ans.
- Une mère peut décéder au plus 85 ans après la naissance de l'enfant alors que le père peut décéder 9 mois avant jusqu'à 85 ans après.
- L'intervalle entre 2 naissances doit être supérieur à 9 mois.
- La différence d'âge entre le père et la mère ne peut excéder 35 ans.

Par exemple, lorsque que nous avons identifié un couple marié qui satisfait à tous ces critères, on peut effectuer le jumelage. Dans ce cas-ci, il faudrait assigner le numéro d'union (ID) du couple marié provenant de la table CONJOINT à la colonne ISSU de la table INDIVIDU pour l'individu Richard Gagnon. Dans le cas où on identifierait un jumelage qui enfreindrait certaines règles, on aurait quand même la possibilité de définir le jumelage après avoir été averti par le programme.

En s'inspirant de cette méthode, nous avons développé, pour Analypop, un module de définition de types de listes. La Figure 3.1 présente l'écran qui permet à l'utilisateur de choisir les champs qu'il désire afficher en choisissant la table à laquelle le champ appartient et le champ voulu. De plus, en ordonnant les champs dans la partie centrale de cet écran, il peut décider dans quel ordre les champs seront triés lors de l'affichage. La partie au bas de cet écran permet de définir le ou les critères qui permettent de faire le lien entre les tables choisies dans le cas où les champs proviennent de plus d'une table. La plupart du temps, on peut définir un lien, appelée *jointure*, entre deux tables en mentionnant le nom des champs qui doivent être égaux. Dans le cas de la Figure 3.1, on voulait obtenir une liste des noms et prénoms de tous les individus mariés dans la paroisse avec la date de mariage. Sachant que l'individu est identifié de façon unique dans la vue V_INDIVIDU par le champ ID et que lors d'une union, son matricule ID se retrouve dans la vue V_CONJOINT dans la colonne HOMME ou FEMME, on peut faire la jointure entre les deux tables avec la commande suivante :

```
v_individu.id = v_conjoint.homme OR v_individu.id = v_conjoint.femme
```

La syntaxe de cette commande est définie par le langage SQL.

Fig. 3.1 Écran de définition de liste.

Une fois que la liste a été définie, nous pouvons la faire afficher à l'écran principal en sélectionnant, dans le menu Options → Types de listes, le nom associé à la nouvelle liste. La Figure 3.2 présente l'écran principal une fois que nous avons sélectionné la liste des mariages des individus.

Pour faciliter le travail d'identification des individus, nous avons prévu un espace au haut de chaque colonne, qui permet de filtrer l'information d'une colonne, c'est-à-dire qui permet de restreindre le nombre de lignes affichées à l'écran. Pour ce faire, nous avons défini quelques opérateurs qui permettent de facilement créer ce filtre. Il y a cinq opérateurs pouvant servir à comparer des données de types numérique ou date : $<$, $<=$, $=$, $>=$, $>$. Ces opérateurs ont la signification qui leur est normalement attribuée. Par contre, nous avons défini un autre opérateur permettant de comparer des chaînes de caractères. L'opérateur *like* est fortement inspiré de l'opérateur du même nom de SQL et permet de comparer des chaînes de caractères selon un certain patron. Ce patron

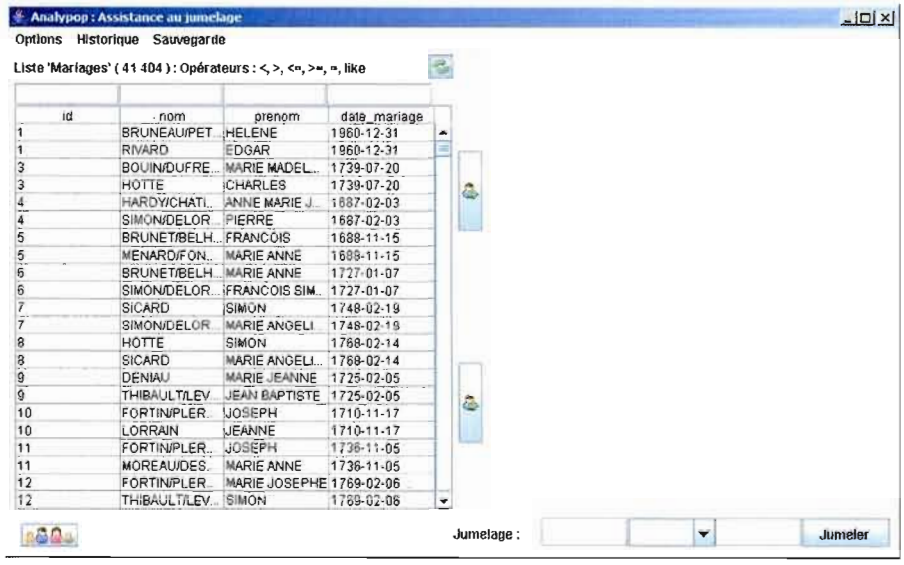

Fig. 3.2 Écran principal du logiciel Analypop.

peut inclure les caractères spéciaux "%" ou "_" pour tester une égalité partielle entre deux chaînes. En plus de ces opérateurs nous avons prévu le cas où l'utilisateur voudrait combiner certains opérateurs et avons ajouté les opérateurs logiques "AND" et "OR".

L'écran présenté à la Figure 3.3 se veut le pivot de l'application Analypop. C'est à partir de cet écran que l'on peut accéder à toutes les autres fonctionnalités du logiciel utiles au processus du jumelage d'individus. Comme nous l'avons mentionné à la Section 3.2, lors du jumelage, il est important de pouvoir comparer l'information concernant plus d'un individu afin de les discriminer. C'est pour cette raison que l'écran principal (Fig. 3.3) est divisé en deux parties. La première permet de présenter une liste quelconque et la deuxième permet de consulter l'information concernant deux individus en particulier, ce qui facilite la prise de décision pour effectuer un jumelage ou non. Pour chaque individu on affiche l'information le concernant tel que définie dans le fichier de propriétés (Section 2.1.9).

Nous avons vu, dans l'introduction, que les fiches de familles regroupent les informations relatives à un individu, son(sa) conjoint(e), ainsi que les enfants de ce couple.

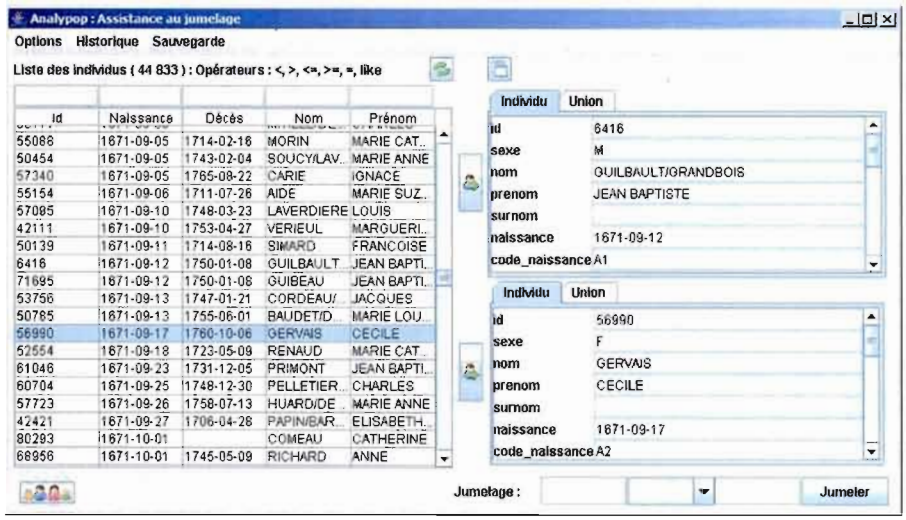

Fig. 3.3 Écran principal du logiciel Analypop où on affiche les informations concernant deux individus.

