Survol des questions touchant l'utilisation d'identificateurs personnels

Document rédigé par Mark Armstrong

DMEM, Statistique Canada

7 juillet 2000

Table des matières

Liste des annexes

A. Variables de l'identification personnelle

1. Introduction

L'objet du présent rapport est de faire le survol des méthodes et techniques existantes qui utilisent les identificateurs personnels en vue de réaliser le couplage des enregistrements. Ce couplage peut être décrit de façon générale comme une méthode de traitement ou de transformation des identificateurs personnels tirés des dossiers personnels enregistrés dans l'une ou plusieurs bases de données opérationnelles afin de jumeler les identificateurs et de créer un dossier composé sur un particulier. Le couplage des enregistrements ne vise pas seulement à identifier les particuliers à des fins opérationnelles, mais à établir les concordances probabilistes de degrés de fiabilité variés à des fins de rapports statistiques. Les techniques utilisées dans le cadre du couplage d'enregistrements peuvent également servir dans les enquêtes afin d'en restreindre le champ dans les bases de données, lorsque des renseignements sur les identificateurs personnels existent.

L'identification de particuliers peut se faire de diverses façons. Pour assurer le suivi d'un particulier au fil du temps ou dans différentes bases de données, il est préférable de lui assigner un identificateur personnel unique. Cet identificateur unique peut être un numéro comme le numéro d'assurance sociale (NAS) utilisé au Canada ou le numéro de sécurité sociale (NSS) utilisé aux États-Unis. En théorie, ces identificateurs sont assignés à un particulier donné et ne changent pas même si celui-ci change son nom, son lieu, sa date de naissance, etc. Les numéros sont habituellement assignés de façon séquentielle et donnent des renseignements précis au sujet d'un particulier. Le premier numéro du NAS, par exemple, correspond à la province où le numéro a été assigné.

Le second type d'identificateur personnel est l'identificateur non unique. Cet identificateur peut servir à établir l'identité d'un particulier, compte tenu d'un certain nombre de critères, qui peuvent se fonder sur des renseignements démographiques, sociaux, physiques ou administratifs. Les renseignements démographiques ou sociaux sont le nom de famille, les prénoms et initiales, la date de naissance, le sexe, l'appartenance ethnique, etc. Les renseignements physiques sont la couleur des yeux et de la peau, certaines caractéristiques distinctives (des doigts manquants), le dactylogramme et d'autres attributs biométriques. Les renseignements de nature administrative comprennent le numéro d'enregistrement des contribuables, le numéro assigné aux détenus, le numéro du permis de conduire, etc. Il est peu probable que les numéros assignés à un particulier donné changent.

L'utilisation d'identificateurs personnels uniques n'est guère généralisée. Dans les pays où des registres de population ont été établis, l'utilisation d'identificateurs personnels uniques est plus probable. À titre d'exemple, les citoyens du Danemark possèdent un numéro unique qui est mis à jour sur une base continue afin de refléter les changements démographiques et le lieu de résidence. Ce numéro unique est utilisé dans des domaines administratifs variés.

Par contre, les Canadiens n'ont pas d'identificateur personnel unique. Toutefois, comme dans la majorité des pays, un certain nombre de variables peuvent permettre d'établir l'identification personnelle unique, sans toutefois la garantir. Ces variables sont utilisées en combinaison pour établir l'identité d'une personne à la fois dans un sens transversal et longitudinal. Beaucoup de ces variables sont utilisées dans une grande variété de fichiers et de bases de données administratives. D'autres variables servent à des fins administratives particulières, comme c'est le cas pour le permis de conduire, le numéro d'empreintes digitales ou d'employé.

La facilité et l'exactitude du suivi des particuliers, de quelque façon que ce soit, sont tributaires de l'existence d'identificateurs uniques et non uniques. L'identification d'un particulier en se servant d'un identificateur non unique comporte plusieurs processus. Il est important de les connaître et de comprendre les conséquences de l'utilisation des différentes variables de l'identification personnelle afin d'identifier des particuliers donnés.

2. Utilisation d'identificateurs personnels dans le système de justice du Canada

À l'heure actuelle, il n'y a aucun système national d'identification qui permette d'identifier précisément un individu qui a des démêlés avec la justice canadienne. Aux États-Unis et en Australie, des systèmes ont été mis sur pied afin d'assurer le suivi des particuliers qui sont enregistrés dans le système de justice pénale. Ces systèmes ont adopté un numéro unique qui est assigné à toute personne appréhendée ou condamnée. La section B comprend de plus amples renseignements sur ces systèmes.

La plupart des bases de données statistiques du système de justice qui visent les particuliers renferment des variables démographiques communes. Ces variables sont enregistrées et mises à jour par les territoires et provinces ainsi que les ministères fédéraux qui s'acquittent de responsabilités judiciaires. Les variables communes qui sont signalées dans les enquêtes judiciaires de Statistique Canada comprennent le nom de famille, le prénom et les initiales, la date de naissance, le sexe, le numéro SED (Système d'empreintes digitales), le code de l'infraction et les dates importantes. Toutefois, ces variables ne sont pas toutes signalées par chaque secteur de compétence dans les différentes enquêtes judiciaires, bien que les identificateurs personnels le soient. La combinaison de ces identificateurs non uniques, en utilisant les techniques de couplage d'enregistrements, permet d'établir les liens entre des particuliers dans un même fichier de données ou entre différents fichiers. Le tableau A1 illustre les renseignements démographiques qui sont recueillis dans cinq enquêtes annuelles à base de microdonnées du Centre canadien de la statistique juridique (CCSJ).

Tableau A1 : *Variables démographiques recueillies dans le cadre des enquêtes à base de microdonnées du CCSJ*

Enquête	Nom de famille	Prénom	Initiale(s)	Chiffrement du nom	Sexe	Date de naissance
DUC 2.0, 2.1	Non	Non	Non	Oui	Oui	Oui
Homicides	Oui	Oui	Oui	Non	Oui	Oui
Tribunaux de la jeunesse	Non	Non	Non	Oui	Oui	Oui
Services correctionnels pour les jeunes	Non	Non	Non	Oui	Oui	Oui
Tribunaux pour adultes	Non	Non	Non	Oui	Oui	Oui

Tableau A1 : *Variables démographiques recueillies dans le cadre des enquêtes à base de microdonnées du CCSJ*

Enquête	Date de l'infraction	Code de l'infraction	N° de réf. du secteur de compétence	Empreintes digitales	Autres attributs biométriques
DUC 2.0, 2.1	Oui	Oui	Oui	Non	Non
Homicides	Oui	Oui	Non	Oui	Non
Tribunaux de la jeunesse	Oui	Oui	Oui	Non	Non
Services correctionnels pour les jeunes	Oui	Oui	Oui	Non	Non
Tribunaux pour adultes	Oui	Oui	Non	Non	Non

Les répondants ne signalent pas toutes les variables pour tous les dossiers entrants. Par exemple, bien que le numéro SED soit signalé dans l'Enquête sur les homicides, tous les dossiers entrants portent un numéro SED. Dans le cadre de l'Enquête sur les tribunaux de la jeunesse, un secteur de compétence signale au CCSJ le nom de famille, le prénom et les initiales des jeunes contrevenants. D'autres écarts mineurs comme ceux indiqués ci-dessus concernent les indicateurs « oui » et « non » du tableau A1. Il est question de ces écarts de façon plus détaillée à l'annexe A. La section 4 et l'annexe H donnent des renseignements supplémentaires sur le numéro SED et son utilisation au Canada.

L'uniformité des données d'enquête signalées varie d'un répondant à l'autre. Par exemple, certains répondants signalent des données démographiques complètes pour chaque personne, tandis que d'autres peuvent indiquer que des données sont « manquantes » ou « inconnues ». Il en découle certaines difficultés au niveau de la qualité des données recueillies qui peuvent ou non avoir des répercussions sur l'assignation d'un identificateur personnel unique.

De façon générale, la plupart des enquêtes importantes à base de microdonnées entreprises par Statistique Canada recueillent des renseignements démographiques de base. Au nombre des variables indiquées au tableau A1 s'ajoutent l'état matrimonial, l'éducation, les indicateurs d'emploi, la langue, l'appartenance à un groupe autochtone, le lieu géographique (province ou emplacement géographique plus précis). Certaines de ces variables sont, dans différentes mesures, recueillies par les répondants aux cinq enquêtes judiciaires à base de microdonnées. Toutefois, le recours à ces variables afin d'identifier précisément des personnes au fil du temps n'est pas très fiable.

Plusieurs études sur le couplage d'enregistrements ont été entreprises à Statistique Canada et s'appuyaient sur les fichiers de données du CCSJ. Ces études ont été documentées et un rapport a résumé le travail accompli. Chaque étude abordait la question de l'utilisation des identificateurs personnels pour réaliser le couplage des enregistrements. La question de la qualité des données a également été abordée. Toutefois, ces études n'ont porté que sur un ou deux secteurs de compétence. Pourtant, malgré les limites des études, il y a aussi été question des difficultés méthodologiques liées au couplage des fichiers de données judiciaires. En 1998, le CCSJ a produit un rapport qui a fait état des travaux réalisés dans le couplage des enregistrements des fichiers judiciaires dans les années 1990 (Centre canadien de la statistique juridique, 1998).

3. Autres méthodes d'établissement d'identificateurs personnels uniques

La dernière colonne du tableau A1 est titrée « Autres attributs biométriques ». La biométrie concerne la mesure, la description et la classification des caractéristiques physiques. À l'heure actuelle, les enquêtes à base de microdonnées du CCSJ ne comprennent aucune donnée biométrique, à l'exception du numéro SED. L'utilisation de données biométriques aux fins de l'identification de personnes dans les systèmes de justice étrangers est en hausse, ce qui est sans doute attribuable au degré d'exactitude des mesures effectuées, aux coûts associés et à l'utilité de ce type de données.

Les mesures biométriques comprennent les différentes techniques dactyloscopiques, la palmiscopie, le balayage rétinien et irridien, le profil de l'acide désoxyribonucléique (ADN), la reconnaissance de l'empreinte vocale, la couleur des yeux et des cheveux, les photos de criminels, la taille et le poids. Certaines de ces méthodes permettent d'identifier précisément un individu, tandis que d'autres ne l'identifient que de façon générale. Bien qu'il soit plus difficile d'obtenir des renseignements biométriques comparativement aux données démographiques, leur utilisation permet l'identification précise des particuliers n'importe quand durant leur vie. Une variable, telle que le profil de l'ADN, permet d'identifier précisément un particulier, et ce, quels que soient les changements qu'il subit au fil des ans. La fiabilité des renseignements biométriques corrects est plus grande que celle des données démographiques, surtout dans le milieu de la justice pénale où des personnes tentent intentionnellement de déjouer les autorités judiciaires qui recueillent des renseignements sur les identités. L'utilisation de pseudonymes, d'une identité ou de documents falsifiés ou altérés et d'autres types de falsification auxquels ont recours les contrevenants restreignent l'usage des identificateurs personnels communs.

Le rapport du gouvernement de l'Ontario (1998) donne une excellente vue d'ensemble des méthodes biométriques utilisées par les systèmes de justice. Le rapport couvre une grande variété de méthodes et fait la lumière sur de nombreux aspects de leur utilisation.

4. Couplage des enregistrements relatifs aux données criminelles au Canada

Il y a des centaines de forces policières au Canada qui maintiennent des fichiers de données sur les actes criminels. Le plus important dépôt de données est tenu par le Centre d'information de la police canadienne (CIPC). Ce Centre fait partie de la Gendarmerie Royale du Canada qui assure les services de police partout au Canada. Environ 40 % des incidents criminels signalés au Canada sont visés par une investigation de la GRC.

Les renseignements dans les casiers judiciaires sont versés dans une banque de données sur l'identité. Cette banque de données est tenue à jour par le personnel des Services d'information et d'identité judiciaire de la GRC, pour le compte des services de police canadiens. Il revient au service de police d'origine d'assurer l'exactitude des données qu'il fournit à la banque de données.

La variable clé pour obtenir des renseignements de la banque de données du CIPC est le numéro SED. Ce numéro est assigné par les Services d'information et d'identité judiciaire lorsqu'ils reçoivent des empreintes digitales qui n'ont pas fait l'objet au préalable d'un classement ou d'une saisie. Les empreintes digitales doivent être assorties d'un code d'infraction criminelle. On obtient les empreintes digitales suivant une procédure normalisée, et celles-ci sont enregistrées sur le formulaire C-216.

Les interrogations dans la banque de données du CIPC ne peuvent être faites qu'au moyen du numéro SED. En utilisant cette seule variable, les dossiers judiciaires complets sont obtenus. Lorsqu'il n'y a pas de numéro SED, les interrogations peuvent s'appuyer sur le Fichier judiciaire nominatif (FJN). Au moyen du FJN, les recherches sur les prénoms et noms de famille, le sexe, la date et le lieu de naissance, la race, la couleur des yeux et les mensurations physiques (taille et poids métriques) peuvent être effectuées. L'interrogation du FJN, en utilisant l'une ou l'autre des variables ci-dessus, peut établir plusieurs jumelages potentiels. Chaque jumelage potentiel est attribué un coefficient de pondération de vraisemblance. La pondération a une valeur maximale suivant les données qui sont saisies à l'interrogation. Le logiciel de recherche utilise l'orthographe phonétique du nom de famille ainsi que des variantes orthographiques. Les recherches qui portent sur des noms composés sont également effectuées en empruntant plusieurs méthodes afin d'optimiser le nombre de jumelages potentiels (GRC, 1999).

