N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

REMANIEMENT DE LA SAISIE DES DONNÉES DU RECENSEMENT DE 2006

Jean-René Boudreau et Timothy Withum[1]

RÉSUMÉ

La saisie des données est l'une des étapes essentielles de toutes les enquêtes. Si on ne l'effectue pas avec minutie, des erreurs non dues à l'échantillonnage risquent d'invalider les analyses. Le Recensement de la population du Canada utilise les installations et le personnel de l'Agence des douanes et du revenu du Canada (ADRC) depuis 1981, une collaboration qui a été bénéfique pour les deux organismes. Jusqu'à présent, la saisie des données, c'est-à-dire le transfert de l'information d'un support papier à un support électronique, était réalisée sous forme d'« entrée directe des données ». Cette opération étant très répétitive, elle est sujette à des erreurs aléatoires. Par conséquent, les organismes statistiques recherchent des technologies permettant d'effectuer par machine le transfert de l'information contenue dans la plupart des champs d'enregistrement. Ces dernières années, les gestionnaires du recensement ont procédé à un remaniement important des activités de collecte et de traitement du Recensement de la population du Canada pour répondre à un besoin exprimé d'intégrer plusieurs méthodes de collecte : retour du questionnaire par la poste, Internet, interviews téléphoniques assistées par ordinateur, etc. L'une des activités qui changera beaucoup est la saisie des données. Par suite d'un concours, Lockheed Martin (LM) a obtenu le contrat du traitement des données du Recensement de 2006, y compris un système de saisie des données fondé sur des technologies optiques. Ce système balaiera, reconnaîtra, traitera et sauvegardera la plupart de l'information presque entièrement sans intervention humaine. Il est important que le système LM permette de maintenir la qualité obtenue par l'ADRC lors des recensements antérieurs. Dans la présente communication, nous discutons du défi que pose la transition de la saisie manuelle à la reconnaissance de caractères.

MOTS CLÉS : Entrée directe des données, reconnaissance intelligente de caractères.

1. INTRODUCTION

Afin de pouvoir expliquer adéquatement aux Canadiens comment leur société évolue, ce qui constitue l'un des objectifs principaux du recensement, l'activité de saisie des données, et de façon plus générale, toutes les opérations du recensement, doivent viser dans la plus large mesure possible la réduction du niveau d'erreurs non dues à l'échantillonnage. L'opération de saisie des données joue un rôle crucial dans ce cas, étant donné qu'elle vise à transférer les données d'un support papier à un support électronique. Nous savons que le changement de support entache l'information de son propre lot de bruit. Si l'on ne fait pas suffisamment attention, le contenu de l'information peut être irrémédiablement perdu. On consacre donc des ressources importantes à cette activité.

Lorsque l'on modifie la méthode de saisie des données, le défi consiste à maintenir la qualité des données, même si presque tous les facteurs qui l'influencent changent. Les technologies, les budgets prévus, les programmes d'assurance de la qualité, la conception des questionnaires, etc., sont tous des facteurs qui influent sur la qualité de la saisie des données et, en dépit de tout ça, nous souhaitons que la nouvelle méthode aboutisse à un produit de qualité équivalente ou supérieure à la précédente. Évidemment, le choix d'une meilleure méthode pour la saisie des données devrait entraîner l'élimination ou une réduction significative des problèmes chroniques engendrés par la méthode utilisée précédemment. Lorsqu'une difficulté est présente dans les deux méthodes, la nouvelle méthode devrait comporter une amélioration réelle de la qualité ou présenter des possibilités d'amélioration, au fur et à mesure des versions subséquentes. De plus, étant donné que rien n'est parfait, même si la nouvelle méthode entraîne de nouveaux problèmes, on s'attend à ce que le nombre de problèmes créés soit inférieur au nombre de problèmes résolus.

1 Jean-René Boudreau, Statistique Canada, Ottawa, Canada, K1A 0T6, Timothy Withum, Lockheed Martin, Beltsville, Maryland, États-Unis, 20705

La présente communication porte sur le défi du remplacement de la méthode d'entrée directe des données (EDD) utilisée pour le Recensement de la population canadienne de 2001 par une méthode de reconnaissance optique de caractères proposée par Lockheed Martin pour assurer la saisie des données du Recensement de 2006. Afin d'aborder cette question de façon intelligente, nous avons adopté la démarche qui suit. Nous expliquons brièvement chaque méthode en mettant l'accent sur leurs points forts. Puis, nous nous penchons sur différents problèmes ou difficultés liés à chacune des méthodes et nous décrivons la façon de les traiter. Enfin, nous examinons les répercussions que cela implique sur la qualité des données. Après cet exercice, le lecteur devrait être en mesure de déterminer si le défi a été relevé ou non.

2. ENTRÉE DIRECTE DES DONNÉES

Statistique Canada a donné à contrat les activités de saisie des données à l'Agence des douanes et du revenu du Canada (ADRC) pour chaque recensement, de 1981 à 2001 inclusivement. Cette décision a été bénéfique pour les deux organismes. Dans le cas de l'ADRC, tous les cinq ans, le recensement entraînait une augmentation de la charge de travail au cours de la période de juin à novembre, période pendant laquelle les activités de l'Agence n'atteignent pas leur niveau maximum. Les avantages pour Statistique Canada étaient liés à l'utilisation de l'infrastructure de l'ADRC (locaux, matériel, ressources des TI), et au recours aux employés d'expérience de l'ADRC. Cette collaboration a entraîné des économies considérables pour les contribuables canadiens.

Quelques chiffres pourraient aider le lecteur à bien saisir le contexte de cette activité. Il peut d'ailleurs se reporter à Boudreau et Liu (2002) pour obtenir davantage de renseignements. Le programme pour le Recensement de 2001 a été lancé le 3 juin 2001 et a pris fin le 25 octobre. Huit centres de traitement de l'ADRC au Canada se sont chargés de son exécution. Environ 1 700 opérateurs expérimentés de saisie au clavier y ont participé. Ils ont traité plus de 13,2 millions de questionnaires, comportant environ 600 millions de champs à saisir. Pour traiter une somme aussi grande d'information, ils ont entré plus de 3,8 milliards de caractères, traitant en moyenne 104 caractères à la minute. La question importante à se poser est la suivante : Ont-ils été efficaces? Selon les normes, ils ont été très efficaces. Le taux d'erreur pour les questions a été de 0,74 %. Au niveau du champ, le taux d'erreur n'a pas dépassé 0,23 %. Une brève digression s'impose ici. Comment calcule-t-on un taux d'erreur? Il s'agit de la proportion d'unités (c.-à-d., champs, questions, questionnaires) incorrectes par rapport au nombre total d'unités traitées ou devant être traitées. Le terme « incorrectes » signifie que ce que l'on retrouve sur support électronique après le traitement diffère de ce qui figurait au départ sur le support papier. En outre, le système traite une unité uniquement si elle contient des renseignements, ce qui fait qu'un champ vide n'est pas traité. Par ailleurs, la mention « devant être traitées » se rapporte à toutes les unités comprenant de l'information qui ont été laissées de côté de façon erronée par le système de saisie des données.