Lorsqu'on tente de jumeler des individus, ces fiches peuvent être très utiles car elles permettent d'avoir une vision plus générale des individus que l'on traite. Par exemple, si on s'aperçoit qu'un couple a déjà 16 enfants, on hésitera peut-être un peu avant de lui en associer un autre. De plus, si on on tente de départager deux couples homonymes (les hommes ont le même patronyme et les femmes ont le même patronyme), il se peut que la consultation de l'information concernant les parents de ces individus, nous aide à les discriminer. C'est pour cette raison que nous avons implanté une fonctionnalité qui permet de consulter la fiche de famille d'un couple. La Figure 3.4 présente l'interface de la fiche de famille du couple composé de Zacharie Perron et Catherine Audet.

En plus de présenter les deux membres du couple, *ego* et sa conjointe, ainsi que leurs enfants, nous avons ajouté la mention des parents d'ego et de sa conjointe pour permettre aux utilisateurs de pouvoir avoir une vue longitudinale de l'information. Aussi, dans le cas où un individu aurait eu plus d'un(e) conjoint(e), deux boutons permettant de naviguer parmi ses différents(es) conjoints(es) sont affichés à la droite de la mention "Conjoint d'ego :". Lorsque l'utilisateur clique pour visionner un autre conjoint, la partie

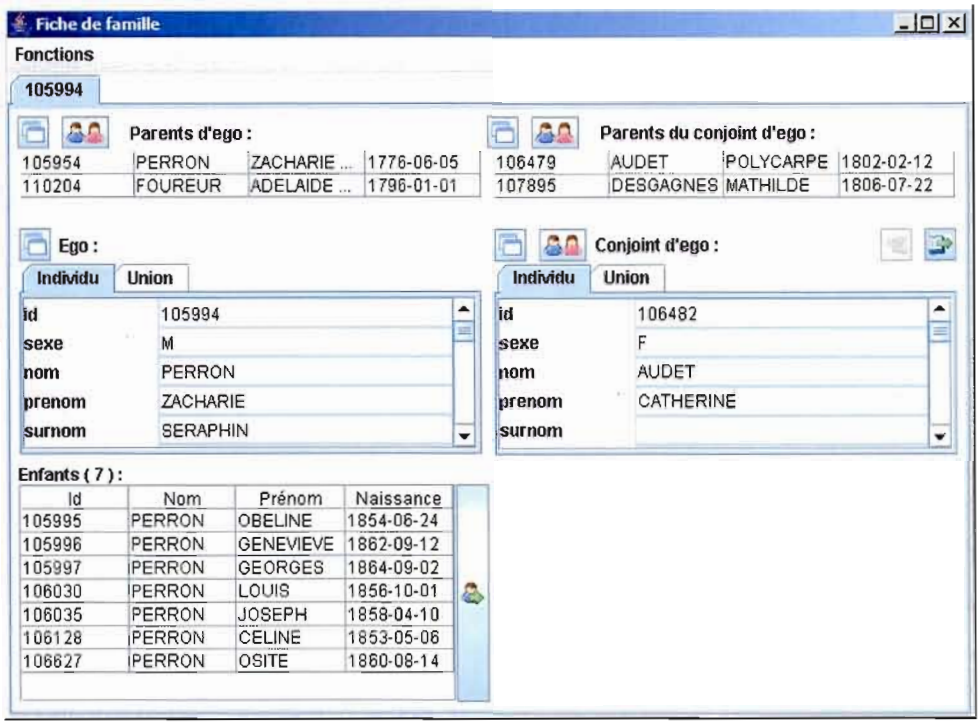

Fig. 3.4 Fiche de famille du couple Zacharie Perron et sa conjointe Catherine Audet.

de droite de cet écran change pour afficher les informations sur le nouveau conjoint et ses parents ainsi que la liste des enfants du couple. La Figure 3.5 présente la fiche de famille de ce même individu mais avec les informations de sa deuxième union.

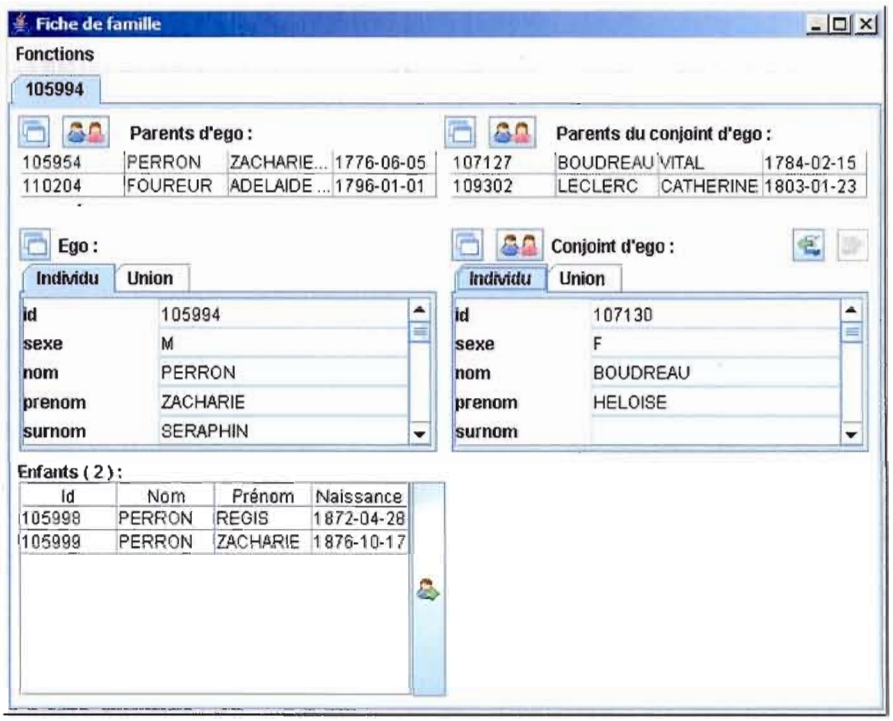

Fig. 3.5 Fiche de famille du couple Zacharie Perron et sa conjointe Héloise Boudreau.

Malgré le fait que l'information présentée dans cet écran soit assez complète, nous sommes conscient qu'il est un peu chargé. C'est pourquoi nous avons prévu la possibilité de consulter l'information sur un seul individu dans une fenêtre à part, afin d'avoir plus d'espace à l'écran. Pour ce faire, il suffit de cliquer sur le bouton affichant une fenêtre par dessus une autre. De plus, la fiche de famille de chacun des couples présents sur cette fiche est accessible lorsqu'on clique sur le bouton représentant un couple. Pour consulter les informations concernant un des enfants, il suffit, toujours à partir de cet écran, de sélectionner l'enfant et de cliquer sur le bouton à la droite de la liste des enfants. La Figure 3.6 présente l'état de la fiche de famille lorsqu'on a sélectionné l'enfant Céline Perron et affiché les informations la concernant.