Les numéros SED sont compris dans d'autres fichiers de données, en plus de ceux de la banque de données du CIPC. De nombreux établissements correctionnels et certaines forces policières incluent le numéro SED. Lorsque le numéro SED est compris dans deux différents fichiers, celui-ci permet de jumeler les données tirées de ces deux fichiers, ce qui est très utile. L'annexe H présente plusieurs autres questions liées au numéro SED.

B. Utilisation d'identificateurs personnels à l'extérieur du Canada

Les types d'identificateur personnel de criminels utilisés dans d'autres pays sont similaires à ceux qui sont utilisés au Canada. Les renseignements démographiques standards comprennent les noms, la date de naissance et le sexe. L'âge est souvent utilisé de façon conjointe à la date de naissance. La variable « âge » doit être définie précisément puisqu'elle change à différents

points du système de justice. Par exemple, la police obtient l'âge au moment de l'arrestation, et les données correctionnelles peuvent indiquer l'âge au moment de l'entrée ou de la sortie d'une installation ou d'un programme correctionnel.

États-Unis

Aux États-Unis cela fait plus de 10 ans qu'on met au point les statistiques sur les transactions relatives aux contrevenants (OBTS). Les systèmes associés sont exploités à l'échelle nationale ainsi que dans les divers États et sont alimentés par les autorités de justice pénale qui fournissent des données sur les contrevenants aux dépôts centraux au cours de l'année. Les lignes directrices sur les OBTS précisent le type de données que chaque autorité doit fournir. Les renseignements sur les contrevenants comprennent l'âge, la race, le sexe, l'origine ethnique, l'organisme qui a procédé à l'arrestation, la date de l'arrestation, l'infraction commise, la date et le type de dispositions policières, les poursuites judiciaires, les actions préparatoires et les activités du tribunal, telles que les dates, les dispositions, le genre de procès et le plaidoyer final. Les résultats du procès et toute disposition prise sont également enregistrés et acheminés aux services centraux d'archivage.

Une importante caractéristique des OBTS est que les enregistrements peuvent établir une concordance entre les données sur les personnes appréhendées, les incidents et les accusations. Les personnes appréhendées sont identifiées au moyen des données démographiques et des empreintes digitales. Un numéro unique leur est assigné, qui peut être jumelé à tout dossier antérieur. Un numéro de cas est également assigné. Ce numéro permet d'établir des liens entre les incidents pour lesquels des accusations multiples ont été portées ou qui comportent de nombreux accusés pour un seul incident.

Plusieurs limites sont associées aux OBTS qui découlent de décisions qui ont été prises lors de l'élaboration et de questions touchant les données. Puisque ce ne sont pas tous les États qui participent aux OBTS, le suivi des contrevenants ne peut être fait de façon intégrale. Une autre contrainte est que les contrevenants ne sont pas tous soumis à la dactyloscopie; donc, ils ne sont pas tous enregistrés dans les OBTS. Puisque les jeunes contrevenants ne font pas partie des OBTS et que seuls les délits graves sont inclus, les OBTS ne comprennent que certains des actes criminels commis. De plus, les OBTS ne contiennent aucune donnée au niveau correctionnel (Ferrante, 1993).

Des répertoires nominatifs principaux (RNP) ont été créés par les États. Ces répertoires contiennent les noms et d'autres identificateurs des personnes qui ont un casier judiciaire. Ils servent à diverses fins, y compris les enquêtes criminelles, la vente d'armes à feu ou l'établissement du montant de la caution. En 1992, presque tous les RNP des États étaient automatisés et stockaient des données presque intégrales.

L'échange de renseignements sur les criminels entre les États se fait au moyen de l'Index d'identification interétatique (III). Le Federal Bureau of Investigation (FBI) maintient son propre index d'identification des personnes qui ont commis des délits graves, aux termes des lois étatiques ou fédérales. L'index comprend le nom, la date de naissance, la race et le sexe de chaque individu. Les recherches sont effectuées au moyen du nom et d'autres identificateurs personnels. L'III comprend le Fichier national des empreintes digitales (FNED). Conformément aux procédures établies pour le FNED, les États n'envoient au FBI que les empreintes digitales des contrevenants primaires ainsi que d'autres identificateurs, tels que le nom et la date de naissance. L'III comprend aussi des caractéristiques biométriques comme les images rétiniennes et les empreintes vocales. Ces méthodes d'identification efficaces sont préférables à l'utilisation de variables non uniques, comme le nom, le sexe et la date de naissance. Le mémoire du ministère de la Justice des États-Unis (1997) récapitule les questions liées aux systèmes étatiques et interétatiques de renseignements historiques sur les criminels.

Une autre méthode d'identifier les personnes qu'utilise le système de justice pénale des États-Unis est le Combined DNA Index System (CODIS), qui a récemment été mis au point. Depuis juin 1998, les 50 États recueillent des échantillons d'ADN, surtout en ce qui concerne les délinquants criminels condamnés. Les États particuliers peuvent recueillir des échantillons d'ADN pour d'autres actes criminels, tels que les meurtres, les homicides involontaires, les voies de fait et les vols qualifiés. Chaque individu a un profil d'ADN propre. Les profils sont saisis dans le CODIS, ce qui permet aux États et aux laboratoires médico-légaux des forces policières locales d'échanger et de comparer les renseignements sur les profils d'ADN. Ces échanges et comparaisons se font par voie électronique, au même titre que l'échange de renseignements sur les empreintes digitales. On ne sait pas au juste l'utilisation que l'on fait des profils d'ADN.

Le CODIS a été mis au point à la suite de l'adoption de la *Loi sur l'identification par les empreintes génétiques* de 1994. Ensuite est venu le programme d'amélioration des laboratoires judiciaires d'analyse de l'ADN qui visait à accroître les capacités de ces laboratoires et à appuyer les enquêtes et la judiciarisation des crimes violents (ministère de la Justice des États-Unis, 2000).

Australie

Dans la foulée de l'élaboration des OBTS et de leur utilisation aux États-Unis, le Crime Research Centre de l'University of Western Australia a mis au point l'Integrated Numerical Offender Identification System (INOIS) à la fin des années 1980. Le principal objectif du projet était d'établir un identificateur commun et unique pour les contrevenants de façon à constituer une base de données longitudinales en Australie de l'Ouest. La base de données permettrait également, au fil du temps, de faire le suivi des contrevenants au sein du système de justice pénale.

Le numéro de l'INOIS est fondé sur le numéro de dossier assigné à un contrevenant primaire. L'identité des individus est validée au moyen des empreintes digitales et des enregistrements connexes. Le numéro de l'INOIS est séquentiel. Un numéro est assigné à chaque contrevenant et peut ainsi être utilisé par les différents systèmes de justice pénale, y compris le système judiciaire pour les jeunes, les établissements correctionnels ainsi que le programme des libérations conditionnelles. L'identification au moyen des empreintes digitales permet d'assurer l'exactitude du numéro de l'INOIS, qui est propre à chaque individu (Ferrante, 1993).

Le fonctionnement de l'INOIS est le suivant : à chaque trimestre, ou sur une autre base régulière, les organismes collaborateurs envoient les dossiers sur les contrevenants qui comprennent les identificateurs nominatifs et d'autres renseignements démographiques. Ces dossiers sont systématiquement comparés aux enregistrements historiques criminels établis par la police. Les dossiers sont ensuite retournés aux organismes de justice pénale, qui comportent maintenant un identificateur INOIS qui a été assigné à chaque individu pour lequel un jumelage a été réalisé. Les organismes n'utilisent que le numéro INOIS lorsqu'ils fournissent par la suite des renseignements (les identificateurs nominatifs ne sont plus utilisés). Le Crime Research Centre ajoute ces enregistrements à la base de données longitudinales en se servant du numéro INOIS à titre de clé.

Lorsqu'un jumelage exact ne peut être réalisé, le système de couplage utilise une approche probabiliste afin de déterminer si les enregistrements de sources variées, qui n'utilisent pas d'identificateurs communs uniques, peuvent être jumelés. À la section D, il est question des concepts de jumelage ou de couplage des enregistrements.

Royaume-Uni

L'Oxford Record Linkage Study (ORLS) a porté sur 10 millions de dossiers se rapportant à cinq millions de personnes et a couvert la période de 1963 à aujourd'hui. Les données intégrées servent à préparer des statistiques sur les services de santé et à la recherche épidémiologique et sur les services de santé. Bon nombre des concepts qui ont servi à élaborer l'ORLS sont également pertinents à d'autres domaines. Au Royaume-Uni, de façon générale, on n'a pas utilisé les identificateurs personnels uniques pour assurer le suivi des particuliers au fil du temps et dans les différents secteurs du système de santé. Les techniques de couplage d'enregistrements sont donc utilisées pour déterminer les différents enregistrements se rapportant à un même particulier.

L'ORLS utilise de nombreuses variables pour faire le lien entre les données se rapportant à une même personne. La principale variable est le nom de famille actuel. Au lieu d'utiliser le nom de famille dans sa forme écrite, celui-ci est transformé ou chiffré. Les variables secondaires et ultérieures comprennent l'initiale du premier prénom, le second prénom ou l'initiale ainsi que le nom de famille à la naissance. Les variables non nominatives qui servent à l'identification comprennent la date de naissance, le sexe, le lieu de naissance et l'adresse domiciliaire.

Le couplage des enregistrements est réalisé en utilisant une méthode probabiliste. Le système OX-LINK relève et élimine également les renseignements en double par jumelage croisé. L'algorithme de compression de nom d'Oxford (ACNO) est décrit à la section 3 (Gill, 1997).

C. Utilisation des noms à titre d'identificateurs uniques

1. Introduction

L'utilisation des noms à titre d'identificateurs personnels est commune parce que toutes les personnes ont un nom, bien que le nom ne soit pas uniforme, c'est-à-dire comportant à la fois un nom de famille et un ou des prénoms. Certaines personnes n'ont qu'un seul nom, tandis que d'autres peuvent avoir plusieurs noms et prénoms officiels. La structure d'un nom est fonction de l'appartenance ethnique d'une personne. L'utilisation de plus d'un prénom est commune dans certaines régions du monde.

Une autre raison pour laquelle on utilise les noms à titre d'identificateurs personnels est qu'ils sont connus d'autres personnes. Le nom d'une personne est l'identificateur personnel le plus universel et c'est le renseignement démographique le plus utilisé et indiqué dans les fichiers de données.

2. Certains problèmes posés par les noms

Il arrive fréquemment qu'on utilise le nom aux fins de l'identification, mais il faut tout de même l'utiliser avec discernement. Cela est particulièrement vrai en ce qui concerne le milieu judiciaire. Bien que toutes les personnes reçoivent un nom à la naissance, leur nom peut changer au fil du temps. Les raisons communes sont le mariage, le divorce, le remariage, l'adoption et le changement légal de son nom. D'autres raisons sont la déception, les variantes d'un nom (le changement de sexe, par exemple), l'inversion des prénoms ou l'utilisation d'initiales au lieu des prénoms. Certains « changements » de nom peuvent découler de l'appropriation ou de la fabrication d'une identité. C'est pourquoi l'utilisation des noms à titre d'identificateurs personnels pose certains problèmes (Newcombe, 1988).

Un problème se présente dans le milieu judiciaire lorsqu'il faut enregistrer les noms et les pseudonymes. Le fait que certaines personnes aient plus d'un nom peut nuire à la capacité de faire le suivi de celles-ci dans un même dossier ou d'un dossier à l'autre.

Contrairement aux autres variables démographiques comme la date de naissance et le sexe, il est difficile de vérifier l'exactitude des noms. Au moment de valider une date de naissance, les systèmes informatiques peuvent être programmés de façon à vérifier que chaque élément de la date corresponde à un ensemble donné de valeurs et que les liens entre ces éléments répondent à diverses conditions. En ce qui concerne le sexe, la variable est habituellement un indicateur numérique pour lequel le système informatique accepte trois valeurs : « féminin », « masculin », « inconnu ». D'ailleurs, suivant l'ethnographie, des centaines de milliers de noms différents existent.