Comment les opérateurs de l'ADRC ont-ils saisi l'information? Pour répondre à cette question, nous devons décrire brièvement la façon dont le questionnaire est conçu. Le questionnaire du recensement est tridimensionnel. La première dimension sert à préciser les pages (recto verso) du questionnaire, à savoir essentiellement un ensemble de questions. La deuxième dimension sert à désigner la personne, dans les pages où les questions s'adressent à cette dernière, ou la page, dans les pages où les questions s'adressent au ménage. Chaque bloc est étiqueté au moyen d'un numéro que nous appelons « numéro d'identification de bloc ». La troisième dimension sert à déterminer la question. Il s'agit essentiellement d'un ensemble de champs. Chaque champ est étiqueté au moyen d'un numéro que nous appelons « numéro d'identification de cellule », afin que chaque champ du questionnaire comporte un ensemble unique de coordonnées : numéro d'identification de bloc et numéro d'identification de cellule. L'écran de saisie ne comporte pas d'espace précis pour inscrire des renseignements particuliers; les opérateurs utilisent ces coordonnées pour entrer les données. La saisie au clavier prend par conséquent la forme de longues chaînes de caractères. L'élément de base d'une chaîne comporte trois composantes : numéro d'identification de bloc (au besoin), numéro d'identification de cellule du champ, contenu du champ (au besoin), et séparateur de champ. Chaque page remplie produit une chaîne. Par exemple, si les données d'un questionnaire prennent la forme suivante :

67. 01 q 02 03 ANGLAIS 04 52 !	68. 01 02 q 03 FRANÇAIS 04 25 !

Il faut entrer la chaîne « 67.01-03ANGLAIS-0452-68.02-03FRANÇAIS-0425- ». Le système utilise le point « . » pour isoler les numéros d'identification de bloc (les deux caractères figurant devant chaque point). Le tiret « - » sert de séparateur de champ. Les deux caractères qui suivent un tiret et qui ne sont pas des numéros d'identification de bloc sont des numéros d'identification de cellule. Les caractères qui suivent les numéros d'identification de cellule correspondent au contenu du champ (la réponse). Si un numéro d'identification de cellule est suivi par un séparateur, cela signifie que le répondant a coché le cercle désigné par le numéro d'identification de cellule.

Ce mode de saisie des données d'un questionnaire n'est pas du tout évident. Nous ne recommanderions pas cette méthode aux opérateurs de saisie au clavier qui ne la connaissent pas bien. Toutefois, les opérateurs de l'ADRC sont familiers avec cette méthode. Elle est naturelle pour eux, étant donné que le système de saisie des données qu'ils utilisent est une version modifiée d'un système qu'ils connaissent. Elle est très rapide, étant donné que les opérateurs peuvent entrer de longues chaînes avant d'envoyer les opérations au système de traitement. En fait, les opérateurs remplissent leurs écrans, qui comprennent plusieurs pages, avant de soumettre les opérations. Ils se concentrent donc sur les questionnaires. Ils n'ont pas à regarder leur écran lorsqu'ils saisissent l'information.

Évidemment, le système de saisie comporte certaines vérifications de la précision des données saisies. Il existe trois niveaux de vérification : la validation des numéros d'identification de bloc et de cellule, le processus de confirmation et la vérification de la saisie.

A. Validation des numéros d'identification de bloc et de cellule

Tous les numéros d'identification de bloc et de cellule sont intégrés au système de saisie des données. Si un opérateur entre un numéro d'identification de bloc qui ne correspond pas à la page traitée, le système ne sait pas comment interpréter la chaîne. Il en va de même pour les numéros d'identification de cellule. Lorsque cela se produit, le système demande à l'opérateur d'entrer à nouveau l'ensemble de la chaîne (tous les renseignements de la page). Lorsque tous les numéros d'identification de bloc et de cellule sont valides, le système analyse la chaîne. Les numéros d'identification de bloc et de cellule d'une chaîne doivent être en ordre croissant. Par exemple, si les numéros d'identification de cellule n'augmentent pas à l'intérieur d'un bloc, cela peut venir du fait que l'opérateur a oublié d'inscrire un numéro d'identification de bloc au moment d'entreprendre un nouveau bloc. Lorsque cette analyse échoue, cela provient toujours d'une erreur de saisie des données. Le système demande alors à l'opérateur d'entrer à nouveau l'ensemble de la chaîne. On pourrait penser que les opérateurs sont pénalisés outre mesure lorsqu'ils doivent saisir à nouveau l'ensemble d'une chaîne. Ce n'est toutefois pas le cas. Il est plus facile et rapide de saisir à nouveau l'ensemble des éléments. Autrement, les opérateurs doivent se concentrer davantage sur leurs écrans, et moins sur la saisie proprement dite.

B. Processus de confirmation

Chaque réponse fait l'objet d'une série de vérifications. Une réponse comme « 80 » pour le nombre de semaines travaillées dans une année, ou un code postal invalide, sont des exemples des erreurs que cette vérification vise à déceler. Lorsqu'une réponse échoue à au moins une vérification, il s'agit soit d'une erreur de saisie, soit d'une réponse qui est « considérée » comme douteuse. Dans tous les cas, le système demande une confirmation. Plus tard dans le processus, le système demandera à un autre opérateur d'entrer à nouveau la même réponse. Si une différence est décelée entre les deux entrées, le système demandera à nouveau au deuxième opérateur d'entrer la réponse. C'est donc dire que le deuxième opérateur agit comme vérificateur et comme arbitre au besoin. Dans le cadre du Recensement de 2001, 630 000 (= 0,11 %) champs comprenaient une réponse qui a échoué à au moins une vérification, mais qui a été confirmée par la suite. Par ailleurs, un peu plus de 1,3 million (= 0,22 %) de champs comprenaient une réponse qui a échoué à au moins une vérification, mais qui en fait avait été saisie de façon erronée. Ce pourcentage se situe dans le même ordre de grandeur que le taux d'erreur final au niveau du champ (grâce à cette vérification!)