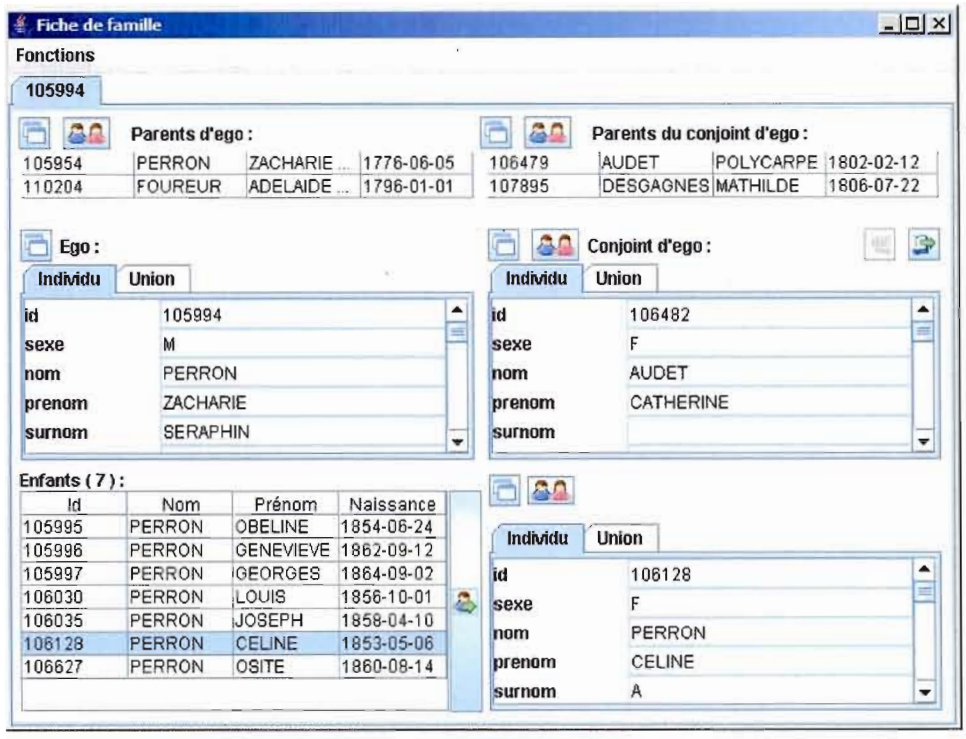

Fig. 3.6 Fiche de famille du couple Zacharie Perron et Catherine Audet où on affiche les informations concernant Céline Perron.

3.3 Types de jumelage

Comme nous l'avons mentionné dans l'introduction, l'action de jumelage permet de lier deux individus dans le registre de population par dépouillement d'actes d'état civil dans lesquels ils font objet de mention. Après avoir analysé les différents cas où les chercheurs avaient à faire un jumelage, nous avons identifié trois cas de base différents qui permettent définir tous les jumelages possibles. Ces trois types de jumelage sont le jumelage d'individus égaux (appelé "égalité" par Analypop), le jumelage de conjoints (appelé "conjoint de") et le jumelage d'un enfant avec ses parents (appelé "issu de").

Le jumelage d'individus égaux permet de réduire la redondance des individus qui sont inscrits à plus d'une reprise dans le registre de population sous un matricule différent.

Lors d'un jumelage de ce type, si les deux individus avaient de l'information complémentaire, elle est ajoutée afin d'enrichir l'information de l'individu résultant. Le jumelage de conjoints permet de définir une nouvelle ligne dans la table CONJOINT en associant deux individus déjà existants dans le registre. Par la suite, si on possède l'information suffisante, on peut associer des enfants à ce couple en ayant recours au jumelage de type "issu de". Ce type de jumelage inscrit dans le champ ISSU de la table INDIVIDU, le numéro de couple spécifié. Ce dernier type diffère des autres procédures de jumelage puisqu'il associe un numéro d'individu avec un numéro de couple, alors que les deux autres associent des numéros d'individus.

3.4 Exemple du processus de jumelage de type "égalité" effectué avec Analypop

Dans le but d'acquérir une vue d'ensemble de la procédure de jumelage et de juger de l'utilité des fonctionnalités présentement offertes par Analypop, nous verrons dans ce qui suit un exemple du processus de jumelage.

Tout d'abord, supposons que nous venons de dépouiller l'information concernant l'acte de décès de l'individu Jean Boivin et lui avons attribué le matricule 220001. Par la suite, dans le but de vérifier si cet individu existe déjà dans notre registre, nous effectuons une recherche à partir du nom Boivin et du prénom Jean, telle que le présente la Figure 3.7.

On peut remarquer, à la Figure 3.7, que le seul individu qui pourrait en fait être le même que celui que nous venons d'ajouter est l'individu au matricule 7482. En effet, sur les 8 individus présents, 3 ont déjà une date de décès enregistrée (68381, 52563 et 5040), un est une femme (68152), un est né 148 ans avant son décès (44581) et l'autre est né après la date de décès enregistrée (66548). Il ne reste donc que 2 individus qui seraient potentiellement le même. Si on affiche leurs informations, on remarque que pour l'individu dont on connaît sa date de naissance, son prénom ne correspond pas parfaitement à celui de Jean. Par contre, d'après les informations que nous avions, nous savions que l'acte de décès dépouillé comprend le surnom Jean Marie. Ceci pourrait

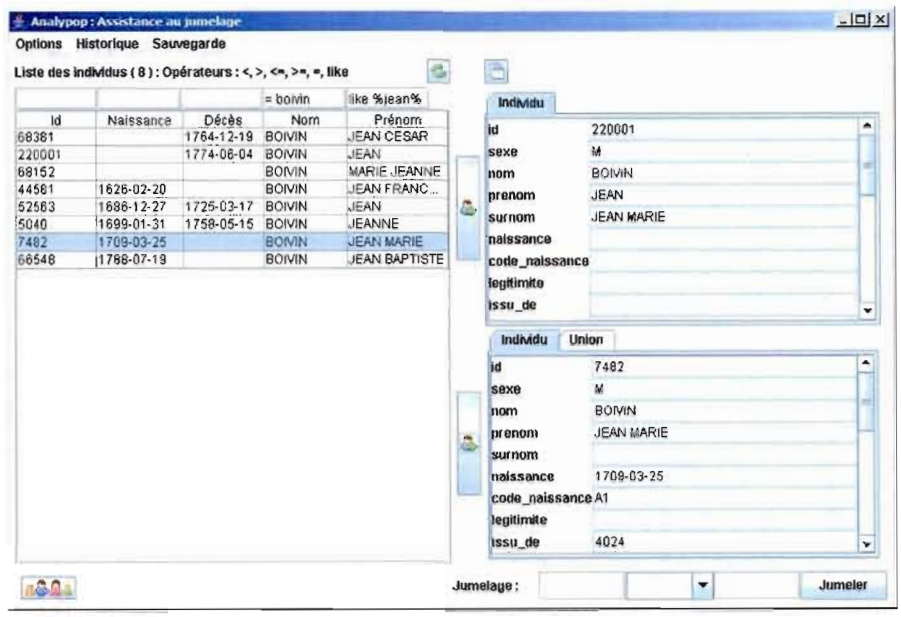

Fig. 3.7 Recherche d'individus s'appelant Jean Boivin.

permettre de penser qu'en fait, ces deux individus sont le même. Nous pourrions alors définir un jumelage de type égalité. Lorsqu'on définit que deux individus sont égaux, le logiciel demande alors à l'utilisateur de sélectionner quelles informations conserver avant d'enrichir l'individu résultant. La Figure 3.8 présente le premier écran qui sera affiché lorsqu'on décide de jumeler l'individu 220001 avec l'individu 7482 selon le type "égalité". Le logiciel présente un écran semblable pour chaque table de la base de données où les individus se retrouvent (configuré à l'aide du fichier de propriétés).

Finalement, lorsque tous les choix ont été faits, l'utilisateur peut entrer un commentaire afin d'identifier les raisons qui l'ont amené à faire ce jumelage. Ensuite, lorsque l'utilisateur décide de concrétiser le jumelage, en cliquant sur le bouton jumeler, l'individu est modifié pour prendre en compte des informations qu'il a sélectionné. On peut vérifier que les modifications ont réellement été faites si, à l'écran principal, on sélectionne l'individu et on affiche l'information retenue. La Figure 3.9 montre l'individu 7482, maintenant connu avec le nom et surnom Jean Marie, et dont la date de décès est 1774-06-04.