3. Problèmes posés par la qualité des données rattachées aux noms

Les questions soulevées ci-dessus semblent indiquer que des problèmes liés à la qualité des données peuvent se poser lorsqu'on utilise les noms à titre d'identificateurs personnels. Si les données sur les noms sont versées dans des fichiers, d'autres difficultés sont possibles. Une difficulté commune relative à la collecte de données est l'interrogation correcte et l'enregistrement de la réponse donnée. Les personnes appréhendées par la police peuvent donner un sobriquet ou seulement leurs initiales, et c'est ce qui est écrit sur le formulaire de collecte de renseignements. Afin d'obtenir des renseignements de qualité, l'agent de police doit poser des questions d'approfondissement ou vérifier le nom donné par un particulier dans les documents officiels et établir le nom véritable des personnes. À titre d'exemple, le prénom d'une personne appréhendée peut être enregistré comme « Bill » ou « William » sur ses papiers d'identité, mais celle-ci peut dire que son nom est « Billy ». Si le système de collecte de données ne permet la saisie que d'un seul prénom, il faut alors saisir le prénom convenable.

D'autres types d'erreurs peuvent se produire relativement aux noms, et se produisent de fait. Une erreur de transcription ou de saisie au clavier peut causer des problèmes au moment d'établir une identité unique. Par exemple, les erreurs de transcription peuvent se produire lorsque la personne qui enregistre ou recueille les données présume à tort que l'orthographe de « Mark » est « Marc ».

Les décisions sur la façon de saisir certains caractères comme les traits d'union, les apostrophes et les accents peuvent également poser problème. En outre, les permutations de certains noms peuvent présenter des difficultés, surtout en ce qui concerne les cultures asiatiques où il est commun d'avoir de multiples noms. Certains noms sont très longs et peuvent dépasser l'espace alloué dans un champ ou un fichier de données. En pareil cas, il faut trancher. On peut décider de laisser tomber les dernières lettres du nom ou adopter un autre moyen.

Le fait de se rendre compte qu'il y a des difficultés associées à l'utilisation du nom à titre d'identificateur personnel n'est pas l'apanage du domaine judiciaire. Les problèmes et les solutions de cette nature sont documentés par des disciplines variées. Des domaines ont adopté des solutions complètes ou partielles au fil des décennies. Les améliorations et les précisions apportées aux solutions qui ont fait leurs preuves sont toujours à l'étude.

Il y a de nombreuses méthodes qui permettent de réduire le nombre d'erreurs typographiques qui sont commises au moment de saisir un nom dans un fichier de données. Certains programmes informatiques sont exécutés sur les noms entrants et produisent un dossier nominatif qui comporte des erreurs d'orthographe ou de saisie. Les trois méthodes suivantes sont des méthodes types : la méthode de comparaison de chaînes Jaro, la méthode Winkler et la méthode Damerau-Levenstein. L'utilisation de ces méthodes est abordée à l'annexe I. Ces méthodes, et d'autres, ne s'appliquent pas seulement aux prénoms et aux noms de famille, mais aussi aux noms de rues et d'entreprises. Parce que le CCSJ reçoit des noms chiffrés, l'utilisation de la méthode de comparaison de chaînes n'est pas applicable. Toutefois, l'utilisation de comparateurs de chaînes peut servir dans les applications de bases de données locales, ce qui permettrait d'améliorer la qualité des données locales et aussi, celle des données chiffrées.

4. Chiffrement des noms

Afin de pouvoir utiliser la variable nominative aux fins de l'identification personnelle, des méthodes ont été élaborées qui réduisent les problèmes décrits ci-dessus. Une des techniques les plus communes est l'utilisation d'un algorithme afin de coder un nom. Ces méthodes de chiffrement sont utilisées sur le nom de famille seulement ou les prénoms et initiales. Une démarche commune pour chiffrer les noms est de les orthographier phonétiquement. Ce faisant, les noms dont la prononciation est la même sont groupés ensemble. Cela permettra d'éliminer un bon nombre de problèmes associés à la collecte de renseignements sur les noms. Pourtant, le chiffrement des noms ne résoudra pas tous les problèmes, tels que les pseudonymes. Toutefois, les méthodes de chiffrement ont permis de régler de nombreux problèmes communs associés aux noms. Ces méthodes diffèrent de celles qui éliminent simplement les derniers caractères. Les algorithmes qui sont fondés sur la compression peuvent différer grandement des méthodes employant la phonétique (Newcombe, 1988).

Il y a deux différentes méthodes de chiffrement des noms qui sont utilisées dans le cadre des enquêtes à base de microdonnées du CCSJ. Toutes deux reposent sur le principe du regroupement phonétique des noms. Ces algorithmes peuvent également être utilisés pour les noms de rues ou d'entreprises. On décrit ci-dessous chacune de ces méthodes sommairement ainsi que d'autres méthodes de chiffrement des noms.

4a. Système de codification Russell-Soundex

Le premier système de codification phonologique des noms largement appliqué a été élaboré par Margaret Odell et Robert Russell et breveté en 1918. Ce système est très utilisé aujourd'hui et on le désigne simplement comme le code « Soundex ». Les règles de la codification Soundex sont bien connues et faciles à appliquer. La codification manuelle peut se faire rapidement suivant un ensemble de règles normalisées. La structure du code Soundex comprend quatre caractères, dont le premier est une lettre suivie de trois chiffres de 0 à 6. Par exemple, en suivant les règles résumées à l'annexe B, le code Soundex du nom de famille « Hilbert » est H460. Le même code Soundex est donné au nom « Heilbronn ». Le code du nom de famille « Rogers » est R262 et « Rodgers », R326.

L'objet de cette méthode de chiffrement est d'inclure des noms semblables dans un même groupe logique et les noms différents dans des groupes distincts. Une difficulté découlant du code Soundex est que ces deux possibilités existent. Dans les exemples donnés ci-dessus, les noms « Hilbert » et « Heilbronn » ont le même code Soundex malgré que leur prononciation diffère. D'ailleurs, la prononciation des noms « Rogers » et « Rodgers » est semblable, et pourtant ces noms appartiennent à des groupes Soundex différents. Ce genre de résultat indésirable est fort commun avec le système Soundex et est mis en évidence dans les algorithmes de chiffrement qui ont été élaborés après l'entrée en vigueur du code Soundex.

À mesure qu'augmentait l'utilisation du code Soundex, les problèmes et les limites associés ont été reconnus. Le code Soundex a servi à coder les noms dès le recensement américain de 1880. Puisque les noms étaient typiquement « américains », le code Soundex convenait fort bien. Les consonnes muettes et les voyelles combinées ne sont pas aussi fréquentes dans les noms américains comparativement, par exemple, aux noms britanniques.

L'application du code Soundex à certains noms de groupes ethniques pose également problème, comme l'indique l'exemple suivant. Les noms « Van der meer » et « Van der berg » ont le même code Soundex (V536). En utilisant la méthode Soundex, il n'est pas facile de distinguer les noms qui commencent pas « van der ». À mesure que s'accroît la diversité culturelle au Canada, les variantes des noms de famille et des prénoms augmentent également. Le code Soundex standard (nom de famille seulement et règles de base) permet de moins en moins de réaliser ce pourquoi il a été élaboré.

Il y a 6 734 différents groupes de noms qui peuvent être créés au moyen du code Russell-Soundex (voir l'annexe B). Les codes Soundex varient de A000 à Z666. Bon nombre des différents codes Soundex ne sont utilisés qu'à l'occasion. D'autres noms peuvent être fort communs. Il en découle qu'un même code Soundex figure fréquemment dans un gros fichier de noms de famille chiffrés. À titre d'exemple, des codes Soundex ont été produits pour les personnes figurant dans l'Enquête sur les tribunaux de juridiction criminelle pour adultes de 1996. Les codes « B626 », « G255 », « L145 » et « T651 » ont figuré des dizaines de milliers de fois alors que les codes « A133 » et « A240 » étaient peu fréquents. D'autres codes Soundex n'ont pas été utilisés.

4b. Méthode Henry

Puisque les règles du système Soundex font qu'il est plus difficile de coder les noms dont la prononciation est semblable et de regrouper les noms dissemblables, une autre méthode a été élaborée à l'Université de Montréal. La méthode Henry a été conçue pour remplacer le code Soundex et convient davantage aux noms français. Comparé à l'algorithme Soundex, le code Henry comporte des règles plus nombreuses. La structure du code utilisé dans le cadre de l'Enquête sur les tribunaux de la jeunesse et de l'Enquête sur les tribunaux de juridiction criminelle pour adultes (ETJCA) du CCSJ comprend six lettres, dont les quatre premières sont tirées du nom de famille et les deux dernières d'un prénom. D'autres variantes ont été programmées afin d'utiliser, par exemple, le nom de famille seulement. Le code Henry sert également à chiffrer les noms d'entreprises qui figurent dans l'ETJCA. Une présentation sommaire de l'algorithme Henry est comprise à l'annexe C.

Il est à noter que le système Henry qui sert au chiffrement phonétique n'a aucun lien avec le système d'empreintes digitales Henry.

4c. Autres méthodes de chiffrement appliquées aux noms

Comme il a été mentionné, des progrès sont réalisés dans l'élaboration de méthodes de chiffrement aux fins de la codification des noms. Les améliorations qui découlent de la méthode Soundex ont produit plusieurs autres algorithmes bien connus. Le code Henry en est un exemple qui est utilisé au Canada. En 1963, le New York State Identification and Information System (**NYSIIS**) a été mis au point. Après plusieurs tentatives pour améliorer la méthode de codification Soundex, le code NYSIIS a été introduit afin d'éliminer certaines des nombreuses difficultés associées au code Soundex. Puisque les noms peuvent être complexes, l'algorithme du NYSIIS a permis d'analyser les améliorations qui ont été apportées depuis le début des années 1920 au code Soundex.

Le code NYSIIS est maintenant utilisé partout dans le monde pour coder les noms dans de nombreux domaines, dont la santé et la justice. L'algorithme est plus complexe que celui du système Russell-Soundex, bien que les objectifs visés par les deux systèmes soient les mêmes. Le code NYSIIS est strictement alphabétique et la longueur du code est fonction de la longueur des noms à coder. Des règles spéciales ont été établies pour des applications particulières lorsque des noms de famille ethniques sont en cause.

Le code NYSIIS est généralement préféré au code Soundex pour plusieurs raisons. Une raison importante est que le nombre de codes NYSIIS valides est supérieur aux variations possibles du code Soundex, ce qui permet une plus grande distinction entre les noms et de retenir des renseignements importants sur les noms. La structure du code NYSIIS conserve la position des voyelles (même si toutes les voyelles sont remplacées par des « A ») au lieu de les éliminer comme dans la méthode Soundex (à moins que la première lettre du nom soit une voyelle).

Le système INOIS utilisé en Australie s'appuie sur le code NYSIIS pour exécuter la routine de jumelage et établir l'identité des contrevenants particuliers (voir la section B). Le code NYSIIS est également utilisé dans les grandes banques de données sur la santé maintenues à Statistique Canada. La démarche NYSIIS est présentée à l'annexe D.

L'algorithme de compression de nom d'Oxford (**ACNO**) est utilisé au Royaume-Uni pour changer le nom de famille des patients en un code chiffré. L'ACNO utilise une version anglicisée de l'algorithme NYSIIS, puis applique l'algorithme Soundex afin de produire un code Soundex normalisé de quatre caractères. Ainsi, la méthode ACNO produit des codes ou des noms dont la grandeur varie suivant que les noms sont plus ou moins communs. La subdivision de ces codes est réalisée en utilisant d'autres informations disponibles. Les chercheurs qui ont élaboré la méthode ACNO ont remarqué que le code Soundex standard n'était pas très utile pour les variantes de noms communs (c.-à-d. Thomson et Thompson qui correspondent à deux codes différents) et qu'il n'était pas très efficace pour les noms courts ou les noms qui comprennent beaucoup de voyelles, tels que les noms d'origine asiatique (Gill, 1997).

D'autres précisions et variations ont été apportées au fil des ans à la méthode Soundex. L'IBM Alpha Search Inquiry System (dans les années 1970) produit une clé phonétique comprenant 14 chiffres tirés du nom de famille. Suivant la méthode IBM, les noms « Rogers » et « Rodgers » seraient tous deux codés « 04740000000000 », alors que le système Soundex codait ces deux noms différemment (R262 et R326 respectivement). Bien que la méthode IBM puisse être perçue comme une amélioration comparativement à la méthode Soundex, des problèmes persistent en ce qui concerne le regroupement et la séparation des noms. Un problème posé par la méthode IBM est qu'il est impossible de savoir quoi que ce soit au sujet des noms, à moins que l'algorithme soit connu. Contrairement aux méthodes Soundex, Henry et NYSIIS, les codes IBM n'indiquent pas la première lettre du nom à coder. Cette information est fort utile pour certains utilisateurs. La méthode du IBM Alpha Inquiry System est résumée à l'annexe E (Moore *et al.*, 1977).

Une autre méthode pour regrouper les noms a été élaborée dans les années 1970 pour le compte de Western Air Lines. Cet algorithme de chiffrement a été conçu pour répondre aux besoins particuliers de la société aérienne. Les algorithmes comme ceux de la méthode Western Air Lines tiennent compte des données auxiliaires qui sont stockées dans les fichiers de données. La méthode de codification est présentée à l'annexe F. La démarche de la méthode de Western Air Line est unique en soi, mais elle semble avoir réglé ce problème. Au lieu d'élaborer une toute nouvelle méthode, on aurait pu changer ou adapter les méthodes de chiffrement existantes en s'appuyant sur les caractéristiques propres à la population visée ou sur d'autres données recueillies (Moore *et al.*, 1977).