C. Vérification de la saisie

L'objectif de la vérification de la saisie (VS) est de contrôler le taux d'erreur au niveau du champ, pour tous les champs qui ont réussi à toutes les vérifications précédentes. Cette vérification constitue un contrôle de la qualité (CQ) officiel, qui vise à faire en sorte que le taux d'erreur demeure en-dessous d'un seuil précis. Comme c'est le cas pour tout CQ d'échantillonnage pour acceptation, les questionnaires sont mis en lots avant la saisie originale. Un certain nombre de questionnaires sont sélectionnés de façon aléatoire dans chaque lot. Ils font l'objet d'une nouvelle saisie complète par d'autres opérateurs (opérateurs de VS). Chaque fois qu'un écart est décelé entre les deux saisies, les opérateurs de VS agissent comme arbitres; le système leur demande d'entrer à nouveau l'information jusqu'à ce qu'il soit possible de déterminer qui a fait l'erreur. Étant donné que les opérateurs de VS agissent comme arbitres, les gestionnaires de l'ADRC ont toujours affecté leurs meilleurs opérateurs à cette tâche. Si, pour un lot particulier, on note un trop grand nombre d'erreurs de la part de l'opérateur original, le système demande à l'opérateur de VS d'entrer à nouveau les questionnaire qui restent. Ces questionnaires sont saisis de la même façon que s'ils faisaient déjà partie de l'échantillon. En 2001, environ 9 % des questionnaires ont été sélectionnés dans l'échantillon. Un peu moins de 11 % des lots ont été rejetés.

Un peu plus de 25 % du nombre total de caractères saisis l'ont été dans le cadre de ces vérifications.

3. RECONNAISSANCE INTELLIGENTE DE CARACTÈRES

Dans la présente section, nous décrivons la méthode de reconnaissance intelligente de caractères (RIC). Une différence est toutefois digne de mention. La description figurant dans la présente section constitue une solution proposée pour la saisie des données du Recensement de 2006, et non pas une description de la méthode déjà utilisée. Il se peut que le niveau de précision de la description ne soit pas comparable. Certaines procédures sont même encore à l'étude. Parfois, nous nous reportons à l'expérience de Lockheed Martin relativement aux recensements du Royaume-Uni et des États-Unis.

L'objectif visé par cette méthode est le traitement facile et répétitif par une machine. On procède ainsi parce que la saisie des données est une activité constituée de millions de petites tâches mécaniques. La solution proposée par Lockheed Martin est conçue afin de permettre le traitement de la majorité de ces tâches de façon très fiable, et de soumettre les tâches plus délicates à un processus auquel un plus grand nombre de ressources sont affectées (c.-à-d. nécessitant une intervention humaine). La méthode se répartit en quatre sections : balayage des questionnaires et traitement des images, reconnaissance des marques et des caractères, saisie à partir des images et du document sur papier, et programme d'assurance de la qualité.

A. Balayage et traitement des images

Le premier processus sert à balayer les questionnaires, afin d'obtenir une image de toutes les pages. Dans certains cas, étant donné que les questionnaires prennent la forme d'une brochure, ils doivent d'abord être massicotés. Une fois que les pages ont été balayées, le système de traitement des images aligne ces dernières sur un modèle, afin que tous les champs soient placés au même endroit pour le traitement subséquent. Dans le cadre de ce processus, on effectue une évaluation afin de déterminer si l'image est de qualité suffisante pour permettre une saisie des données très précise et efficace. Dans les cas où l'évaluation fait ressortir un problème de qualité, l'erreur peut être corrigée par un opérateur, ou le document sur papier peut être récupéré et balayé à nouveau ou saisi au clavier.

Lorsque l'image d'une page est satisfaisante, le système effectue une série de traitements pour éliminer les marques et les couleurs inutiles à l'extérieur des zones de réponse. Cela se fait en deux étapes. La première étape est intégrée à la conception des formulaires. Le scanner utilise les couleurs servant à établir le contraste et élimine les autres. La deuxième étape est effectuée à partir de l'image en noir et blanc obtenue. Elle permet de superposer le modèle d'image sur l'image balayée et d'effacer essentiellement le modèle de l'original, afin qu'il ne reste que les marques, les chiffres et les mots, à l'intérieur ou à l'extérieur des zones de réponse, qui ne font pas partie du modèle.

B. Reconnaissance des marques et des caractères

Deux moteurs servent à reconnaître les réponses : la reconnaissance optique des marques (ROM) et la reconnaissance intelligente de caractères (RIC). La ROM est utilisée pour déterminer si une case à cocher l'a été