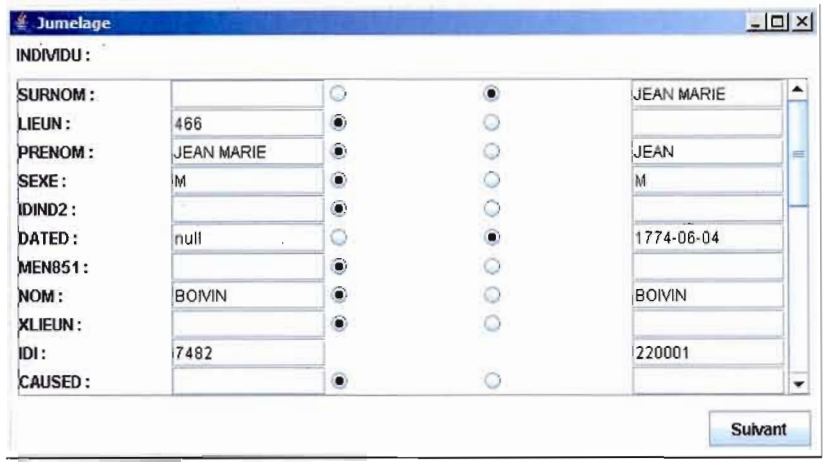

Fig. 3.8 Demande de sélection des informations à conserver.

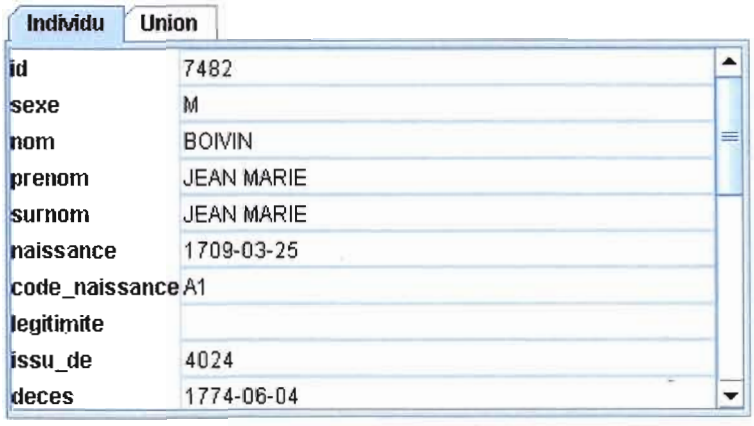

Fig. 3.9 Détail de l'information concernant l'individu 7482, Jean Marie Boivin.

3.5 Consultation de l'historique

Comme nous l'avons déjà mentionné à la Section 2.1.8, il est important de conserver un historique des décisions de jumelage. Il arrive parfois que le chercheur obtienne de nouvelles informations et s'apperçoive qu'un jumelage qu'il a jugé valide auparavant, ne l'est plus suite à la connaissance d'informations supplémentaires. Puisque le registre de population est l'outil sur lequel se basent toutes les analyses, son état doit toujours être le plus près possible de la réalité. Il fallait donc prévoir une fonction qui permettrait

aux utilisateurs, le cas échéant, de récupérer des informations non conservées ou avant qu'elles soient combinées lors du jumelage. Nous avons donc développé une interface permettant de récupérer les informations telles qu'elles se présentaient avant le jumelage. La Figure 3.10 présente l'interface qui permet aux utilisateurs d'accomplir cette tâche.

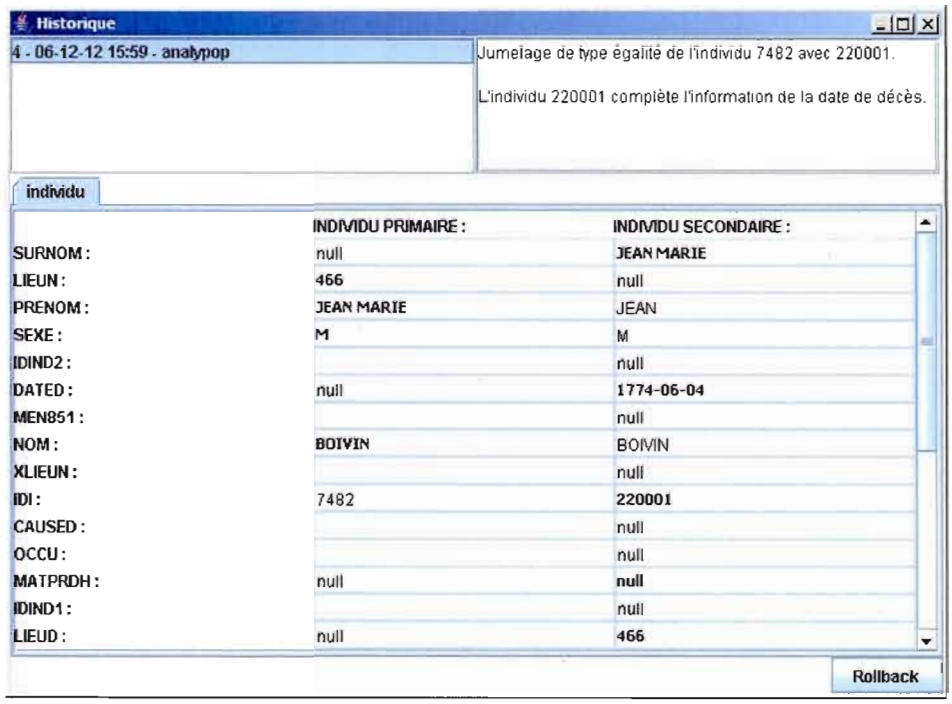

Fig. 3.10 Écran permettant de consulter l'historique.

Tout d'abord, on peut remarquer que l'utilisateur doit sélectionner un jumelage parmi la liste qui lui est présentée au haut de l'écran. Pour faciliter la tâche d'identification d'un jumelage en particulier, lorsqu'il sélectionne un élément de la liste, le commentaire qui a été associé au jumelage s'affiche à droite de celle-ci. Lorsque l'utilisateur s'est assuré d'avoir sélectionné le jumelage voulu, il peut consulter les informations telles qu'elles étaient avant le jumelage. En effet, la partie au bas de l'écran montre les informations concernant l'individu *primaire*, individu de base dont on a inscrit le matricule à gauche de la liste de sélection des types de jumelage, et l'individu *secondaire*, celui qui est venu compléter l'information de l'individu primaire. Il est à noter que nous avons inscrit en

gras l'information qui a été retenue à l'étape de la sélection (Fig. 3.8).

Si l'utilisateur décide d'annuler ce jumelage, il doit cliquer sur le bouton Rollback au bas de l'écran. Ceci a pour effet de modifier les individus afin de leur redonner les informations qu'ils possédaient avant le jumelage. On peut constater, à partir de l'écran principal, l'effet qu'a cette opération à la Figure 3.11.

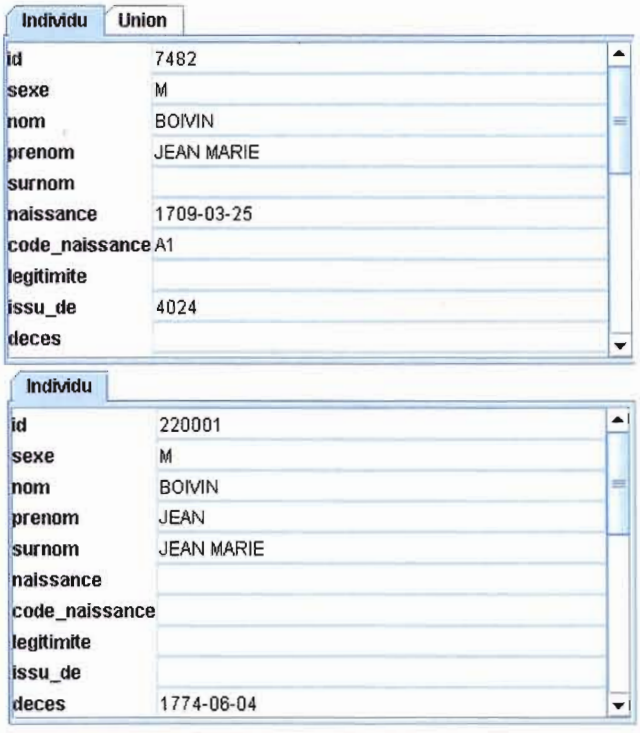

Fig. 3.11 Présentation des deux individus tels qu'ils étaient avant le jumelage.