Plus près de nous, des méthodes culturellement adaptées de jumelage des noms sont en cours d'élaboration. Lorsqu'il y a beaucoup de noms composés, de variantes d'orthographe et d'accents, les algorithmes de chiffrement standards sont inadéquats. La société LAS aux États-Unis commercialise un logiciel et des bibliothèques nominatives qui appliquent les règles linguistiques pour convertir l'orthographe des noms en transcription phonétique des prononciations possibles. Le logiciel applique ensuite les principes de la phonétique articulatoire pour jumeler les noms (p. ex., les noms « Leigh » et « Li »). La société a créé un logiciel qui permet de classer les noms en fonction de l'appartenance ethnique. Le traitement spécialisé des noms arabes, hispaniques, mandarins et anglo-européens peut être effectué, et des outils et des bibliothèques pour d'autres ethnies sont en cours d'élaboration. (Nota : bien qu'il soit question de ce logiciel dans le présent rapport, les documents n'ont fait mention d'aucune évaluation ou d'applications particulières de son utilisation.)

Un exemple d'une récente amélioration apportée au système Soundex visant la codification des noms ethniques particuliers est le système **Daitch-Mokotoff Soundex**. Ce système a été créé par Randy Daitch et Gary Mokotoff dans les années 1990 afin de coder les noms de famille slaves et yiddish. Le système a également apporté des précisions à la méthode Soundex initiale. La structure du code D-M Soundex comprend six chiffres, et des zéros sont ajoutés aux noms qui n'ont pas suffisamment de lettres pour former un code de six chiffres. Un tableau de codification détaillé est utilisé pour coder chaque lettre du nom de famille. Le tableau de codification Daitch-Mokotoff est présenté à l'annexe G.

D. Collecte d'identificateurs personnels

1. Facteurs de la qualité des données

Plusieurs éléments permettent de déterminer ce qui constitue des données de bonne qualité, dont l'exactitude, l'opportunité, la pertinence, la cohérence et l'interprétabilité. Pour obtenir des données de la meilleure qualité possible, ces éléments doivent être équilibrés. Il se peut, par exemple, qu'on mette beaucoup de temps pour produire les meilleures données possibles, ce qui signifie que les statistiques produites ne seront plus aussi pertinentes.

Pour assurer des données de qualité, il faut poser les bonnes questions et faire preuve d'observation. L'enregistrement des données démographiques dans le milieu policier doit tenir compte de certaines questions, telles que l'usage de pseudonymes, par exemple. Un sténographe judiciaire ne doit pas présumer de l'orthographe d'un nom. Pour que les données qui figurent au fichier statistique soient véridiques, il importe de bien traiter ou gérer ces données. De façon typique, des lignes directrices déterminent les données à recueillir et les procédures à suivre. Des documents complets et à jour sont produits à cette fin. Les situations qui se présentent et qui ne sont pas abordées dans les lignes directrices ou les procédures normalisées peuvent causer des problèmes et empêcher l'utilisation de ces données plus tard.

Qu'il s'agisse d'un recensement complet ou d'un échantillonnage restreint des enregistrements, les problèmes liés à la qualité des données s'intensifieront à moins que des méthodes soient adoptées pour réduire les erreurs. L'enregistrement et la saisie exacts des données sont essentiels afin d'assurer que les données conservent leur valeur à l'avenir. Les erreurs de transcription et de saisie des données peuvent être mineures ou causer des problèmes majeurs en ce qui concerne leur utilité. De nombreux fichiers et bases de données stockent des données qui ne se présentent pas sous leur forme initiale ou brute. Des programmes informatisés de mise en forme et d'imputation des données sont couramment élaborés et appliqués aux données d'entrée. Ces programmes font en sorte que les valeurs des données correspondent à une échelle préétablie, que ces données sont cohérentes d'un point de vue logique ou que les données à blanc sont assignées une certaine valeur. Les processus de mise en forme et d'imputation augmenteront l'utilité des données en réduisant certains types d'erreur. Toutefois, ces programmes peuvent également introduire des erreurs dans les données.

La clé de la qualité des données est de faire en sorte que celles-ci sont saisies à la source de façon exacte et cohérente. On recourt ainsi beaucoup moins aux processus de mise en forme et d'imputation des données, et les coûts de vérification connexes sont réduits.

Le niveau de confiance dans la qualité des données découle de leur traitement à différents intervalles. Bien que le milieu opérationnel puisse ne pas se prêter à la collecte des données statistiques, l'utilité de celles-ci est uniquement tributaire du personnel au sein des organismes qui recueillent les données les plus exactes possibles. Il ne sert à rien d'analyser les données à répétition si les données initiales sont viciées. Des renseignements exacts seront produits suivant l'exactitude des données recueillies au départ, et des pratiques de gestion des données doivent être en place pour que celles-ci parviennent aux analystes statistiques.

2. Facteurs de confidentialité

Les notions d'identification personnelle unique et de confidentialité des renseignements personnels sont habituellement conflictuelles. Cela est particulièrement vrai lorsque le couplage des enregistrements est envisagé. On estime confidentielles les données d'enquête et administratives recueillies sur des particuliers, et seules les personnes autorisées peuvent les consulter. Il n'est pas nécessaire aux fins d'analyses et de rapports statistiques que les indicateurs personnels soient compris dans les fichiers de données. Par contre, dans un milieu opérationnel, il se peut que cela soit nécessaire.

Il y a de nombreuses utilisations statistiques potentielles des données de justice qui se fondent sur l'intégration des fichiers de données sur les particuliers. Au lieu d'utiliser un nom, il suffit d'utiliser un numéro d'identification. Ce qui importe ici, c'est qu'un seul numéro ou code soit assigné au même individu. Cela vaut pour les études transversales, longitudinales ou éventuelles. La capacité de s'appuyer sur un identificateur unique est un important facteur dont il faut tenir compte au fil du temps parce que les particuliers peuvent faire l'objet de nombreuses saisies dans les systèmes de justice.

Au Canada, il est rare que les données soient produites au niveau des particuliers. Statistique Canada et d'autres organismes gouvernementaux tâchent d'assurer que les renseignements sur des particuliers, entreprises ou établissements donnés ne sont pas présentés ou induits de façon certaine. Certaines méthodes, telles que la suppression des données, l'arrondissement aléatoire, la perturbation des cellules, le regroupement des données, etc. doivent être utilisées afin d'assurer la confidentialité des données.

Dans les enquêtes à base de microdonnées du CCSJ, les noms d'entrée fournis par les répondants sont chiffrés selon les méthodes Russell-Soundex ou Henry. Chacune de ces méthodes permet de créer une version codée des noms qui figurent dans le fichier de données. Les variantes orthographiques des noms peuvent produire un code différent. Ces méthodes de chiffrement ne fonctionnent que dans un sens, c'est-à-dire qu'une fois qu'un nom de famille est codé, le code ne peut être converti dans le nom initial. Cette méthode ne peut pleinement garantir la confidentialité du nom d'un particulier. De façon générale, plus gros sont les fichiers de données et plus communs les noms d'entrée, plus les méthodes de chiffrement assureront la confidentialité. D'un autre côté, la capacité de distinguer les différents particuliers aux fins du couplage des enregistrements ou de l'intégration des données peut poser problème.

En raison de la nécessité, d'une part, d'assurer et de maintenir la confidentialité et, d'autre part, de faire le suivi ou le couplage de renseignements sur les particuliers, différentes méthodes de couplage des enregistrements ont été créées. Ces méthodes s'appuient généralement sur des théories statistiques valables et ont permis d'obtenir les résultats escomptés. La section suivante donne un aperçu de certaines démarches communes de couplage des enregistrements. Elles s'appliquent au couplage des enregistrements sur les particuliers auxquels un identificateur personnel unique ou non unique a été assigné. Les méthodes peuvent s'appliquer à presque tous les domaines dans lesquels la collecte des données au niveau des particuliers est réalisée.

3. Couplage des enregistrements

1. Démarche de base

Le concept du couplage des enregistrements est plutôt élémentaire. Une procédure est élaborée afin de jumeler ou d'intégrer des données tirées d'un ou de plusieurs fichiers qui se rapportent à une même entité. Le « suivi » s'applique également au couplage des enregistrements, bien qu'il concerne habituellement l'identification des entités dans un même fichier de données.

Le couplage d'enregistrements type comprend les cinq étapes suivantes (Newcombe *et al.*, 1992) :

1. Établir des correspondances exactes entre deux fichiers.
2. Créer des « cases », « blocs » ou « groupes » d'enregistrements semblables qui n'ont pas été jumelés.
3. Créer des paires d'enregistrements pour lesquelles il est possible d'établir un couplage à partir des cases.
4. Pondérer ou évaluer les couplages probables pour chacune des paires.
5. Calculer les valeurs de pondération seuil et classer chaque paire d'enregistrements.

Par jumelage exact, on entend le regroupement de deux enregistrements qui satisfont tous les critères en matière de couplage des enregistrements. En ce qui concerne les données qui ont un identificateur unique, la procédure de couplage porte sur ce critère. Tous les enregistrements pour lesquels un jumelage exact a été établi suivant cette seule variable sont jumelés. Le numéro SED est un exemple de variable unique qui produit un jumelage exact. Toute différence par rapport à l'identificateur unique signifie que le couplage peut être réalisé ou non. Une fois que des enregistrements ont été jumelés, on passe à l'étape 2.

Puisque les fichiers de données peuvent être volumineux, des cases de données sont créées. Ces cases s'appuient habituellement sur des variables de base, telles que les noms chiffrés, le sexe, la date de naissance ou des données géographiques. Ensuite, chacun des enregistrements pour lesquels un jumelage exact n'a pas été établi est placé dans une case. Cette démarche simplifie le processus de couplage puisqu'on tente ainsi d'établir des liens au sein des cases et non dans le fichier en entier. Cette stratégie permet de considérablement réduire le nombre de combinaisons pour lesquelles on tente d'établir un jumelage. Avant d'utiliser les variables se rapportant aux cases, il faut au préalable s'assurer de leur qualité et les comprendre.

Ensuite, on compare les enregistrements dans les cases en utilisant une méthode qui peut comprendre un certain nombre de variables. À titre d'exemple, si dans une case on a déterminé que la variable est le sexe, le nom chiffré, la date de naissance ou la date de l'infraction peuvent être des variables du couplage. La qualité des données d'entrée de ces variables permettra d'effectuer le couplage de deux enregistrements apparentés. Habituellement, plusieurs essais sont élaborés et évalués avant qu'une méthode de couplage particulière soit utilisée.

Suivant les résultats du couplage, les pondérations ou probabilités associées à chaque paire d'enregistrements dans une même case sont calculées. Plus élevée est la valeur de la pondération, plus il est probable que deux enregistrements soient jumelés. De façon générale, les valeurs de pondération sont fondées sur les probabilités. Une valeur de pondération de 1,0 indique que deux enregistrements se rapportent de fait à une même entité. Des valeurs de pondération négatives peuvent également être obtenues dans certaines méthodes de couplage. Ces valeurs servent à indiquer qu'il n'y a définitivement pas de correspondance entre deux enregistrements.

Les valeurs de pondération exprimées en binits peuvent également être utilisées pour certaines applications du couplage des enregistrements. Une pondération en binits exprime le degré de concordance ou de discordance entre les deux variables faisant l'objet du jumelage. Une discordance extrême peut donner lieu à une pondération en binits négative. C'est l'analyste qui détermine les pondérations en binits, et il peut en résulter que la date de naissance aura le plus de poids en ce qui concerne le jumelage (10 binits) et la variable sexe donnera un pointage de 2 binits seulement. À titre d'exemple, deux enregistrements comportant deux codes différents liés au sexe pourraient obtenir un pointage en binits de -5. Deux enregistrements ayant un pointage binit élevé seraient considérés comme apparentés (Gill, 1997).

La dernière étape du processus de couplage d'enregistrements est l'examen des valeurs de pondération attribuées à chaque paire jumelée. Des valeurs seuil qui sont fondées sur le degré de confiance nécessaire pour produire un couplage sont établies. Les valeurs seuil comportent des points limites. À un point limite, on distingue entre les enregistrements dont le jumelage est presque certain et les enregistrements susceptibles d'être jumelés. Le second point limite distingue les enregistrements susceptibles d'être jumelés et ceux pour lesquels une concordance est peu possible.

Le classement final, effectué à l'étape 5, se fait suivant l'une des trois possibilités suivantes : les deux enregistrements sont jumelés, les deux enregistrements ne sont pas jumelés, un jumelage est possible. De façon générale, les enregistrements de ce dernier classement posent les plus grands problèmes parce qu'il faut beaucoup de temps pour réaliser le couplage. Parfois, pour déterminer si les enregistrements du groupe « jumelage possible » sont de fait apparentés, il faut communiquer avec le répondant et se reporter aux documents d'origine. Il faut alors modifier les données et exécuter de nouveau le couplage.