effectivement, et la RIC sert à déchiffrer les réponses écrites en toutes lettres. Les algorithmes de reconnaissance sont fonction du nombre de pixels établis (= 1) et de leur emplacement dans la zone de réponse. La ROM enregistre uniquement le nombre de pixels établis. On utilise plusieurs niveaux de détection des marques pour obtenir des résultats précis. Le premier niveau correspond à un algorithme simple de seuillage. Le moteur convertit le nombre de pixels établis en une valeur comprise dans une fourchette déterminée. Cette valeur est appelée « degré de certitude ». Si la valeur est supérieure au seuil précisé, le moteur enregistre une marque. Si la valeur est inférieure à un autre seuil précisé, le système n'enregistre pas de marque (le champ est considéré comme vide ou en blanc). Si la valeur se situe entre ces deux seuils, une intervention humaine est nécessaire. Dans le cas de la RIC, on utilise le même type de processus, mais ce dernier est beaucoup plus complexe. Tout d'abord, la RIC doit segmenter la réponse en caractères. La conception du questionnaire est très utile à cet égard. Chaque zone de réponse remplie est fractionnée en cases de dimensions égales, afin de décourager l'utilisation de lettres cursives. On demande aux répondants d'utiliser des majuscules. Une fois que les caractères sont séparés, chacun comporte un degré de certitude. On effectue une série de vérifications au niveau du champ, afin de rehausser ou de diminuer le degré de certitude des caractères. Par exemple, en anglais et en français, certaines combinaisons de lettres augmentent le degré de certitude (p. ex., un « q » suivi par un « u » dans la chaîne). Il est rare que l'on voie trois « i » accolés (« iii »), ce qui fait que cette série diminuera le degré de certitude. En outre, si le nombre de réponses est limité, par exemple, une province ou un pays de naissance, le système peut consulter les réponses d'un tableau de référence, aussi appelé dictionnaire. Ce tableau comprend des séries de caractères qui s'apparentent aux réponses possibles. Lorsqu'un appariement est trouvé, cela augmente considérablement le degré de certitude du caractère. Il convient de souligner que la RIC produira souvent des deuxièmes choix pour chaque caractère. Par exemple, la lettre « C » peut être prise pour un « C », le second choix comportant un degré de certitude plus faible étant « O ». Le traitement au moyen du dictionnaire tiendra compte de ces deuxièmes choix au moment de l'évaluation. Une fois ces analyses effectuées, si le degré de certitude du champ (le degré de certitude le plus faible parmi tous les caractères du champ) est supérieur à un seuil déterminé, la RIC enregistrera la chaîne de caractères comme étant la réponse. Si le nombre de pixels établis dans toutes les cases de réponse est insuffisant, la RIC n'enregistrera pas de réponse (le champ sera considéré comme vide ou en blanc). Dans d'autres cas, la RIC nécessite une intervention. Comme pour la ROM, il s'agit d'une intervention humaine. Voilà à peu près comment les moteurs de reconnaissance fonctionnent. Selon notre expérience passée, 0,1 % des marques et 20 % des réponses écrites nécessitent l'intervention humaine; toutefois, ce dernier pourcentage dépend dans une large mesure des types de champs.

C. Saisie à partir de l'image et saisie à partir du document sur papier

Lorsqu'une intervention humaine est nécessaire pour la ROM ou la RIC, le champ doit être saisi par un opérateur. Cette opération est appelée saisie à partir de l'image (SI), ou « correction des données ». Elle peut être perçue comme une « file d'attente de service », chaque champ faisant la queue pour obtenir un service (saisie). L'écran qui sert à cette saisie est très particulier. Il est destiné à aider les opérateurs à se concentrer. Il se divise en trois sections horizontales identiques. La section du milieu comprend des renseignements concernant le champ à traiter. La section supérieure comprend des renseignements concernant le dernier champ traité et la section inférieure, des renseignements concernant le champ qui sera traité par la suite. Chaque fois qu'un opérateur traite un champ, les sections sont mises à jour. Cet écran permet aux opérateurs de réévaluer le champ qui vient d'être traité ou d'avoir un aperçu rapide de ce qui suit. Cette méthode s'est révélée très efficace. Dans chaque section, on retrouve une fraction d'image autour du champ et une zone de saisie. L'opérateur doit entrer ce qu'il se voit dans le champ compris dans la fraction d'image de la zone de saisie. Encore une fois, les champs qui ont été soumis à ce processus n'étaient pas suffisamment clairs pour être reconnus, ce qui fait qu'on peut s'attendre à des réponses étranges. Cela signifie que des règles de saisie bien définies sont nécessaires (p. ex. Que doit-on entrer lorsque deux chiffres sont présents dans un champ?).

Certains questionnaires ne peuvent être balayés parce qu'ils ont été endommagés d'une façon ou d'une autre. D'autres questionnaires ne produisent pas une image claire, parce que, par exemple, les réponses figurant au recto ont transpercé le papier et interfèrent avec les réponses figurant au verso. Lorsque c'est le cas, toutes les pages du questionnaire doivent faire l'objet d'une entrée directe. Cette opération est appelée saisie à partir du document sur papier (SP). Pour entrer les données, le système fournira une version modifiée des écrans Internet utilisés par les répondants. On demandera aux opérateurs de répondre aux questions figurant dans les écrans Internet à partir des données figurant dans les questionnaires.

D. Programme d'assurance de la qualité

Le programme d'assurance de la qualité vise à assurer le même niveau de précision que le système d'EDD du Recensement de 2001. Le seuil de confiance décrit ci-dessus, c'est-à-dire le point qui sépare la reconnaissance automatique de la saisie manuelle, revêt une importance cruciale. Si la valeur de ce seuil est trop élevée, on aura recours trop souvent à la SI, même si la reconnaissance est très précise. Le temps et les ressources poseront un problème. Par ailleurs, si la valeur du seuil est trop faible, la qualité pourrait en souffrir. Les seuils de confiance doivent être établis de façon appropriée pour maintenir un équilibre entre la qualité et les ressources. La technologie disponible permet d'augmenter la valeur de ces seuils suffisamment pour établir un équilibre entre les deux objectifs. Étant donné que la saisie de l'information est principalement effectuée par une machine, cette dernière doit être bien réglée. Par ailleurs, la qualité des opérations de SI et de SP doit faire l'objet d'un contrôle, afin de garantir la qualité des données au niveau prescrit.

Comment contrôle-t-on les moteurs de ROM et de RIC dans le cadre de l'assurance de la qualité? Le système doit effectuer un deuxième passage pour les champs qui ont fait l'objet d'une reconnaissance automatique. À cette fin, on sélectionne un échantillon de champs. Dans le système, les degrés de certitude de tous les champs de l'échantillon qui ont été reconnus de façon automatique sont réduits. Puis, ces champs sont soumis à un processus de SI et font tous l'objet d'une nouvelle saisie. La saisie initiale et la deuxième saisie sont soumises à un algorithme d'arbitrage, afin de déterminer si la saisie initiale est correcte ou erronée. On calcule des taux de précision, et des mesures sont prises si ces taux sont trop faibles.