3.6 Exemple du processus de jumelage de type "conjoint de" effectué avec Analypop

Comme nous l'avons vus précédemment, Analypop permet de définir trois types de jumelage. Nous venons de voir un exemple du processus de jumelage pour définir un jumelage par "égalité". La présente section présentera un deuxième exemple, cette fois-ci d'un processus de jumelage de type "conjoint de".

Supposons que nous venons de dépouiller, d'un registre paroissial, l'acte de mariage d'Angélique Lavoye et de Jean Marie Boivin. Sur cet acte, en date du 12 avril 1734, nous avons en plus l'information concernant les dates de naissances des deux époux : Angélique Lavoye est née en le 10 juillet 1716 et Jean Marie Boivin est né le 25 mars 1709. Ceci nous permet de vérifier si ces deux individus existent déjà dans le registre de population informatisé. Étant donné que nous connaissons déjà l'existence de Jean Marie Boivin (7482) dans le registre de population, nous allons rechercher, à l'aide de l'écran principal, la présence d'un individu qui se nomme Angélique Lavoye (Fig. 3.12).

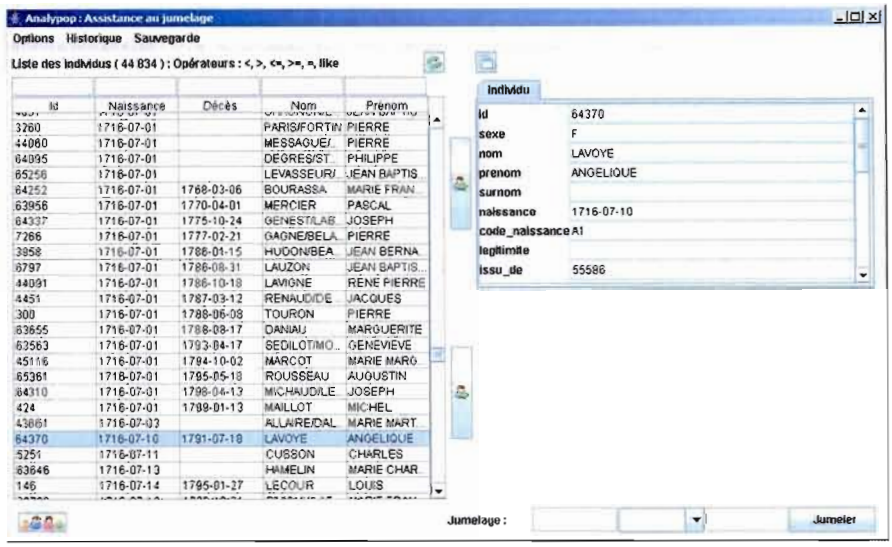

Fig. 3.12 Présentation des information de l'individu Angélique Lavoye trouvé dans le registre de population.

Par la suite, étant donné que nous avons effectivement déjà ces deux individus, nous consultons leur fiche de famille afin de détecter toute incohérence possible (Fig. 3.13). Par exemple, si l'individu Angélique Lavoye du registre de population est déjà mariée et nous avions l'information qu'elle a eu des enfants pendant la période couvrant l'année 1709, on peut juger qu'il est très improbable qu'elle se soit mariée à cette époque. Nous serions donc en présence d'un deuxième individu nommé Angélique Lavoye et il faudrait créer un nouvel individus dans notre registre informatisé.

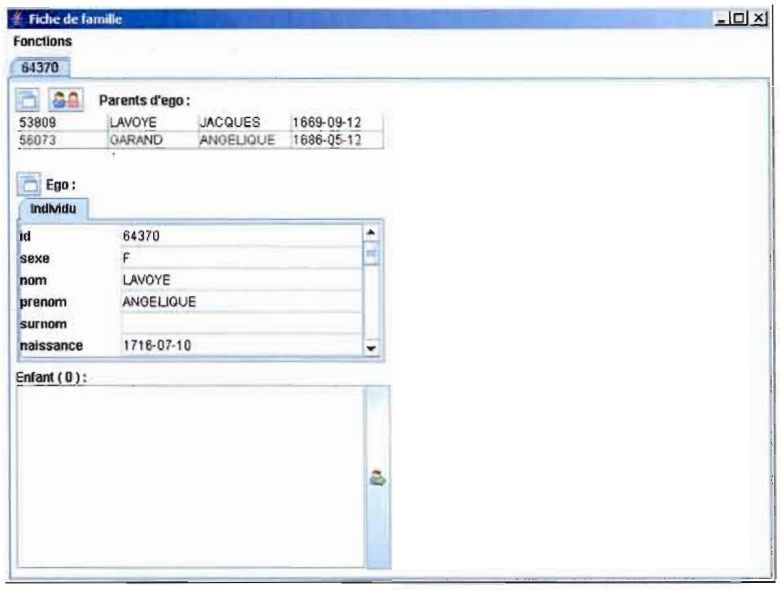

Fig. 3.13 Présentation de la fiche de famille d'Angélique Lavoye (64370).

Lorsque nous avons consulté les fiches de familles des deux individus impliqués par le jumelage et que nous nous sommes assurés que l'information contenue dans le registre de population informatisé est cohérente avec le jumelage que nous désirons effectuer, nous pouvons, à partir de l'écran principal, définir un jumelage de type "Conjoint de"ẽntre ces deux individus (Fig. 3.14).

Par la suite, Analypop demande d'entrer les informations relatives à l'union de ce couple (Fig. 3.15). Dans notre cas, nous n'avions que l'information concernant la date de mariage (1734-04-12). Veuillez noter que le numéro d'union (ID) est un numéro séquentiel généré automatiquement par le programme mais qu'il demeure possible pour l'utilisateur de le modifier pour un numéro qui n'est pas déjà utilisé.

Lorsque l'utilisateur a terminé d'entrer les informations relatives à l'union et qu'il appuie sur le bouton Suivant, une autre fenêtre apparaît pour lui permettre d'écrire des commentaires à propos du contexte qui a fait en sorte qu'il prenne la décision de jumeler les deux individus (Fig. 3.16).

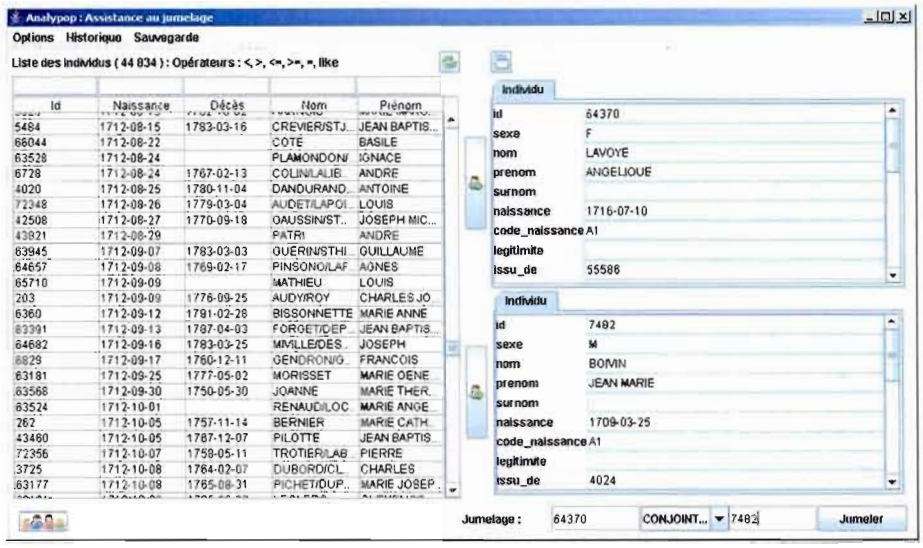

Fig. 3.14 Définition d'un jumelage de type "Conjoint de" entre les individus Angélique Lavoye (64370) et Jean Marie Boivin (7482).