2. Facteurs et applications liés au couplage des enregistrements

Suivant la terminologie du couplage des enregistrements, une erreur de type I se produit lorsqu'un couplage « faux-positif » est réalisé. Cela signifie que deux enregistrements ont été jumelés alors qu'en fait ils se rapportent à deux entités différentes. Une erreur de type II se produit lorsqu'un couplage « faux-négatif » est réalisé. Cela signifie que deux enregistrements qui se rapportent à une même entité n'ont pas été jumelés. L'établissement de valeurs seuil déterminera le nombre d'erreurs de types I et II qui sont commises. Si l'on insiste pour que les seuils d'erreur de types I et II soient minimes, il en découlera un grand nombre de jumelages d'enregistrements possibles. Comme il a été mentionné, il sera difficile d'attribuer correctement les enregistrements de ce groupe soit dans le groupe des enregistrements jumelés ou non jumelés.

Un certain nombre de documents abordent la question du couplage des enregistrements. Le mémoire précurseur de Fellegi et Sunter (1969) présentait les bases statistiques du couplage des enregistrements. Le document de Newcombe, Fair et Lalonde de 1992 donnait l'historique du couplage probabiliste d'enregistrements et indiquait certains résultats empiriques obtenus en utilisant la démarche de Fellegi et Sunter. Bon nombre des applications logicielles utilisées aujourd'hui pour réaliser le couplage des enregistrements s'appuient sur les démarches présentées dans ces deux documents. Si l'on vise seulement des jumelages exacts, il n'est pas nécessaire de recourir à la méthode en cinq étapes. On peut utiliser différents logiciels d'usage commun (SAS, MS Access, etc.) pour réaliser des jumelages exacts.

Statistique Canada se sert du Système généralisé de couplage des enregistrements (SGCE) pour effectuer le couplage d'enregistrements à grande échelle en employant une méthode de jumelage probabiliste (non exacte). Le SGCE a été utilisé pour réaliser la contre-vérification des enregistrements du Recensement de la population et des logements de 1996, le couplage des enregistrements de l'Enquête sur la santé et des dossiers administratifs de la Base canadienne de données sur la mortalité, de la Base canadienne de données sur le cancer et de la Base canadienne de données sur la santé, ainsi que le couplage des données agricoles afin de créer le Registre central des fermes. On s'est également servi du SGCE pour faire le couplage des dossiers judiciaires maintenus par le CCSJ. La version actuelle du SGCE (4.1) est exploitée dans un environnement client-serveur au moyen d'un compilateur Oracle et C. Le logiciel peut également être exploité sur un ordinateur personnel ou un poste de travail équipé du système d'exploitation UNIX. Le SGCE convient particulièrement aux applications dans lesquelles les dossiers à jumeler ne comprennent pas d'identificateur unique (Fair, 1997). On aurait intérêt à rappeler que les versions antérieures du SGCE s'appelaient CANLINK et SGCEI (Système généralisé de couplage d'enregistrements itératifs).

Le choix des variables utilisées dans le couplage des enregistrements des particuliers est très important d'un point de vue de la qualité des données et de leur traitement. Parce que les identificateurs personnels tels que le nom de famille et les prénoms ne sont pas habituellement disponibles, le code du nom chiffré sert habituellement de variable de case. D'autres variables communes sont le sexe, la date de naissance, la date de l'infraction et l'infraction commise. Les variables géographiques peuvent également servir dans le cadre des études sur le couplage d'enregistrements qui portent sur une région géographique étendue. De façon générale, plus grand est le nombre de variables utilisées pour réaliser le couplage d'enregistrements, plus il y aura de problèmes. Cela est surtout attribuable aux questions liées à la qualité des données et aux délais d'exécution des systèmes. Les couplages d'enregistrements typiques se rapportant à des fichiers judiciaires comprennent les noms chiffrés, la date de naissance et la date de l'infraction. Des données supplémentaires peuvent être utilisées pour confirmer les couplages. Celles-ci ne sont pas nécessaires pour effectuer les couplages comme tels mais servent à obtenir des renseignements supplémentaires sur les deux enregistrements, surtout en ce qui concerne les enregistrements du groupe « jumelage possible ».

Applications

Le couplage d'enregistrements a connu un certain succès en utilisant les statistiques sur la santé de Statistique Canada. Par le passé, on a déjà accompli du travail considérable afin de regrouper les données sur la santé en utilisant les techniques du couplage d'enregistrements. Deux exemples de travaux antérieurs qui ont été entrepris dans le domaine de la santé ainsi qu'un aperçu de certains travaux en cours sont présentés ci-dessous.

Le premier exemple se rapporte au couplage des données du Registre canadien du cancer (RCC) et de la Base canadienne de données sur la mortalité (BSCM). Le RCC est une base de données longitudinales sur des particuliers qui contient tous les renseignements sur les cancéreux et leurs tumeurs. Cette base de données existe depuis 1992. Un important volet du RCC est le module de confirmation des décès. Ce module a été conçu pour utiliser les enregistrements sur les décès de la BCDM afin de confirmer le décès des cancéreux inscrits dans le RCC qui s'est produit durant une période prédéterminée. Deux types de couplage d'enregistrements ont été utilisés en raison de différences entre les dossiers provinciaux d'enregistrement des décès et les données de la BCDM. Le jumelage direct a été réalisé en utilisant la date de décès, la province, le territoire ou le pays où le décès a eu lieu et le numéro d'enregistrement du décès. Pour effectuer les jumelages probabilistes, on a eu recours au code NYSIIS du nom de famille de la personne décédée et du nom de famille du père de la personne décédée. Le couplage des enregistrements du RCC et de la BCDM permet de calculer le taux de survie des personnes atteintes du cancer, et facilite les études épidémiologiques en utilisant la cause du décès (LaBillois *et al.*, 1997).

Dans le second exemple, trois fichiers de données ont été fusionnés à une fin précise. L'objectif visé par le couplage des enregistrements était s'assurer une compréhension accrue des mécanismes qui exercent une influence sur l'utilisation des services de santé au Manitoba. En utilisant les données du Recensement de la population et des logements de 1996, de l'Enquête sur la santé et les limitations d'activité (ESLA) de 1986-1987 et de Santé Manitoba, l'analyse de l'association entre les caractéristiques sociodémographiques et de l'utilisation des soins de santé et médicaux dans la province a été réalisée. Dans le présent exemple, un taux de jumelage de 74 % a été affiché en ce qui concerne les ménages particuliers. Une étude ultérieure a analysé le taux de concordance global des enregistrements jumelés, qui s'élevait à près de 96 %. Les identificateurs qui ont servi à jumeler les personnes dans le cadre de cette étude étaient le sexe, la date de naissance, le mois de naissance et le code postal. Il se peut qu'on ait tenu compte de l'âge à une étape ultérieure du processus. Les noms et adresses particuliers n'ont pas été utilisés afin d'assurer la confidentialité et la protection des renseignements personnels (Houle *et al.*, 1997).

L'élaboration du Carnet de route de l'information sur la santé a récemment été amorcée. Il s'agit d'une initiative de quatre ans visant à moderniser le système d'information sur la santé du Canada. Par cette initiative, on tente de combler les écarts qui existent actuellement et de regrouper des données d'une variété de sources afin de permettre aux chercheurs du domaine de la santé d'accéder à un vaste ensemble de données.

L'une des sources de données est la base de données axées sur la personne. Cette base de données est dérivée des fichiers du système de morbidité hospitalière (SMH). Ces fichiers sont produits tous les ans et comprennent des renseignements sur les départs des hôpitaux. Ces fichiers de données sont axés sur les événements, et il revient donc aux utilisateurs de regrouper les dossiers qui se rapportent à une même personne. Le projet de la base de données axées sur la personne transforme les données du SMH, au moyen d'une série de couplages, en données axées sur les personnes. Les dossiers de personnes comptant plus d'un départ d'hôpital peuvent ainsi être regroupés. Ce genre de couplage donne de l'information utile aux chercheurs. En raison de la différente structure des dossiers provinciaux et territoriaux sur la morbidité, il est très important d'effectuer un prétraitement (y compris la normalisation) avant d'entreprendre le couplage.

Les récentes percées qui touchent les logiciels ont permis d'effectuer facilement des couplages d'enregistrements exacts et des jumelages probabilistes sur une base élémentaire. Newcombe (1988) aborde de nombreuses questions sur le couplage d'enregistrements, et de nombreuses applications (et certaines théories) ont été abordées dans le Proceedings of an International Workshop and Exposition on Record Linkage Techniques, qui s'est tenu à Washington en mars 1997.

Ces deux ouvrages contiennent de nombreux autres renvois à des aspects particuliers du couplage d'enregistrements. On y discute en profondeur des questions telles que le choix des variables aux fins du jumelage, y compris les noms et le chiffrement des noms, ainsi que le choix des règles de jumelage. La communication de Scott Meyer (Proceedings) « Using Microsoft Access to Perform Exact Record Linkages » porte sur l'utilisation des fichiers judiciaires du CCSJ. Les méthodes et les taux de couplage sont présentés aux fins du jumelage des enregistrements de l'Enquête sur les tribunaux de juridiction criminelle pour adultes et des dossiers du Programme de déclaration uniforme de la criminalité 2.0 de la ville de Regina. Bien que la communication soit restreinte en ce qui concerne la portée de l'étude, la région géographique et les genres d'infraction, on y aborde de nombreuses questions relatives au couplage d'enregistrements. Les variables utilisées dans le cadre du couplage comprenaient la codification Russell-Soundex du nom de famille de l'accusé, la date de naissance, le sexe, la date de l'infraction et le genre d'infraction commise. Sept différentes stratégies de couplage ont affiché des taux de couplage variant de 62 à 85 %.

4. Facteurs statistiques et facteurs opérationnels

Les exigences associées au couplage des fichiers de données varient suivant les besoins statistiques et opérationnels. Souvent, les données recueillies aux fins d'analyses peuvent ne pas être disponibles ou disponibles seulement sous une forme différente de celle qui est requise. L'exhaustivité et l'exactitude des données peuvent également comporter différents niveaux d'importance. Par exemple, des renseignements peuvent être recueillis sur l'âge d'un particulier pour des motifs opérationnels, mais la date de naissance est plus pertinente aux fins du couplage des enregistrements et de l'analyse des données. Si la date de naissance est fournie par les répondants à leurs propres fins opérationnelles, sa présentation peut ne pas être normalisée à des fins statistiques.

Un problème commun lié à la collecte de données statistiques est que les documents sur le mode de collecte des données et sur ce que les données représentent peuvent ne pas être consultables ou à jour. Il arrive fréquemment que des valeurs codées soient utilisées à titre d'indicateurs au lieu d'une description. La variable « sexe » est souvent codée sous trois différentes valeurs, c.-à-d. « masculin », « féminin » et « inconnu ». À moins que l'analyste ait une description de ces codes, certaines ambiguïtés peuvent persister. Cette difficulté s'aggrave lorsqu'une variable comporte plus de trois résultats.

Les changements apportés aux procédures et aux systèmes opérationnels peuvent avoir une incidence sur une application statistique. À titre d'exemple, la collecte de certaines variables peut cesser ou encore celles-ci peuvent être modifiées à un moment donné. Les utilisateurs des données statistiques peuvent ne pas être au courant de ces changements, et des difficultés peuvent se produire au moment de traiter ou d'analyser les données. Une difficulté particulière en ce qui concerne les données judiciaires est le manque d'uniformité dans le signalement de l'appartenance ethnique des contrevenants. Bien que cette variable puisse être très utile pour identifier des particuliers, cette information est en elle-même de nature délicate et il faut faire preuve de discernement au moment de la recueillir et de la signaler. L'exactitude des données sur l'appartenance ethnique peut poser problème suivant la façon dont la question a été posée ou la réponse donnée. En outre, les données peuvent être induites et ne pas découler d'une question précise, ce qui peut causer des erreurs.

On aurait intérêt à normaliser les données d'entrée aux fins statistiques. Cela peut se faire au moment de recueillir les données ou de leur traitement par un analyste statistique. La majorité des administrateurs opérationnels assurent le contrôle de fichiers et de bases de données particuliers. Un poste de police dans une région donnée du pays peut assigner un code numérique au sexe d'un contrevenant, tandis qu'un autre poste lui assignera un code alphabétique. En outre, un troisième poste peut indiquer le sexe dans la base de données en tant que « masculin » ou « féminin ». Ces variantes de codification et de stockage des données ont une incidence sur l'identification des particuliers au fil du temps ou par les différents organismes. Une importante question liée à la normalisation des données est les coûts associés. Les coûts de la modification des programmes informatiques actuels aux fin de la saisie et de la mise en forme des données peuvent être considérables. Il n'est pas très compliqué de modifier le code d'une variable dans un fichier, mais cela peut prendre du temps. Le personnel opérationnel peut ne pas voir l'utilité d'apporter ces changements, bien qu'ils comportent d'importants avantages pour l'analyste statistique.