Le deuxième élément à contrôler est la qualité des opérateurs de SI. À cette fin, le système sélectionne de façon aléatoire des champs à partir des lots. Dans ce cas, le lot est l'unité de traitement, c'est-à-dire un ensemble de questionnaires. Si un champ sélectionné comporte un degré de certitude faible, il est renvoyé à la SI, afin d'être saisi à nouveau par un autre opérateur de SI. Puis le système récupère ce que le moteur a reconnu en premier (même si les degrés de certitude sont faibles), la saisie initiale et la deuxième saisie de SI, et soumet ces trois éléments à un algorithme d'arbitrage, afin de déterminer si la première saisie de SI était correcte ou erronée. Si le taux de précision observé dans l'échantillon d'un lot est trop faible, tous les champs comportant de faibles degrés de certitude sont entrés à nouveau par SI. La subtilité de ce processus vient du fait que les opérateurs de SI ne savent pas s'ils saisissent, vérifient ou arbitrent les données. L'assurance de la qualité de l'opération de SI est similaire, sauf que l'unité d'échantillonnage correspond à un questionnaire complet plutôt qu'à un champ. (Même si l'on préfère l'échantillonnage au niveau du champ pour obtenir un échantillon plus efficace, cela n'est pas rentable pour la SP.) Une fois qu'une série de lots de questionnaires ont été entrés par un opérateur de SP, on sélectionne un échantillon de questionnaires qui sont saisis à nouveau par un deuxième opérateur. Au moment de cette nouvelle saisie, le système soumet la saisie initiale et la deuxième saisie des questionnaires sélectionnés à un algorithme d'arbitrage. Si le taux de précision de la première SP est trop faible, les lots font l'objet d'une nouvelle saisie complète par un arbitre de la SP. Cet opérateur spécial de SP saisit à nouveau tous les champs de tous les questionnaires des lots et agit comme arbitre au besoin. Si la saisie d'un champ donné diffère de la SP originale, le système demande à l'opérateur de saisir à nouveau le champ en question. Le mécanisme d'évaluation des lots des opérations de SI et de SP est un peu plus complexe que l'explication fournie, mais les simplifications faites n'affectent pas l'idée générale.

4. ANALYSE DES CHANGEMENTS TOUCHANT LA MÉTHODE DE SAISIE DES DONNÉES

Comme nous l'avons indiqué précédemment, les répercussions sur la qualité des données seront déterminées par suite d'un examen de la façon dont chaque méthode permet de résoudre les problèmes ou les difficultés qui se posent. Nous pouvons répartir cette section en trois parties. La première partie porte sur les difficultés liées à la méthode d'EDD, qui devraient être éliminées grâce à la RIC. La deuxième aborde les problèmes qui se posent pour les deux méthodes. Dans cette partie, nous mettons l'accent sur la simplicité des solutions possibles. La dernière partie aborde les difficultés liées à la RIC, qui ne posent pas réellement de problèmes pour l'EDD.

4.1 Difficultés liées à l'EDD éliminées par la RIC

A. Erreurs de coordonnées (numéro d'identification de bloc et numéro d'identification de cellule)
La première erreur que la RIC permet d'éliminer est la saisie d'un numéro d'identification de cellule erroné qui n'est pas décelée par le système. Si nous utilisons le même exemple que précédemment, l'opérateur devrait avoir entré ce qui suit : « 67.01-03ANGLAIS-0452-68.02-03FRANÇAIS-0425- », mais a plutôt entré « 67.02-03ANGLAIS-0452-68.02-03FRANÇAIS-0425- ». Le numéro d'identification de cellule « 02 » dans le numéro d'identification de bloc « 67 » est un numéro d'identification de cellule valide. Cette situation entraîne deux erreurs de saisie des données. La première erreur correspond à un champ omis (le numéro d'identification de cellule correct n'est pas entré), la deuxième, à l'ajout d'un champ (le mauvais numéro d'identification de cellule est entré). À partir de l'échantillon de CQ, nous avons observé que 17 % de toutes les erreurs qui n'ont pu être décelées par le système découlaient en fait de l'ajout de champs. Nous croyons qu'un nombre significatif de champs excédentaires ont été produits de cette façon en 2001. De façon plus générale, le traitement manuel peut entraîner la création de données qui ne figurent pas réellement dans le questionnaire. Il se peut que l'opérateur ait consulté le mauvais champ au moment de la saisie, etc. Dans le cas de la RIC, il est à peu près impossible de créer des données à partir de rien ou de supprimer complètement des données présentes. Au pire, le champ est envoyé à la SI pour être corrigé.

B. Caractères spéciaux qui ne peuvent être saisis
La RIC peut saisir les points « . » et les tirets « - ». L'EDD ne permet pas la saisie de ces caractères, parce qu'ils sont utilisés par le système pour désigner les numéros d'identification de bloc et pour séparer les champs. Cela est important parce que cela réduit la somme d'instructions à l'intention des opérateurs de SI et de SP. La consigne « Entrez ce que vous voyez » inclut tous ces cas dans la RIC. Toutefois, grâce à la méthode d'EDD, les opérateurs ont comme instructions d'entrer une espace « » chaque fois qu'un point doit être entré. Cela prête particulièrement à confusion dans les champs comportant de nombreuses abréviations (p. ex., en anglais, pour U.S.A. devrait-on entrer « U S A » ou USA? Ou devrait-on saisir deux espaces si un point est suivi par une espace?). Ces types de situation ne produisent pas beaucoup d'erreurs, mais les opérateurs ont de la difficulté à se conformer à la consigne « Entrez ce que vous voyez ». La principale contrainte liée à l'incapacité de saisir un tiret est la suivante : « Que fait-on avec les nombres négatifs? ». Étant donné que la RIC permet la saisie de ces caractères, elle réduit le nombre de cas spéciaux à traiter, ce qui est particulièrement important pour les systèmes de saisie des données.

C. Réponses tronquées
Les réponses tronquées présentent une autre difficulté qui pourra être résolue par le méthode de RIC. Les exemples qui suivent sont propres au système d'EDD élaboré par l'ADRC. Du fait de la méthode utilisée pour exporter les données à Statistique Canada (un enregistrement par document), le nombre d'octets possibles était limité. Chaque champ alphanumérique comportait un nombre d'octets déterminé par suite de négociations. On a saisi plus de 50 millions de champs alphanumériques. De ce nombre, 0,19 % étaient tronqués. Toutefois, les champs tronqués n'étaient pas répartis également entre les questions et les langues. Par exemple, 0,99 % des champs « Main Activity » en anglais et 1,68 % des champs « Activité principale » en français étaient tronqués. La question des champs tronqués est importante, étant donné que certains d'entre d'eux peuvent comprendre des termes importants à la fin. Il s'agit d'un problème différent de celui du nombre insuffisant de cases servant à entrer l'information, qui est propre à la méthode de RIC. Si cela se produit, un répondant peut inscrire deux lettres ou plus dans les cases et continuer à écrire la réponse à côté. Ces cas comportent de faibles degrés de certitude et doivent être confiés aux opérateurs de SI. Ceux-ci doivent être en mesure d'entrer le contenu de toutes les réponses.