Ensuite, lorsqu'il appuie sur le bouton Suivant, un message l'informant si le jumelage a été correctement effectué ou non est affiché. Par la suite, l'utilisateur est ramené à l'écran principal d'où il peut s'assurer du bon fonctionnement du jumelage. Dans notre cas, nous vérifions la fiche de famille d'Angélique Lavoye afin de s'assurer qu'elle est belle et bien unie à Jean Marie Boivin (Fig. 3.17).

Finalement, ce chapitre nous a permis de mieux comprendre les difficultés reliées au jumelage d'individus ainsi que les méthodes généralement utilisées afin d'en faire leur gestion. Malgré le grand nombre de types de liens qui peuvent unir deux individus de la même famille, nous avons vu qu'un sous-ensemble restreint de trois types de relations permet de représenter toute la complexité des liens qui peuvent exister dans une population donnée. Par ailleurs, deux exemples de jumelage nous ont permis de voir comment ce processus a été intégré au logiciel. Analypop permet non seulement de définir des jumelages mais fait une gestion de leur historique qui, dans le cas échéant, permet d'annuler une décision de jumelage et de revenir l'état antérieur à celui-ci.

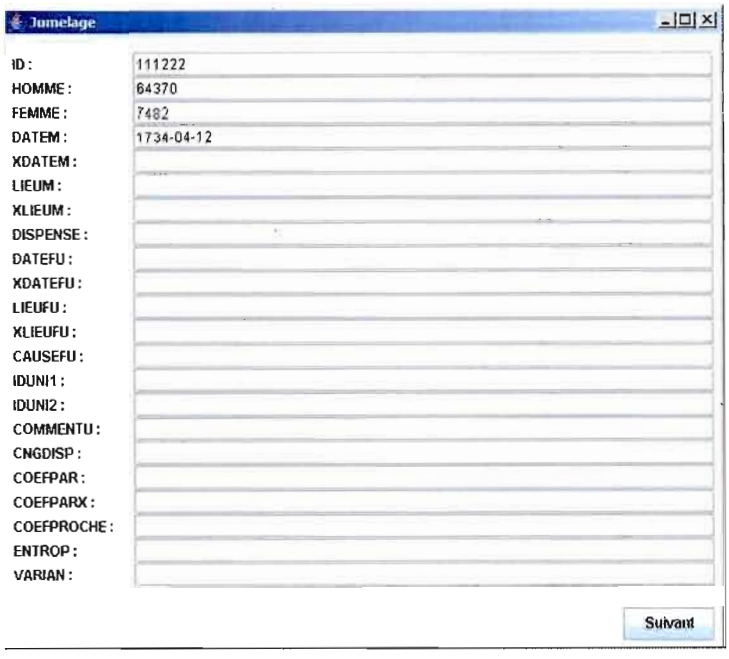

Fig. 3.15 Écran où l'on doit entrer les information que l'on connaît à propos l'union qui est créée.

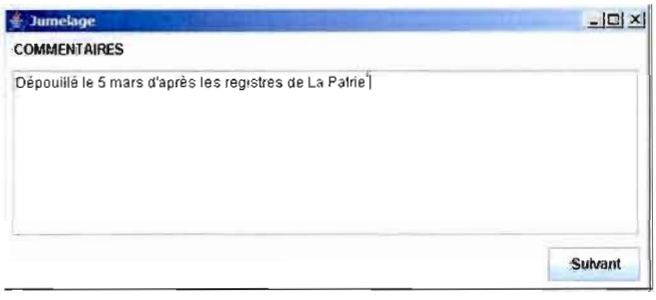

Fig. 3.16 Écran où l'utilisateur peut noter le contexte qui lui a permit de prendre la décision de jumelage.

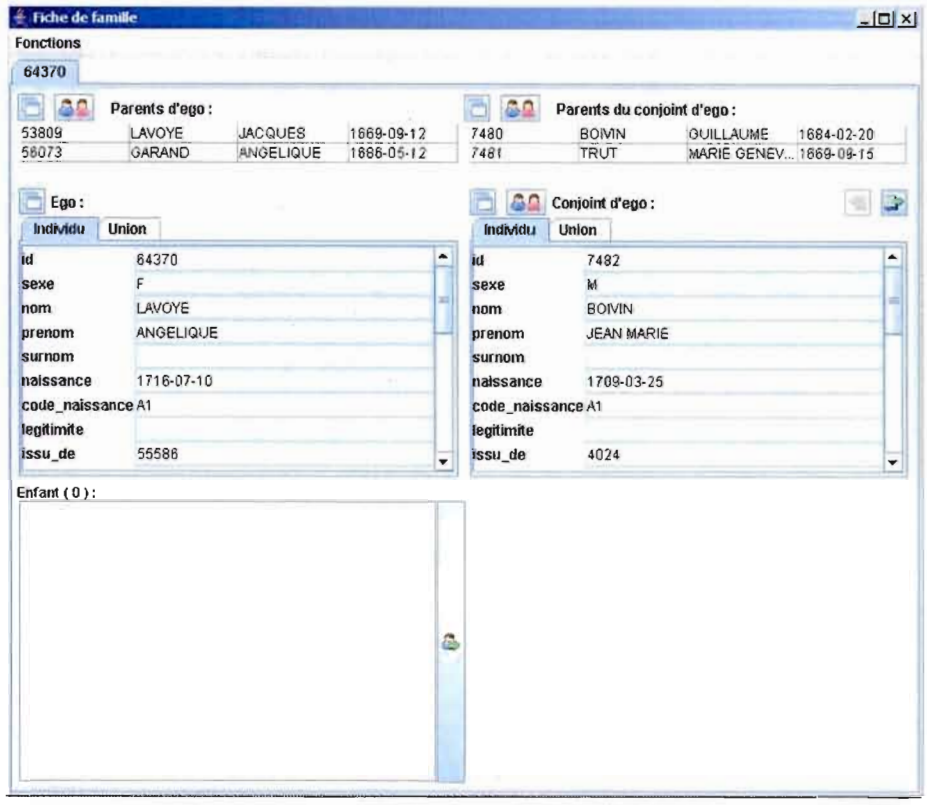

Fig. 3.17 Fiche de famille d'Angélique Lavoye, mariée à Jean Marie Boivin.

CHAPITRE IV

FICHIERS EXTERNES GEDCOM ET KML

Dans le but de rendre les données du logiciel Analypop encore plus accessibles, nous avons prévu des fonctionnalités d'exportation des données pour qu'elles puissent être utilisées avec d'autres programmes. Dans ce chapitre, nous décrirons les deux formats de fichiers qu'il est présentement possible d'obtenir à partir du logiciel Analypop. La première section abordera le format de fichier GEDCOM, qui nous permet de visionner une généalogie complète grâce à des programmes spécialement conçus à cet effet, alors que la section suivante traitera du format de fichier KML, qui permet de visionner des données de types géographiques avec le logiciel Google Earth.

4.1 Fichiers GEDCOM

Il existe, sur le marché, une multitude de programmes traitant des informations généalogiques. Malgré le fait qu'ils soient trop limités pour gérer ou effectuer des analyses sur un groupe important d'individus, ils offrent quand même des fonctionnalités intéressantes si on traite un petit nombre d'individus. La fonctionnalité qui, pour l'instant, nous intéresse le plus, est la possibilité de visionner une généalogie avec un support graphique. Étant donné qu'il aurait été plutôt long de développer, pour Analypop, notre propre module de visionnement de généalogies, nous avons préféré tirer profit de l'existence de ces programmes, sachant que plusieurs d'entre-eux sont disponibles gratuitement.