Les données d'entrée non normalisées peuvent entraîner des erreurs durant le traitement. Le fait de devoir composer avec des codes inconnus ou problématiques pose des difficultés pour l'analyste et retarde son travail. Les programmes informatiques mis au point par les analystes permettent habituellement de cerner les difficultés posées par les données d'entrée, mais la mise en forme peut ne pas les corriger. Les erreurs de traitement des données démographiques peuvent prévenir le jumelage d'enregistrements ou produire des erreurs de types I et II. Par exemple, un enregistrement qui ne comporte pas de code de la variable sexe peut signifier que toutes les données se rapportant à cet enregistrement seront limitées. Les tableaux créés et dont la variable de contrôle est le sexe ne comprendraient pas l'enregistrement mentionné ci-dessus à moins que la catégorie « sexe-inconnu » soit utilisée. Ce genre de difficulté réduit l'utilité des données et les analystes ne peuvent habituellement pas les régler de façon certaine. À moins de communiquer avec le répondant, tout ce que l'analyste peut faire, c'est de voir si l'enregistrement peut être jumelé à un autre enregistrement sur la même personne et se servir du code sexe de cette source. Il faut beaucoup de temps pour faire cela, et ce n'est pas une pratique commune des enquêtes judiciaires du CCSJ. Cette technique d'imputation peut être utilisée au besoin, mais les valeurs induites ou imputées ne sont pas nécessairement correctes.

5. Conclusion

L'objet du présent rapport était de donner un aperçu des questions touchant les identificateurs personnels. L'utilisation de ces identificateurs dans le milieu judiciaire est très intéressante, bien que le rapport n'ait pas été rédigé pour ce domaine. Il n'était pas également prévu dans le présent rapport d'examiner à fond tous les aspects qui peuvent être liés aux questions des identificateurs personnels, des méthodes de chiffrement, du couplage des enregistrements, de la confidentialité des renseignements, etc. La majorité, sinon toutes, les questions abordées dans le rapport ont été traitées pendant de nombreuses années par des personnes partout dans le monde. Leur travail a été documenté dans des monographies détaillées. Le présent rapport a été rédigé en tenant compte de considérations statistiques au lieu d'opérationnelles. Il faut donc aborder aussi ces nombreuses questions à un point de vue opérationnel.

Les conclusions tirées se rapportent aux principales questions touchant l'utilisation des identificateurs personnels. Bon nombre des ouvrages cités sont d'excellentes sources de renseignements pour mener des recherches plus poussées sur des sujets donnés. Internet est également une bonne source d'information sur l'historique, l'élaboration et l'utilisation des procédures se rapportant aux domaines dont il a été question dans le présent rapport. Certains des sites Internet visités durant la rédaction du rapport sont indiqués à titre de référence. On recommande de vérifier l'information sur ces sites avant de l'utiliser plus avant.

Il est fort commun d'utiliser les noms aux fins de l'identification personnelle, mais cela peut poser des problèmes. Pour des questions de confidentialité et de qualité des données, il est préférable d'employer des méthodes de chiffrement des noms. Bien que cette option ne soit pas très utile à un point de vue opérationnel, elle peut contribuer à l'identification précise de particuliers à partir de fichiers de données statistiques. Le fait d'utiliser un nom de famille chiffré irréversible ainsi que d'autres renseignements démographiques permet le jumelage ou le suivi de particuliers.

En outre, l'utilisation des mesures biométriques dans le contexte de l'identification précise de personnes est de plus en plus usitée. La méthode la plus commune est sans doute les empreintes digitales et d'autres méthodes, telles que les profils d'ADN, et l'archivage électronique de photos de criminels aide également à jumeler des renseignements sur les particuliers. Bien que l'utilisation de ces techniques soit évidente en ce qui concerne le système de justice, ces renseignements peuvent également être utilisés à des fins statistiques. Puisque ces méthodes récentes permettent d'identifier précisément les particuliers, l'utilisation de ces variables à titre de « données » peut permettre aux analystes de jumeler les données beaucoup plus rapidement et exactement que ce que permettent les données démographiques traditionnelles.

Pourtant, certaines complications associées au jumelage statistique se produisent lorsque des correspondances exactes ne peuvent être établies pour des particuliers. Il importe de recueillir et de traiter les meilleures données possible afin de pouvoir réaliser le plus grand nombre de jumelages exacts. Les principes statistiques servent à élaborer des procédures de jumelage d'enregistrements qui devraient concorder parfaitement, mais qui ne concordent pas. De nombreux algorithmes existent afin d'augmenter le nombre de données jumelées ou regroupées, de façon à ce que des conclusions plus fiables puissent être tirées dans le cadre des analyses de données. Les procédures sont fort bien documentées et appuyées de preuves empiriques afin d'indiquer les réussites et les difficultés connexes. Des logiciels généralisés ou destinés à des applications précises sont largement disponibles afin d'exécuter des couplages d'enregistrements probabilistes. L'actuel Système généralisé de couplage des enregistrements (V3) de Statistique Canada est le résultat de plus de 20 ans de recherche et de développement, et est utilisé dans un grand nombre d'études importantes et pour créer de grosses bases de données.

Il importe de recueillir des données sources de bonne qualité afin d'obtenir de bonnes données opérationnelles et statistiques. Bien que l'utilisation des données à ces deux fins diffère, la question de la qualité des données reste la même. Le développement de bases de données judiciaires intégrées à des fins statistiques ne diffère pas tellement des systèmes qui ont été conçus dans le milieu judiciaire aux États-Unis et de l'INOIS en Australie de l'Ouest. Ces expériences rendent possible l'intégration des données au moyen d'identificateurs personnels dans de nombreuses applications, y compris celles du système de justice pénale du Canada.

Annexe A

Écarts par rapport au tableau 1 : Variables démographiques recueillies dans le cadre des enquêtes à base de microdonnées du CCSJ

Les notes suivantes précisent les variantes relatives aux variables démographiques signalées dans chacune des cinq enquêtes à base de microdonnées du CCSJ présentées au tableau A.1.

Programme de déclaration uniforme de la criminalité (DUC) (2.0 et 2.1)

1. Les quatre accusations les plus sérieuses sont enregistrées dans le cadre de l'enquête. La plus sérieuse est assignée le code MSO1, et la moins sérieuse, MSO4. Les autres types d'accusation ne sont pas signalés par les policiers dans le cadre du DUC 2.0/2.1.

2. Les noms de famille des personnes appréhendées sont chiffrés au moyen du code Russell-Soundex, sauf au Québec où la méthode Henry est utilisée.

3. Les variables de la date de naissance et de l'âge figurent dans les enquêtes du DUC 2.0 et 2.1. Le CCSJ attribue une variable liée à l'âge aux dates de naissance signalées par les répondants. En 1998, environ 1,9 % des enregistrements du DUC 2.0 ne comprenaient pas la date de naissance. L'âge du contrevenant avait été signalé pour ces enregistrements.

Enquête sur les homicides

Différentes versions de l'Enquête sur les homicides ont été tirées de la version initiale de 1961. Plusieurs modifications ont été apportées aux questions de nature démographique à compter de 1991. Dans les notes ci-dessous, on résume les différences dans les données démographiques recueillies avant 1991 et de 1991 à aujourd'hui.

1. Avant 1991, le nom de famille de l'accusé ne comptait que dix caractères. Deux autres caractères correspondaient au prénom. La version en vigueur de 1991 à aujourd'hui, comprend le nom de famille, les prénoms et initiales au complet.

2. Avant 1991, la date de naissance de l'accusé n'était pas disponible. Plutôt, l'âge de l'accusé au moment de l'appréhension était indiqué. La version ultérieure de l'enquête comprenait la date de naissance.

3. Avant 1991, les données relatives à l'événement ne comprenaient que le mois et l'année. Les version ultérieures incluaient la date de l'événement.

4. En ce qui concerne les homicides commis avant 1974, on ne distinguait pas s'il s'agissait d'un meurtre, d'un homicide involontaire ou d'un infanticide aux termes du *Code criminel du Canada.*

5. Les variables « état matrimonial », « situation professionnelle » et « appartenance à un groupe autochtone » ont toujours figuré dans l'Enquête sur les homicides.

6. Les variables « pays de résidence » et « profession » ont été ajoutées en 1991.

Enquête sur les tribunaux de la jeunesse (ETJ)

1. Au Québec, on signale le nom de famille et le prénom de l'accusé. Toutefois, le nom est chiffré dans les fichiers des bases de données de l'ETJ. Le code Henry à 7 caractères est utilisé au Québec [le code Soundex est utilisé dans les autres provinces].

2. Il arrive parfois que la date de naissance manque même après un suivi auprès du répondant (moins de 1 %).

Enquête sur les services communautaires et le placement sous garde des jeunes (SCPSGJ)

1. L'algorithme Soundex est utilisé pour chiffrer tous les noms. À l'heure actuelle, le Québec ne fournit aucune donnée dans le cadre de l'enquête SCPSGJ.

Enquête sur les tribunaux de juridiction criminelle pour adultes (ETJCA)

1. Les noms et prénoms des personnes sont chiffrés en employant le code Russell-Soundex. Toutefois, la méthode de chiffrement Henry à sept caractères est employée au Québec.

2. Les raisons sociales des entreprises sont également chiffrées au moyen des méthodes Soundex et Henry.

3. Lorsque le contrevenant est une entreprise, la variable sexe est codée « 3 » et la date de naissance, « 0 ».

4. Environ de 15 % à 20 % des codes liés au sexe ne figurent pas dans les dossiers des contrevenants du Québec.

Annexe B

Règles de codification du système Russell-Soundex

1. Conserver la première lettre du nom (habituellement le nom de famille) et omettre les lettres suivantes : A, E, H, I, O, U, W et Y [nota : le nom peut être le nom de famille seulement ou comprendre un prénom. En pareil cas, le prénom suit directement le nom de famille sans être séparé par un espace.]

Assigner les chiffres suivants aux autres lettres du nom :

Lettres	Codes
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

2. Si deux ou plus de deux lettres ayant le même code sont contiguës dans le nom (c.-à-d. avant qu'il ne soit codé), il faut les omettre toutes sauf la première.

3. Le nom codé prend la forme d'une lettre suivie de trois chiffres. Si le code compte moins de trois chiffres, des zéros sont ajoutés. S'il y a plus de trois chiffres, il faut supprimer la position 4 et les positions suivantes.

On effectue habituellement un prétraitement des noms à chiffrer au moyen du code Soundex et des autres méthodes décrites ci-dessous. Le prétraitement peut comprendre la suppression des traits d'union, des espaces, des lettres portant un accent, etc.

Exemples de codes Soundex :

Nom de famille	Soundex
Lloyd	L300
Ladd	L300
Harper	H616
Livingston	L152
Rogers	R262
Rodgers	R326
Ho	H000
Jackson	J250

Présentation des diffférents codes Soundex

La première lettre du code est l'une des 26 lettres alphabétiques.
Le premier chiffre est l'un des suivants : 0, 1, 2, 3, 4, 5, 6
Le second chiffre est l'un des suivants : 0, 1, 2, 3, 4, 5, 6
Le troisième chiffre est l'un des suivants : 0, 1, 2, 3, 4, 5, 6.

Il existe certaines combinaisons impossibles en ce qui concerne le code Soundex. Par exemple, les codes D306 et T061 sont invalides parce que les chiffres 1 à 6 ne peuvent suivre un zéro.

Suivant les règles ci-dessus, les différents codes Soundex possibles sont les suivants :

Il y a 26*7*7*7 ou 8 918 différents codes, les codes inacceptables exceptés. Les codes inacceptables peuvent prendre l'une ou l'autre des formes suivantes :

A00X pour X = 1, 2, 3, 4, 5, 6 (156 combinaisons)
A= A à Z

A0X0 pour X = 1, 2, 3, 4, 5, 6 (156 combinaisons)
A = A à Z

A0XY pour X = 1, 2, 3, 4, 5, 6 et
Y = 1, 2, 3, 4, 5, 6 (936 combinaisons)
A = A à Z

AX0Y pour X = 1, 2, 3, 4, 5, 6 et
Y = 1, 2, 3, 4, 5, 6 (936 combinaisons)
A = A à Z

Le nombre global de codes Soundex inacceptables est 2 184. En retranchant les codes inacceptables des 8 918 codes acceptables, cela donne 6 734 différents codes Soundex valides.

Annexe C

Règles de codification de la méthode Henry

Le code Henry utilisé dans les enquêtes à base de microdonnées du CCSJ compte sept lettres. Les cinq premières lettres se rapportent au nom de famille et les deux dernières, au prénom. On ajoute des blancs lorsque les noms sont trop courts pour donner un code complet.

La méthode Henry propose trois méthodes de codification du nom de famille et du premier prénom :

a) pour la lettre initiale

b) pour les lettres médianes (il faut effectuer le balayage de chaque consonne)

c) pour les consonnes à la fin du nom.

Dans la méthode Henry, les voyelles sont : A, E, I, O, U et Y.