D. Activités de renvoi
La RIC ne comporte pas d'activités de renvoi (activités qui permettent de modifier les données avant la saisie proprement dite). Dans le contexte de l'EDD, les commis au renvoi doivent « corriger » les données avant de les saisir, parce que leur format n'est pas acceptable. Si le « loyer mensuel » demandé est fourni pour l'ensemble de l'année, l'opérateur doit demander au commis au renvoi de procéder à la conversion appropriée. La réponse originale est barrée et le montant converti est inscrit à côté de la zone de réponse. Les commis au renvoi inscrivent les corrections dans les questionnaires, afin d'éviter d'avoir à reprendre cet exercice lorsque les champs sont saisis (au besoin) dans le cadre d'un des processus de vérification. Cela fait en sorte que le processus de vérification et d'arbitrage est très dépendant, et probablement trop dépendant, de la saisie originale. Avec la RIC, étant donné que les opérateurs travaillent à partir d'images, des processus indépendants se déroulent lorsque les mêmes champs sont

saisis une deuxième fois. Par ailleurs, la tâche des commis au renvoi ne se limite pas à la conversion. Ils doivent appliquer toutes les règles de pré-traitement figurant dans les guides des opérateurs, par exemple : Que doit-on saisir lorsque deux sources de revenu sont déclarées pour une composante du revenu? Ils ont même le dernier mot lorsque des termes ne sont pas admissibles. Dans l'ensemble, en 2001, 0,13 % (= 425 000) de toutes les questions ont nécessité l'intervention de commis au renvoi. Il s'agit d'un nombre appréciable de corrections. Même si la solution de la méthode de RIC pour ces problèmes de saisie n'est pas encore déterminée avec précision, toutes les solutions qui se rapprochent de la consigne « Entrez ce que vous voyez » devraient avoir la préférence, par exemple, « Entrez les données et laissez le système faire les calculs ou, de façon plus générale, le traitement ». Dans l'exemple des deux sources de revenu, la saisie des deux chiffres et l'inscription d'un signe « + » entre les deux permettent de résoudre le problème. L'intervention du commis au renvoi est remplacée par l'insertion des deux entrées et par l'introduction d'un code source permettant d'additionner les deux chiffres au moment du traitement postérieur à la saisie.

4.2 Difficultés communes

A. Réponses illisibles

Les méthodes de saisie des réponses illisibles sont parfois complexes, parce qu'elles comprennent des éléments contradictoires. Par exemple, dans le cadre du Recensement de 2001, la procédure pour les réponses illisibles était la suivante : « *Le commis au renvoi doit tenter de déterminer la réponse et supprimer les parties illisibles. Si la réponse est complètement illisible, il faut supprimer l'ensemble du champ* ». Toutefois, plus loin dans le guide, il est aussi dit : « *N'essayez pas d'interpréter la réponse si le répondant n'a répondu qu'en partie* ». Les mots « déterminer » et « interpréter » ont une signification similaire dans ce contexte. Dans le cadre du Recensement de 2006, on s'attend à ce que tous ces cas soient envoyés à la SI pour être saisis à nouveau. Si ces champs sont saisis à nouveau par d'autres opérateurs de SI, des évaluations indépendantes seront effectuées. Deux opinions valent parfois mieux qu'une seule.

B. Conversions

Des conversions sont effectuées lorsque les montants déclarés n'utilisent pas l'unité de mesure requise; le loyer annuel plutôt que le loyer mensuel, la rémunération horaire plutôt que la rémunération annuelle, etc. constituent des exemples. Selon la méthode d'EDD, le commis au renvoi doit procéder à ces conversions. En 2001, on a enregistré 180 000 conversions numériques (= 0,17 % de tous les champs numériques). En 2006, ces champs ne pourront être identifiés que si des unités de mesures sont indiquées dans les zones de réponse ou à proximité de celles-ci (p. ex. à l'heure, par semaine, etc.). Sans ces unités de mesure, on ne pourra pas procéder aux conversions. Par ailleurs, des conversions peuvent être effectuées si les réponses s'écartent de certains contrôles par intervalles. Le taux d'erreur observé pour la conversion a été de 3,97 % en 2001. Si la RIC permet de déterminer qu'une conversion est nécessaire, il se peut que l'opérateur de SI ne puisse l'effectuer parce que les écrans de SI n'affichent que des fractions de questions. L'opérateur de SI peut entrer uniquement le montant déclaré et indiquer la conversion requise. Encore une fois, la solution de la méthode de RIC pour la conversion est encore à l'étude, mais comme nous l'avons mentionné dans la section précédente, toutes les solutions qui prévoient la saisie des données par les opérateurs et les conversions par le système sont supérieures à celles qui consistent à demander aux opérateurs d'épurer les données.

C. Erreurs de substitution

Dans le contexte de la RIC, on parle le plus souvent d'« erreurs de substitution », c'est-à-dire que le moteur de la RIC reconnaît de façon incorrecte un caractère. Cela se produit aussi dans le contexte de l'EDD, mais dans une moindre mesure. En 2001, le taux d'erreur estimé découlant d'une mauvaise saisie dans le cas des champs de réponses écrites était légèrement inférieur à 0,07 %, soit un taux de précision de 99,03 %. Ces erreurs dépendent directement de la fiabilité des opérateurs chargés de la saisie pour la méthode d'EDD. Dans le cadre de la méthode de RIC, la fiabilité du moteur de RIC s'ajoute à la fiabilité des opérateurs de SI et de SP. L'expérience du recensement des États-Unis a démontré que, pour le moteur de RIC et les opérateurs de SI, la précision a atteint 99,5 % et 98 % respectivement. Cela donne un niveau de précision globale de 99,4 %. Nous devons cependant être prudents avec ces chiffres. Ils ont été calculés à partir de définitions des erreurs pour des champs alphanumériques de réponses écrites différentes de celles utilisées par les opérateurs de l'ADRC. Il n'est pas possible de les rajuster pour procéder à une comparaison juste.