Pour ce faire, il a fallu développer une fonction d'exportation des données sous le format

GEDCOM (3). Le format de fichier GEDCOM, acronyme pour GEnealogy Data COMmunication est un format de fichier texte utilisé par la plupart des applications traitant de données généalogiques. Il fut développé par les mormons afin de créer un format de fichier standardisé permettant d'échanger des données entre plusieurs applications. La spécification GEDCOM 5.5 spécifie la syntaxe des étiquettes qui peuvent être utilisées pour décrire les individus ainsi que les évènements reliant ces individus. Les fichiers GEDCOM sont composés de 3 sections, une section en-tête, une section où sont écrits les champs et une section de fin de fichier. Chaque ligne d'un fichier GEDCOM débute par un numéro indiquant le niveau de l'étiquette. Ce numéro de niveau indique à quelle étiquette s'applique l'étiquette courante. La section en-tête comporte les informations relatives au fichier : sa version, son ou ses auteur(s), sa date de création etc. La section des champs contient l'information qui décrit les individus (numéro d'individu, nom, prénom, date de naissance, occupation...) ainsi que certaines étiquettes spéciales qui permettent de relier cet individu à d'autres individus présents dans le fichier (l'étiquette FAM permet de spécifier de quelle famille fait partie cet individu, l'étiquette HUSB permet de spécifier que l'individu est le père de famille, l'étiquette MARR permet de décrire un évènement de mariage entre 2 individus). Finalement, la section de fin de fichier est en fait une ligne qui indique la fin de fichier (0 TRLR). Veuillez vous référer à l'Annexe A.

Quoique utilisée par la majorité des applications de généalogies, la spécification GEDCOM 5.5 n'est pas respectée telle quelle par tous (4). Chaque application a plutôt étendu et adapté la spécification selon ses propres besoins. C'est pour faciliter l'échange d'information entre les différents programmes que la spécification GEDCOM 6 XML a été proposée. Cette nouvelle spécification suggère l'utilisation de XML (Extensible Markup Language), un langage permettant d'échanger de l'information de manière structurée. D'ailleurs, XML est déjà utilisé par plusieurs autres domaines scientifiques qui ont fait face aussi à la difficulté d'échange d'information (7) Malgré tout, étant donné que, dans le cas qui nous intéresse, l'information que nous voulons exporter est plutôt standard (nom, prénom, date de naissance, date de décès...) il est possible d'échanger de l'infor-

mation en utilisant ce format sans risquer de perdre de l'information.

À l'intérieur d'Analypop, la fonction d'exportation de généalogies est accessible à partir de l'écran de fiche de famille. Lorsqu'on sélectionne cette fonction, à partir du menu Fonctions → Exporter en format GEDCOM, l'utilisateur doit choisir un emplacement où sauvegarder le fichier. Ensuite, il peut visionner ce fichier à l'aide d'un logiciel tel que Le Généalogiste. On présente, à la Figure 4.1, l'ascendance de l'individu Monique Leclerc (33294).

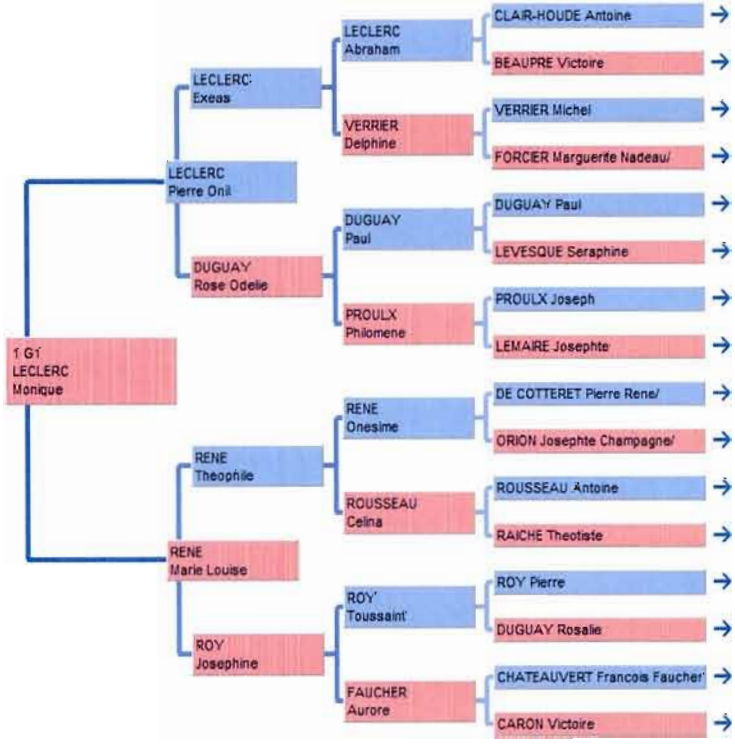

Fig. 4.1 Ascendance de l'individu Monique Leclerc telle que présentée par le logiciel Le Généalogiste.

4.2 Fichier KML

La représentation de données démographiques à l'aide d'un système d'information géographique est déjà assez répandu dans le domaine des sciences sociales. Ces systèmes permettent entre autre de mieux comprendre les interactions de structures spatiales. D'ailleurs de tels systèmes sont implantés par plusieurs pays tels que le Canada, les États-Unis, la France, le Japon et le Royaume-Uni afin de visualiser les données provenant de leur recensement respectifs (15). Le format de fichier KML (Keyhole Markup Language) est le format utilisé par Google pour représenter des informations de type géographiques. Ce langage permet de représenter des points, des lignes et des polygones. Dans le cas qui nous intéresse, nous utilisons ce type de fichier afin de représenter des points sur une carte géographique.

La syntaxe des fichiers de ce type est relativement simple, si l'on connaît les balises de base. Tous les fichiers KML doivent commencer par les balises suivantes :

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http ://earth.google.com/kml/2.0">
```

La première balise permet de définir que le format de fichier est de type XML (Extensible Markup Language) et la deuxième donne l'information où trouver les règles de syntaxe pour valider le document. En fait, KML est un type particulier de fichier XML. Ce qui vient ensuite permet de définir des points sur une carte. Ces points sont définis par un nom, une longitude, une latitude et l'altitude à laquelle nous voulons observer ce point. Voici comment définir un point :

```
<name>UQÀM - Coeur des sciences</name>
    <LookAt>
        <longitude>-122.0839597145766</longitude>
        <latitude>37.42222904525232</latitude>
        <range>500.6566641072245</range>
    </LookAt>
```

Finalement, le fichier doit terminer par la balise suivante :