Règles :

A. CONSONNES

1. La lettre initiale (PH ou CH) [H exceptée parce qu'elle est omise dans la méthode Henry]

* B, D, F, J, K, L, M, N, R, T, V sont retenues

La codification des autres lettres initiales est la suivante :

Lettre	Code
C devant A, O, U, L, R	K
C devant E, I, Y	S
CH devant une voyelle	C
CH devant L, R	K
G devant A, O, U, L, R	G
G devant E, I, Y	J
GN	N
H	Non existant
P devant H	F
P devant une autre voyelle	P
Q devant UE, UI, UY, E, I, Y	K
Q devant UA, UO, A, O	K
S	S [sauf dans Saint(e), Sainct(e), Sct(e), St(e) et Sain, est codé « X »].
W	V
X	S
Z	S

2. Les lettres médianes (2e, 3e et 4e positions du nom de famille)

a) Consonne simple entre deux voyelles

B, D, F, J, K, L, M, N, P, R, T, V sont retenues

Dans les autres cas, les règles suivantes s'appliquent :

Lettre	Code
C devant A, O, U	K
C devant E, I, Y	S
G devant A, O, U	G
G devant E, I, Y	J
H	Non existant
Q devant UE, UI, UY, E, I, Y	K
Q devant UA, UE, A, O	K
S	S [sauf dans Saint(e), Sainct(e) et Sain, qui sont codés « X » après DE].
W	V
X	S
Z	S

b) Groupe de consonnes

Les consonnes doubles sont considérées simples et sont représentées par une seule lettre (NN=N, MM=M et LL=L).

Les autres lettres sont codées comme suit :

Lettre	Code
C devant L, R	K
CH devant L, R	K
CH devant une voyelle	C
PH	F

S est muet devant toutes les consonnes (même H)

H est non existant dans tous les autres cas

Toute consonne autre que L ou R devant une consonne autre que L ou R est muette

L est muet devant M ou N

ST(E) et SCT(E) après DE sont codés « X »

3. Lettres finales

Lettre simple (précédée d'une voyelle) :

C, D, H, J, M, N, S, T, V, W, X, Z sont muettes

B, F, K, L, P sont retenues

Les autres cas sont codés comme suit :

Lettre	Code
G	G
Q	K
R précédé par E	muet
R précédé par une autre voyelle	R

b) Groupe de consonnes

Les consonnes doubles finales sont traitées comme des consonnes simples finales

Dans les autres cas, on procède comme suit :

Z = S

Un S à la fin d'un nom est considéré non existant, qu'il précède ou suive une ou plusieurs consonnes

H est toujours muet, sauf dans CH, qui est codé « C », et dans PH, qui est codé « F »

Lorsque R et P sont les avant-dernières lettres, la lettre finale est muette

L devant F ou M est muet, et F et M sont codés

CQ = K

Dans tous les autres cas, les dernières consonnes sont considérées muettes

B) VOYELLES

La voyelle initiale seulement est codée.

2. Lorsqu'une voyelle est suivie d'une consonne simple ou de deux ou plusieurs consonnes, dont la première n'est ni M ni N :

A, E, I, O et U sont retenues
Y = I

3. Lorsqu'une voyelle est suivie d'une ou de plusieurs consonnes, dont la première n'est ni M ni N :

Lettre	Code
A	A
E	A
I	E
O	O
U	E
Y	E

4. Une voyelle est suivie d'une autre voyelle :

1) Diphthongues : AE = E EI = E OI = O
AY = E EY = E OY = O
AU = O EU = U OU = O

2) Autres : Voir point 1

Exemples de codes suivant la méthode Henry (nom de famille) :

Nom de famille	Code
Clarke	Klrk
Cyr	Sr
St. Germain	Xjrm
Thivierge	Tvrj
Saint-Denis	Xdn
St. Denis	Xdn
Deschamps	Dc
Auerbach	Orbc
Rodgers	Rj
Rogers	Rj
Szamavitz	Ssmv
Montgomery	Mgmr
Ladd	L
Worthy	Vrt
Harper	Arp

Nota : Aucun des noms ci-dessus n'était suffisamment long pour produire un code à cinq caractères.

Annexe D

Règles de codification du NYSIIS

I. Si les premières lettres du nom sont :

MAC	remplacer ces lettres par MCC
KN	remplacer ces lettres par NN
K	remplacer cette lettre par C
PH	remplacer ces lettres par FF
PF	remplacer ces lettres par FF
SCH	remplacer ces lettres par SSS

2. Si les dernières lettres du nom sont :

EE	remplacer ces lettres par Y*
IE	remplacer ces lettres par Y*
DT ou RT ou RD ou NT ou ND	remplacer ces lettres par D* [* représente un blanc]

3. Le premier caractère du code NYSIIS est la première lettre du nom.

4. Suivant les règles, les caractères d'un nom sont balayés. Il s'agit en fait d'une boucle. Un pointeur est utilisé pour pointer la position visée dans le nom. À l'étape 4, il faut régler le pointeur pour qu'il pointe le deuxième caractère du nom.

5. Pour chaque position successive indiquée par le pointeur, seulement un des énoncés suivants s'applique :

   a. Si un blanc, passer à la règle 7.
      si la position est une voyelle (AEIOU)
      et s'il s'agit d'un E suivi d'un V, remplacer par AF
      sinon remplacer la position par A.

   b. Si la position est la lettre :
      Q, remplacer par la lettre G
      Z, remplacer par la lettre S
      M, remplacer par la lettre N

   c. Si la position est la lettre K
      et que la lettre suivante est N, remplacer par N
      sinon remplacer la position par C

   d. Si la position est la série de lettre :
      SCH, remplacer par SSS
      PH, remplacer par FF

e. Si la position est la lettre H et que la lettre précédente ou suivante n'est pas une voyelle (AEIOU), remplacer la position par la lettre précédente.

f. Si la position est la lettre W et que la lettre précédente est une voyelle, remplacer la position par la lettre précédente.

Si aucune des règles ci-dessus ne s'applique, la position conserve sa valeur.

6. Si la position est la même que la dernière lettre insérée dans le code, alors pointer la lettre suivante et passer à l'étape 5.

   Le caractère suivant du code NYSIIS est la position courante.

   Régler le pointeur pour qu'il pointe la lettre suivante.

   Passer à l'étape 5.

7. Si le dernier caractère du code NYSIIS est la lettre S, la supprimer.

   Si les deux derniers caractères du code NYSIIS sont les lettres AY, les remplacer par la lettre Y.

9. Si le dernier caractère du code NYSIIS est la lettre A, la supprimer.

Exemples de codes NYSIIS :

Nom de famille	Code NYSIIS
Worthy	Warty
Ogata	Ogat
Montgomery	Mantganary
Costales	Castal
Tu	T

Annexe E

Algorithme de codage de noms personnels de l'IBM Alpha Inquiry System

Les règles de codification produisent une clé phonétique à 14 chiffres comme suit :

1. Le premier caractère du code équivaut à la première lettre ou à la combinaison de lettres initiales et les valeurs suivantes sont accordées :

Lettre	Valeur
A	1
E	1
GF	08
GM	03
GN	02
H	2
I	1
J	3
KN	02
O	1
PF	08
PN	02
PS	00
U	1
W (WR exceptés)	04
Y	5

Un zéro est attribué si le premier caractère ou caractère(s) n'est pas indiqué dans le tableau ci-dessus.

2. Les lettres ou les combinaisons de lettres qui sont phonétiquement équivalentes sont attribuées une même valeur. On ne tient pas compte de la voyelle Y ou des lettres H et W.

Code	Lettre(s)
0	Z, S, CI, CY, CE, TS, TZ
1	D, T
2	N
3	M
4	R
5	L
6	J, SH, SCH, CH
7	C, G, K, Q, X, DG
8	F, V, PH
9	B, P

3. Il y a certaines exceptions aux règles de base énoncées à l'étape 2. Lorsque des lettres ou des groupes de lettres ont des sons différents, un deuxième ou un troisième passage en algorithme est exécuté.

Lettres	1er passage	2e passage	3e passage
CZ	70	6	0
CH	6	70	0
CK	7	7	6
C	7	7	6
K	7	7	6
DS	0	10	10
DZ	0	10	10
TS	0	10	10
TZ	0	10	10

Exemples de codes IBM :

Nom de famille	Code alpha
Rodgers	04740000000000
Rogers	04740000000000
Kant	02100000000000
Knuth	07210000000000

Annexe F

Démarche de notation de couplage de la Western Air Lines (1977)

La démarche élaborée par la société Western Air Lines est la suivante :

1. Omettre toutes les voyelles, sauf si une voyelle est le premier caractère du nom de famille.

2. Supprimer toutes les consonnes doubles en éliminant la seconde.

3. Le code produit comprend six caractères au maximum. Seuls les trois premiers et les trois derniers caractères encodés sont conservés.

4. La longueur de chaque paire de nom encodé est examinée (le nom d'entrée et le nom saisi dans le fichier de données [ce nom est désigné comme l'identificateur numérique personnel ou IPN]). S'il diffère par plus de deux caractères, aucune comparaison de similarité n'est réalisée. Une notation de similarité minimale acceptable est établie pour chaque paire de nom codé comme suit :

> Longueur maximale est de 4 caractères ou moins; notation de 5
> Longueur maximale est de 7 caractères ou moins; notation de 4
> Longueur maximale et de 11 caractères ou moins; notation de 3
> Longueur maximale est de 12 caractères; notation de 2.

5. On compare ensuite les noms codés. La comparaison procède de gauche à droite, caractère par caractère. Les paires de caractères apparentés sont supprimées. La comparaison se poursuit jusqu'à ce que chaque nom codé n'ait plus de caractères restants.

6. Les caractères non apparentés de chaque nom codé sont placés à la droite et la comparaison procède de droite à gauche. Une fois les comparaisons terminées, le nombre de caractères non apparentés dans le nom le plus long est soustrait de la valeur six, le résultat est la notation de similarité de cet INP.

7. Chaque INP dont la notation de similarité est égale ou supérieure à la notation minimale (voir l'étape 4) est considéré comme un jumelage potentiel d'un enregistrement entrant.

Par exemple : le nom d'entrée (HARPER) sera comparé au nom codé compris dans le fichier de données.

Étape 1. HARPER devient HRPR
Étape 2. HRPR reste HRPR
Étape 3. HRPR reste HRPR.

Si certains des enregistrements codés dans le fichier des données ressemblent à : HLDN, HRPR, HRPRD, HRP, HBLTWNS... donc

Étape 4. Les différences dans les longueurs entre HRPR et ceux indiqués ci-dessus sont : 0, 0, 1, 1, 3. Suivant la règle de la différence maximale énoncée à l'étape 4, HRPR serait comparé à chacun des premiers noms dans les fichiers de données existants. [Le résultat de la comparaison avec HBLTWNS serait 3].

Les étapes 5, 6 et 7 donnent les résultats suivants :

A. Les caractères non apparentés à HLDN sont LDN, et une notation de similarité de 6-3 ou 3 est accordée.
B. Il n'y a aucun caractère non apparenté à HRPR, et une notation de similarité de 6-0 ou 6 est accordée.
C. Le caractère non apparenté à HRPRD est D, et une notation de similarité de 6-1 ou 5 est accordée.
D. Le caractère non apparenté à HRP est R, et une notation de similarité de 6-1 ou 5 est accordée.

Puisque le nom d'entrée codé compte quatre lettres, tous les jumelages aux noms dans la base de données dont la valeur de similarité est 5 ou 6 seraient conservés pour comparaison ultérieure.

Ainsi, le nom d'entrée (HARPER) serait comparé aux enregistrements correspondant aux scénarios B, C et D indiqués ci-dessus.

Annexe G

Daitch-Mokotoff Système Soundex

1. Les noms sont codés et comprennent six chiffres, chacun correspondant à un son figurant dans le tableau de codification. Voir le tableau G.1.

2. Lorsque le nom ne compte pas suffisamment de sons pour former un code de six chiffres, il faut insérer des zéros pour faire un code à six chiffres. GOLDEN qui compte seulement quatre sons codés [G-L-D-N] est codé : 583600.

3. Les lettres A, E, I, O, U, J et Y sont toujours codées lorsqu'elles sont la première lettre du nom, par exemple : Alpert devient 087930. Dans toutes les autres situations, on n'en tient pas compte sauf lorsque deux de ces sons forment une paire et que la paire précède une voyelle, comme dans Breuer (791900), mais non dans Freud.

4. La lettre H en début du nom est codée, exemple Haber (579000), ou lorsqu'elle précède une voyelle comme dans Manheim (665600), sinon elle n'est pas codée.

5. Lorsque des sons contigus se combinent pour former un son plus large, ils prennent le chiffre codé du son plus large. Mintz n'est pas codé MIN-T-Z mais MIN-TZ (664000).

6. Lorsque des lettres contiguës ont le même chiffre codé, elles sont codées comme un seul son, par exemple : TOPF, qui n'est pas codé TO-P-F (377000) mais TO-PF (370000). Les exceptions à cette règle sont les lettres combinées MN et NM qui sont codées séparément, comme dans Kleinman qui est codé 586660 et non 586600.

7. Lorsqu'un nom de famille est combiné, il est codé comme s'il s'agissait d'un seul nom, tel que Ben Aron qui est codé comme Benaron.