D. Réponses annulées par les répondants

Devrait-on saisir les champs qui ont été supprimés par les répondants? La réponse est non. En 2001, 2,5 millions de questions comprenaient au moins un champ supprimé (= 0,75 % du nombre total de questions). Même si la méthode d'annulation des champs était claire pour l'EDD, elle a suscité beaucoup de confusion, parce qu'il n'est pas toujours facile de déterminer si un champ particulier est annulé. La RIC comporte le même problème. Si l'annulation se produit dans un champ de réponse écrite, on s'attend à ce que le degré de certitude soit faible et à ce que l'opérateur de SI qui traite le champ puisse décider si ce dernier est annulé. Dans le cas des champs à cocher, cela n'est pas aussi facile. Le système extrait des éléments plus complexes à partir de chaque marque, afin de déterminer si elles sont suffisamment réalistes pour être considérées comme la réponse fournie par le répondant. Chacun de ces facteurs est évalué, afin de classer chaque marque comme valable ou non. Afin de pousser plus loin l'évaluation, on évalue collectivement les marques multiples figurant dans une page, afin de déterminer une tendance quant à la façon dont le répondant inscrit les marques. La précision de la ROM dans le cadre de l'expérience du recensement des États-Unis a été de 99,98 %, ce qui est similaire à l'expérience d'EDD de 2001.

E. Symboles visuels

Parfois, les répondants inscrivent des symboles visuels plutôt que des réponses. Étant donné que les questionnaires du recensement de la population sont conçus sous forme de tableau, un répondant peut tracer une flèche ou utiliser des guillemets de répétition, pour indiquer que la réponse s'applique à d'autres personnes dans la même page. Il peut aussi répondre à une question au moyen d'un point d'interrogation, lorsqu'il ne sait pas la réponse. En 2001, 0,20 % de toutes les questions comportaient ce type de symbole dans les zones de réponse. Dans le cas des flèches ou des guillemets de répétition, on a demandé à l'opérateur de saisie d'entrer ce que le signe signifiait. Cette façon de faire n'a pas produit une saisie précise pour ces champs. Le taux d'erreur pour ces questions a été de 11,38 %. Les méthodes de saisie n'étaient pas uniformes; certaines différaient pour des questions différentes. La fréquence de ces symboles dépend de la conception du questionnaire. Le formulaire sous forme de tableau facilite cette façon de faire, parce que chaque zone de réponse pour toutes les personnes figure sur la même ligne ou rangée. Les zones de réponse segmentées, pour aider à la segmentation des caractères pour la RIC, pourrait résoudre ce problème en partie. Étant donné que le recensement est une enquête par autodénombrement, les questions trop détaillées (comme les composantes du revenu au dollar près) augmentent la fréquence des points d'interrogation dans les champs numériques. En 2006, si un de ces symboles est présent dans un champ, nous nous attendons à ce que son degré de certitude soit faible et à ce que l'opérateur de SI puisse prendre les mesures appropriées. Nous croyons qu'il est possible de réduire le taux d'erreur.

F. Facteur humain

L'utilisation de la méthode de RIC réduit en outre la nécessité de recruter et de former un effectif temporaire important. La méthode d'EDD nécessite du personnel très bien formé, et l'ADRC pourrait bien, en 2006, ne pas pouvoir fournir un effectif de taille et de qualité égales à celui des recensements précédents. Même si le recrutement et la formation d'un effectif chargé de la saisie des données continuent de poser un problème, la méthode de RIC, étant donné qu'elle repose sur un effectif réduit, l'atténue de façon significative. La particularité des moteurs de RIC/ROM est qu'ils ne sont pas affectés par le stress, les longues heures, etc. En outre, une fois la formation assurée, le moteur peut être reproduit à plusieurs endroits dans le système (sans courbe d'apprentissage), afin d'augmenter le volume de traitement. (Le recrutement d'autres personnes entraînera souvent de la formation additionnelle.) Étant donné qu'ils sont automatisés, les moteurs peuvent utiliser l'ensemble des données du questionnaire pour améliorer la précision. Par exemple, les algorithmes peuvent utiliser tous les noms de famille figurant dans un questionnaire — qui sont parfois les mêmes — pour déterminer le nom de famille approprié à saisir. Les opérateurs peuvent aussi le faire, mais cela est plus long.

4.3 Nouveaux problèmes pour la RIC

A. Données à l'extérieur des zones de réponse

Pour 1,2 million de questions (= 0,20 %), une partie de la réponse, ou l'ensemble de la réponse, se trouvait à l'extérieur de la zone de réponse. Même si le taux d'erreur pour ces questions était élevé, soit 4,9 %, seulement 14 % de ce taux était attribuable à des questions omises. L'EDD est une bonne méthode pour les déceler. On a même mis en place des procédures à suivre si le répondant avait répondu à une question en indiquant des exemples dans la zone de question. Comment la RIC peut-elle déceler les réponses à l'extérieur des zones de réponse? Lorsque l'on met une

page en image, plusieurs processus éliminent les données comprises dans le modèle, et le formulaire disparaît, comme nous l'avons indiqué précédemment. Après ces processus, si la page est bien positionnée dans le scanner, l'image obtenue ne montrera que les données inscrites par les répondants. Une analyse du nombre de pixels à l'extérieur des zones de réponse est effectuée. Le système peut décider si une page particulière comporte un trop grand nombre de ces pixels. Puis, la page peut être balayée à nouveau ou l'ensemble des pages du questionnaire peut être soumis à la SP pour une meilleure saisie.

B. Intégrité des questionnaires

La difficulté dans ce cas consiste à garder ensemble les pages séparées d'un même questionnaire. Si chacun des questionnaires comportait un code à barres unique, il n'y aurait pas de problème. Les pages pourraient être traitées de façon appropriée, même si elles ont été mélangées. Toutefois, cette façon de procéder est souvent trop coûteuse. D'autres méthodes sont disponibles pour assurer l'intégrité des documents, et elles sont presque aussi précises tout en étant beaucoup moins coûteuses. L'explication de ces méthodes, qui sont encore à l'étude, dépasse toutefois la portée du présent document.

C. Reconnaissance des accents en français

La reconnaissance des accents en français n'a jamais posé de problème dans la méthode d'EDD. Les opérateurs de saisie devaient entrer uniquement les lettres, sans accent. Pour la RIC, cela ne sera peut-être pas aussi facile. Il se peut que les accents en français perturbent l'algorithme de segmentation des caractères, étant donné que l'« univers » des caractères à reconnaître est plus vaste.