```
</kml>
```

Avec Analypop, on peut générer un tel fichier pour afficher les lieux de naissance des individus d'une ascendance donnée. Pour ce faire, on doit connaître les coordonnées géograhiques de l'information que l'on veut représenter. La tâche consistant à relever les coordonnées géographiques de tous les lieux mentionnées pour les individus du registre de population n'a par contre pas encore été effectuée. Nous sommes actuellement en mesure de créer ce fichier, pour une ascendance donnée, mais avec des coordonnées prises au hasard. Cette fonctionnalité sera pleinement développée lors de la prochaine étude qui identifiera la transmission des biens immobiliers, à l'aide d'information provenant de l'index aux immeubles contennant toutes les transactions effecutées pour les différents immeubles.

La Figure 4.2 est l'image correspond aux coordonnées données à l'exemple précédent.

Fig. 4.2 Image correspondant aux coordonnées spécifiées par le fichier KML donné en exemple.

CONCLUSION

Comme nous avons pu le constater, cette nouvelle version d'Analypop constitue un outil extrêmement intéressant pour le laboratoire de recherche de Francine M. Mayer. En plus d'être assez souple pour permettre de s'adapter à un modèle de données spécifique, grâce aux fichiers de propriétés ainsi que par la flexibilité de son modèle de données, la composition de l'interface graphique permet une consultation organisée de l'information du registre de population.

D'autre part, la méthode implantée pour effectuer les jumelages permettant de sauvegarder l'historique des décisions offre aux utilisateurs l'assurance que les donnés ayant servies à la constitution du registre de population ne sont jamais perdues, elles sont plutôt enrichies au fur et à mesure de son utilisation. L'utilisateur peut, à tout moment, consulter l'historique des décision de jumelage et décider de revenir à l'état précédent si de nouvelles informations le porte à penser qu'une erreur de jumelage a été commise.

Par ailleurs, les deux formats de fichiers de sorties (GEDCOM et KML) permettent de développer de nouveaux moyens de représenter l'information du registre de population et pourront offrir aux chercheurs l'opportunité de développer de nouvelles initiatives de recherche. À ce propos, il est important de mentionner que le laboratoire de Francine M. Mayer prévoit déjà utiliser la fonctionnalité d'exporter certaines données sous le format KML afin d'identifier le mode de transmission de la propriété foncière de 6 familles fondatrices de La Patrie, dans les Cantons de l'Est, pour détecter les différentes stratégies utilisées pour l'établissement des enfants.

Malgré le fait que la version actuelle n'implante pas encore en totalité des fonctionnalités de l'ancienne version, nous considérons que l'effort produit à la réingéniérie d'Analypop permettra d'enrichir cette nouvelle version plus facilement. Le modèle de données flexible, l'utilisation d'un langage de programmation répandu, l'utilisation de techniques

d'abstraction des données, par l'utilisation de fichiers de propriétés, ainsi que le recours à des logiciels matures gratuits, feront en sorte qu'Analypop pourra être adapté pour répondre aux besoins grandissants des chercheurs. Bref, la version actuelle d'Analypop constitue un atout important pour le laboratoire de Francine M. Mayer qui prévoit continuer à investir dans son développement afin de lui donner un large éventail de fonctionnalités pour supporter les recherches les plus diverses.

Finalement, même si avec ce mémoire se termine une phase importante du développement d'Analypop, mon implication au sein de l'équipe de recherche de Francine M. Mayer se poursuivra. Je demeure intéressé aux sujets de ses recherches et j'espère pouvoir continuer à y contribuer légèrement.

ANNEXE A

FORMAT DE FICHIER GEDCOM

Voici un exemple d'un fichier GEDCOM. Ce fichier permet de décrire une famille composée de trois individus : le père étant Philippe Trudel, la mère Catherine Gariépy et leur enfant Marie-Marguerite Trudel.

Étiquette GEDCOM	Définition
0 HEAD	indique le début du fichier
1 SOUR Analypop	créateur de ce fichier
1 DATE 13 SEPT 2006	date de création du fichier
1 FILE Exemple_Gedcom.txt	nom du fichier
0 @I1@ INDI	début de la définition de l'individu 1
1 NAME Philippe Trudel	nom
1 SEX M	sexe
1 FAMS @F1@	indique qu'il est l'époux ("spouse") de la famille 1
1 BIRT	début de l'information sur la naissance
2 DATE 6 FEV 1704	date de naissance
2 PLAC Trois-Pistoles	lieu de naissance

0 @I2@ INDI	deuxième individu
1 NAME Catherine Gariépy	nom
1 SEX F	sexe
1 FAMS @F1@	épouse de la famille 1
1 DEAT	début de l'information sur le décès
2 DATE 14 JAN 1783	date de décès
0 @I3@ INDI	troisième individu
1 NAME Marie-Marguerite Trudel	nom
1 SEX F	sexe
1 FAMC @F1@	enfant ("child") de la famille 1
0 @F1@ FAM	définition de la famille 1
1 HUSB @I1@	le père ("husband") de la famille 1 est l'individu 1
1 WIFE @I2@	la mère ("wife") de la famille 1 est l'individu 2
1 CHIL @I3@	l'enfant de la famille 1 est l'individu 3
0 TRLR	indique la fin du fichier

BIBLIOGRAPHIE

(1) Boetsch, G., Prost, M., Rabino-Massa, E. *Evolution of consanguinity in a French Alpine Valley : The Vallouise in the Briancon Region (17th-19th centuries)*, Human Biology 74.2 (2002) 285-300.

(2) Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection*, London : John Murray, 496 p.

(3) "GEDCOM". 2007. In RootsWeb. En ligne.
http ://homepages.rootsweb.com/~pmcbride/gedcom/55gctoc.htm
Consulté le 16 mai 2007.

(4) "GEDCOM Testbook". In National Genealogical Society. En ligne.
https ://www.ngsgenealogy.org/ngsgentech/projects/TestBook2001/index.cfm
Consulté le 5 avril 2007.

(5) Griffiths, A., Wessler, S., Gelbart, W., Lewontin, R., Miller, J., Suzuki, D. 2006. *Introduction à l'analyse génétique*, Paris : Éditions De Boeck, 800 p.

(6) Jacquard, A. 1978. *Éloge de la différence, la génétique et les hommes*, Paris : Éditions du Seuil, 217 p.

(7) Kennedy, J., Kukla, R., Paterson, T. *Scientific Names Are Ambigous as Identifiers for Biological Taxa : Their Context and Definition Are Required for Accurate Data Integration.* In Data Integration in the Life Sciences : Second International Workshop (Californie, Juillet 2005) 80-95.

(8) Kroger, P., Bry, F. *A computational biology database digest : Data, data analysis, and data management*, Distributed and Parallel Databases 13.1 (2003) 7-42.

(9) Leboutte, R. 1998. *Du registre de population au registre national*, Italie : European University Institute, 50 p.

(10) "Licence GPL". In Free Software Foundation. En ligne.
http ://www.fsf.org/licensing/licenses/gpl-faq.html
Consulté le 5 avril 2007.

(11) Lima, M., Mayer, F., Coutinho, P., Abade, A. *Origins of a mutation : Population genetics of Machado-Joseph disease in the Azores (Portugal)*, Human Biology 70.6 (1998) 1011-1023.

(12) Lyell, C. 1830. *Principles of Geology. Being an Attempt to Explain the Former Changes of the Earth's Surface, by Reference to Causes now in Operation*, London : John Murray, 512 p.

(13) Morency-Bachand, E., Kamwena J.,*Analypop : Modernisation et convivialité d'un outil de gestion et d'analyse d'un registre de population par reingiénérie*, Entretiens Jacques-Cartier, Septembre 2004.

(14) "MySQL". 2007. In MySQL. En ligne.
http ://www.mysql.com/industry/
Consulté le 5 avril 2007.

(15) Nations Unies. *Les systèmes d'information géographique appliqués aux statistiques démographiques*, États-Unis : Édition des Nations Unies (1998) 32 p.

(16) Nault, F., Desjardins B. *Recent Advances in Computerized Population Registers*, Historical Methods 21 (1988) 29-33.

(17) Rabino-Massa, E., Prost, M., Boetsch, G. *Social Structure and Consanguinity in a French Mountain Population (1550-1849)*, Human Biology 77.2 (2005) 201-212.

(18) Skolnick, M. *A Computer Program for Linking Records*, Historical Methods Newsletter : quantitative analysis of social, economic, and political development 4.4 (1971) 114-125.

(19) Skolnick, M., Bean, L., Dintelman, S., Mineau, G. *A computerized family history data base system*, Sociology and social research 63.3 (1979) 425-612.

(20) "Tabulating Machine". 2007. In Wikipedia. En ligne.
http ://en.wikipedia.org/wiki/Tabulating_machine
Consulté le 5 avril 2007.

(21) Thomas Malthus. 1798. *An Essay on the Principles of Population*, London : John Murray, 542 p.

(22) Thomas Morgan. 1919. *A Critique of the Theory of Evolution*, Princeton : University Press, 197 p.

(23) Wesley, H., Chiaramella, Y., Dintelman, S., Maness, A., Mineau, G., Bean, L., Williams, R., Skolnick, M. *Record Linking Using a Genealogical Database System*, Information Systems for Differential Demographic Analysis. In International Union for the scientific Study of Population (IUSSP) General Conference (Italie, Juin 1985) 5-12.