8. Plusieurs lettres et lettres combinées peuvent se prononcer d'une ou de deux façons. Les lettres et lettres combinées CH, CK, C, J et RS sont assignées deux chiffres codés possibles.

Tableau G.1 *: Tableau de codification Daitch-Mokotoff*

« AC » signifie qu'aucun changement n'est apporté à la lettre

Lettre	Orthographe de rechange	En début du nom	Devant une voyelle	Toute autre situation
AI	AJ, AY	0	1	AC
AU		0	7	AC
A		0	AC	AC
B		7	7	7
CHS		5	54	54
CH	KH (5) + TCH (4)			
CK	K (5) + TSK (45)			
CZ	CS, CSZ, CZS	4	4	4
C	K (5) + TZ (4)			
DRZ	DRS	4	4	4
DS	DSH, DSZ	4	4	4
DZ	DZH, DZS	4	4	4
D	DT	3	3	3
EI	EJ, EY	0	1	AC
EU		1	1	AC
E		0	AC	AC
FB		7	7	7
F		7	7	7
G		5	5	5
H		5	5	AC
IA	IE, IO, IU	1	AC	AC
I		0	AC	AC
J	Y (1) + DZH (4)			
KS		5	54	54
KH		5	5	5
K		5	5	5
L		8	8	8
MN			66	66
M		6	6	6
NM			66	66
N		6	6	6
OI	OJ, OY	0	1	AC
O		0	AC	AC
P	PF, PH	7	7	7
Q		5	5	5
RZ, RS	RTZ (94) + ZH (4)			
R		9	9	9
SCHTSCH	SCHTSH, SCHTCH	2	4	4
SCH		4	4	4
SHTCH	SHCH, SHTSH	2	4	4
SHT	SCHT, SCHD	2	43	43

Lettre	Orthographe de rechange	En début du nom	Devant une voyelle	Toute autre situation
SH		4	4	4
STCH	STSCH, SC	2	4	4
STRZ	STRS, STSH	2	4	4
ST		2	43	43
SZSZ	SZCS	2	4	4
SZT	SHD, SZD, SD	2	43	43
SZ		4	4	4
S		4	4	4
TCH	TTCH, TTSCH	4	4	4
TH		3	3	3
TRZ	TRS	4	4	4
TSCH	TSH	4	4	4
TS	TTS, TTSZ, TC	4	4	4
TZ	TTZ, TZS, TSZ	4	4	4
T		3	3	3
UI	UJ, UY	0	1	AC
U	UE	0	AC	AC
V		7	7	7
W		7	7	7
X		5	54	54
Y		1	AC	AC
ZDZ	ZDZH, ZHDZH	2	4	4
ZD	ZHD	2	43	43
ZH	ZS, ZSCH, ZSH	4	4	4
Z		4	4	4

Exemples de codes Daitch-Mokotoff :

Nom de famille		Code
Auerbach	A-UE-R-B-A-CH	097500
Lipshitz	L-I-P-SH-I-TZ	874400
Lippszyc	L-I-P-P-SZ-Y-C	874400
Ohrbach	O-H-R-B-A-CH	097500
Shlamowicz	SH-L-A-M-O-W-I-CZ	486740
Szlamavitz	SZ-L-A-M-A-V-I-TZ	486740

(source : http ://www.jewishgen.org/infofiles/Soundex.txt)

Annexe H

Numéro du Système des empreintes digitales (canadien)

Le numéro SED utilisé au Canada compte six chiffres au maximum suivis d'un caractère alphabétique. Le caractère alphabétique est assigné de façon séquentielle en commençant par la lettre « A ».

Les numéros du SED sont assignés seulement aux particuliers, à la fois aux jeunes et aux adultes. Une personne qui entre de nouveau dans le système de justice pénale serait identifiée par le même numéro SED. Une fois qu'un numéro SED est entré dans la base de données du Centre d'information de la police canadienne, il y est conservé jusqu'à trois ans suivant l'avis de décès d'une personne ou lorsque la personne atteint 80 ans. Plusieurs exceptions touchent le critère des 80 ans, y compris si une personne purgeait une détention à perpétuité. Les numéros SED peuvent également être supprimés de la base de données dans diverses situations, y compris lorsqu'une personne est déclarée non coupable et qu'elle demande la suppression de son numéro SED.

L'utilisation du SED à titre de variable de jumelage statistique présente divers facteurs intéressants. La simplicité et l'unicité du numéro rendent possible un jumelage exact. On considère la qualité du numéro très bonne au moment de la saisie. Des problèmes peuvent se produire lorsque l'accusé tente d'altérer ses empreintes digitales. Le degré d'utilisation ou de couverture du numéro SED est sans aucun doute la principale faiblesse associée au jumelage des fichiers de données. De façon typique, les suspects d'infractions non criminelles ne sont pas soumis à la dactyloscopie. Ainsi, le couplage des renseignements sur un particulier entre les données que conservent la police et les tribunaux peut ne pas être possible en ce qui concerne de nombreuses infractions. Dans les études sur le récidivisme, cela peut constituer un facteur important.

On peut interroger la banque de données du Centre d'information de la police canadienne (CIPC) en utilisant seulement le SED. Si le numéro SED n'est pas disponible, d'autres variables telles que le nom, la date et le lieu de naissance, le sexe, la race et d'autres caractères physiques peuvent être utilisées. Dans le tableau H.1, les variables et leurs attributs qui servent à réaliser un couplage d'enregistrements dans la banque de données du CIPC sont présentés de façon sommaire (GRC, 1999).

Tableau H.1 : ***Variables de couplage utilisées dans la banque de données du CIPC (liste partielle)***

Variable	Longueur	Type	Notes
Prénom	10	Alpha	Une recherche peut porter sur cinq prénoms à la fois.
Nom de famille	25	Alpha	L'orthographe phonétique et les variantes orthographiques sont possibles. Une recherche peut également être faite sur des noms composés.
Sexe	1	Alpha	Féminin, masculin, inconnu
Date de naissance	8	Num.	AAAA.MM.JJ
Lieu de naissance	4	Alpha	Les provinces et territoires sont codés (c.-à-d. Ontario devient ONT.). Tous les États américains sont codés. Des codes pour le Royaume-Uni, l'Europe et d'autres pays sont disponibles.
Race	1	Alpha	Blanc, non blanc, inconnu
Âge	3	Num.	Suivant la date de naissance
Couleur des yeux	7	Alpha	Il y a au moins sept types/codes
Taille (métrique)	3	Num.	Suivant les renseignements les plus récents
Poids (métrique)	3	Num.	Suivant les renseignements les plus récents

Annexe I

Comparateurs de chaînes

Des erreurs au moment de saisir et d'enregistrer les noms dans un fichier de données occasionneront des difficultés lorsqu'on réalise le couplage des enregistrements. Suivant la nature de l'erreur typographique, il se peut que deux dossiers ne soient pas jumelés ou jumelés lorsqu'ils ne devraient pas l'être. L'utilisation de comparateurs de chaînes réduit l'incidence des erreurs typographiques qui peuvent modifier la longueur ou l'orthographe d'un nom. Les comparateurs de chaînes sont essentiellement des algorithmes appliqués aux données d'entrée, et l'extrant est une valeur numérique de 0 à 1. Plus élevée est la valeur, plus grande est la probabilité que les deux noms soient les mêmes.

En 1989, Jaro a introduit une méthode de comparaison de chaînes afin de quantifier le nombre d'insertions, de suppressions et de transpositions des lettres dans les noms. D'autres études empiriques ont suivi et ont modifié la méthode Jaro. À titre d'exemple, une précision apportée par Winkler en 1990 accorde une valeur supérieure aux premiers caractères concordants d'une chaîne. D'autres précisions ont tenu compte de la longueur des noms. Les comparateurs de chaînes ne s'appliquent pas seulement aux noms de personnes, mais peuvent aussi être utilisées pour les noms de rue, d'entreprise, etc.

Une autre méthode de comparaison de chaînes est fondée sur le principe des bigrammes. Un bigramme est simplement deux lettres consécutives dans une chaîne. Le mot « string » compte cinq bigrammes : « st », « tr », « ri », « in » et « ng ». La fonction associée aux bigrammes accorde une valeur entre 0 et 1, la valeur la plus élevée indique une plus grande concordance entre deux noms.

Des précisions sur la méthode et l'utilisation des comparateurs de chaînes sont présentées dans les documents de Porter et Winkler (1997) et de Winkler (1995).

Deux chaînes de noms de famille, une entrant et une ayant déjà fait l'objet d'un enregistrement dans une base de données, peuvent être comparées de diverses façons. Le tableau I.1 présente les résultats numériques de cinq différentes méthodes de comparaison des variantes de deux noms de famille. D'autres méthodes existent, et la majorité d'entre elles peuvent s'appliquer aux prénoms, aux noms de rue et d'entreprise.

Tableau I.1 : *Exemples de méthodes de comparaisons de chaînes*

Nom d'entrée	Nom dans la base de données	Méthode Jaro	Winkler	McLaughlin	Lynch	Bigramme
Jones	Johnson	0,790	0,832	0,860	0,874	0,000
Massey	Massie	0,889	0,933	0,953	0,953	0,845
Brrookhaven	Brookhaven	0,933	0,947	0,947	0,964	0,975
Abroms	Abrams	0,889	0,922	0,946	0,952	0,906
Hardin	Martinez	0,000	0,000	0,000	0,000	0,000

Bibliographie

Canadian Centre for Justice Statistics (1998). *Record Linkage in the Canadian Centre for Justice Statistics, 1996-1997.* (Internal Statistics Canada report).

Fair, M. E. (1997). *Record Linkage in an Information Age Society*. Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 427-441.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association,* 64, 1183-1210.

Ferrante, Anna. (1993). Developing an Offender-Based Tracking System : The Western Australia INOIS Project, *Australian and New Zealand Journal of Criminology*, 26, 232-250.

Gill, L.E. (1997) OX-LINK : *The Oxford Medical Record Linkage System*. Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 15-33.

Hart, G.E. (1968). ADP- Police Records and Automatic Data Processing Name Indexes, *Police Research Bulletin*, 8, 14-20.

Houle, C., Berthelot, J-M., David, P., Wolfson, M.C. , Mustard, C. and Roos, L. (1997). *Matching Census Database and Manitoba Health Care Files.* Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 305-318.

Gill, L.E. OX-LINK : (1997). *The Oxford Medical Record Linkage System*. Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 15-33.

Government of Ontario. (1998) Common Positive Identification Technology Assessment and Best Practices. Report prepared by the Ontario Integrated Justice Project.

Labillois, T., Wysocki, M. and Grabowiecki, F. (1997). *A Comparison of Direct Match and Probabilistic Linkage in the Death Clearance of the Canadian Cancer Registry*. Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 203-211.

Moore, G.B., Kuhns, J.L., Trefftzs, J.L. and Montgomery, C.A. (1977). *Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers*. U.S. Department of Commerce (National Bureau of Standards), Washington.

Newcombe, H.B. (1988). *Handbook of Record Linkage : Methods for Health and Statistical Studies, Administration, and Business.* Oxford : Oxford University Press.

Newcombe, H.B., Fair, M.E. and Lalonde, P. (1992). *The Use of Names for Linking Personal Records.* Journal of the American Statistical Association, 87, No. 420. Dec. 1992.

Porter, E. H. and Winkler, W. E. (1997). Approxim*ate String Comparison and its Effect on an Advanced Record Linkage System.* Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 190-199.

Royal Canadian Mounted Police (1999). CPIC Reference Manual (Revision 33). Prepared by the Technical Information Services Section.

US Department of Justice, Bureau of Justice Statistics (2000). *Survey of DNA Crime Laboratories, 1998.* Bureau of Justice Statistics Special Report. Washington.

US Department of Justice, Bureau of Justice Statistics (1997). *Survey of State Criminal History Information Systems, 1977.* Washington.

Winkler, W. E. (1995). *Matching and Record Linkage.* Business Survey Methods. John Wiley & Sons, Inc.

Sites Internet consultés

Nota : Divers sites Internet ont été consultés durant la recherche. Certains des sites comprenaient des renseignements qui divergeaient des renseignements d'autres sites ou de documents publiés. Les renseignements divergeants sur les différents sites et dans la documentation d'origine peuvent être cause d'inquiétude. Il est toujours préférable de se reporter aux documents source s'ils sont indiqués sur le site Internet.

1. www.las-inc.com/tools.htm – ce site porte sur les logiciels MetaMatch, NameClassifier et NameHunter.

2. home.gnofn.org/~nopl/guides/genguide/Soundex.htm – ce site porte sur le système Soundex.

3. www.bradandkathy.com/genealogy/overviewofSoundex.htm – ce site porte sur la méthode phonétique du système Soundex. Il y a de nombreux autres sites Internet liés à la généalogie où il est question de couplage ou de jumelage de noms.

4. www.gcis.net/cjhs/aguideto.htm – ce site décrit le système Soundex Daitch-Mokotoff et montre l'algorithme.

5. www.jewishgen.org/infofiles/Soundex.txt – ce site donne un aperçu du système de codification Russell-Soundex et du système Daitch-Mokotoff Soundex.