5. RÉPERCUSSIONS SUR LA QUALITÉ DES DONNÉES

La qualité des données doit être évaluée après chaque activité majeure dans le cadre d'une enquête. La saisie des données pour le Recensement de 2006 ne fera pas exception à cette règle. Une étude sera effectuée après la production, afin d'évaluer les taux d'erreur et de déterminer les raisons de ces erreurs. Toutefois, nous devons maintenant évaluer les répercussions des changements de méthodologie sur la qualité des données, ce qui est différent. La qualité des données est habituellement mesurée du point de vue de la variance et du biais. Dans ce contexte, la variance est fonction de la fiabilité des méthodes pour la reproduction des mêmes résultats dans des conditions semblables. Un exemple simple est la mesure de la dispersion de la saisie, dans les cas où les opérateurs d'EDD doivent saisir le chiffre « 5 ». Pour ce qui est du biais, une autre mesure est requise. Dans ce cas, il s'agit de la précision des méthodes dans une situation donnée. Dans le contexte de l'EDD, un exemple de biais serait le suivant : tous les commis au renvoi d'un emplacement de traitement font systématiquement la même erreur d'interprétation des règles de saisie. Dans le contexte de la RIC, un exemple de biais serait le suivant : les réponses figurant à l'extérieur des zones de réponse sont très différentes de celles figurant à l'intérieur et sont toutes laissées de côté par la RIC. Nous abordons maintenant la pertinence de réduire ou d'améliorer la qualité des données en modifiant la méthode. Nous séparons l'analyse en deux parties. La première partie traite de la saisie simple des données, qui ne nécessite aucune correction. La deuxième partie traite des cas où des corrections sont requises.

A. Saisie simple

Lorsque les opérateurs de saisie de l'EDD font des erreurs, à quelle composante ces erreurs appartiennent-elles? Produisent-elles principalement une variation ou un biais? Selon la position de la main sur le clavier numérique et la position des touches numériques, pouvons-nous observer certaines tendances dans la saisie des chiffres par les opérateurs? Il est difficile de l'affirmer. Les erreurs semblent être aléatoires. En 2001, 66 % de toutes les erreurs (= 1,6 million) commises par des opérateurs de l'EDD avaient trait à l'omission ou à l'ajout de champs. Ces erreurs sont le fait du manque d'expérience des opérateurs qui entrent les numéros d'identification de bloc et de cellule, ou de leur inattention. Ces champs ne représentent pas des cas complexes. Ces erreurs appartiennent principalement à la composante de la variance. Toutefois, elles peuvent entraîner du biais pour les questions comportant très peu d'occurrences, comme l'« ascendance autochtone ». À partir de l'échantillon de CQ de la vérification de la saisie en 2001, nous avons observé que la catégorie « Inuit » avait été saisie 907 fois dans l'échantillon par les opérateurs de l'EDD au moment de la saisie originale. Le CQ a révélé que 99 de ces cas appartenaient dans les faits à la catégorie « Non autochtone », ce qui signifie que 10 % des Inuits n'étaient pas réellement des Inuits. La majorité des ces erreurs ont été corrigées par le CQ ou ont été décelées plus tard dans le processus de recensement, au moment de la

comparaison avec d'autres variables, comme l'immigration, la langue, etc. Dans le cas de ces questions, nous pouvons déceler un certain biais. Heureusement, toutes ces situations disparaîtront avec la méthode de RIC.

Ce qui subsistera essentiellement, ce sont des erreurs de substitution quant au contenu des réponses. En 2001, 30 % des erreurs (= 725 000 erreurs) effectuées par des opérateurs de l'EDD découlaient de la saisie erronée des données. Nous croyons qu'elles appartenaient toutes à la composante de la variation. Le moteur de RIC affiche les mêmes tendances. Par exemple, les chiffres « 8 » et « 6 » sont assez similaires. La distribution des chiffres erronés n'est toutefois pas uniforme. Néanmoins, si l'on décèle un biais au cours de la production, une modification des algorithmes de traitement est possible et permettra de corriger presque instantanément le problème. Par ailleurs, si un biais est décelé après la production, on saura que la RIC a traité les cas de la même façon chaque fois. Il n'y aura pas de variation d'un opérateur à l'autre. Une telle situation devrait être facile à corriger.

Dans le cas d'une saisie simple, qui s'applique à 98,17 % de toutes les questions traitées, nous concluons ce qui suit.

1. La méthode d'EDD mène principalement à des erreurs de variation. Les deux tiers d'entre elles sont des champs omis ou des champs ajoutés. Nous ne croyons pas que les champs en question sont plus difficiles à saisir. La méthode de RIC permettra de les éliminer pour la plupart, sinon en totalité. Il s'agit là de la principale force de la méthode de RIC;

2. La méthode de RIC produit principalement des erreurs de substitution. Ces erreurs se concentrent dans les champs qui sont plus difficiles à déchiffrer. Étant donné que la distribution des chiffres erronés n'est pas uniforme, un biais est présent, mais peut être « tolérable ». Toutefois, on ne peut procéder à une évaluation équitable à partir des données disponibles, du fait de la présence d'un trop grand nombre d'aspects non contrôlés (y compris des calculs dont la précision diffère, les règles de saisie à appliquer, le poste de travail de l'opérateur), et il est trop tôt pour conclure qu'une méthode est meilleure que l'autre. Cette question fera l'objet d'un examen plus poussé au cours de la répétition générale, lorsqu'un plus grand nombre de données seront disponibles;

3. Il existe une possibilité d'amélioration de la qualité des données si les méthodes de saisie des opérateurs de SI suivent le plus possible la consigne « Entrez ce que vous voyez ».

B. Corrections nécessaires avant la saisie

En 2001, 425 000 questions ont dû être corrigées avant la saisie. Parmi elles, environ 85 000 (= 20,12 %) n'ont pas été saisies correctement parce que les corrections n'ont pas été effectuées de façon appropriée. Si nous partons du principe qu'un seul champ par question avait besoin d'être corrigé, les corrections erronées représenteraient 6 % de tous les champs erronés, ce qui n'est pas négligeable. Nous n'avons pas de données pour déterminer les types de réponses qui suscitent des corrections, mais nous ne pouvons pas présumer qu'elles sont similaires à celles qui n'en suscitent pas. Par exemple, si une personne déclare son salaire sur une base horaire, on peut s'attendre à une rémunération annuelle faible, même si cela n'est pas toujours vrai. Si des corrections ne sont pas effectuées, un biais sera présent, probablement minime au niveau du Canada, mais peut-être important pour des petites régions géographiques. On n'a pas encore trouvé de solution pour traiter ces cas au moyen de la méthode de RIC. Si certaines fonctions de SI étaient élaborées en conformité du présent document : touches de fonction spéciales pour traiter les conversions, résolutions d'expressions algébriques, etc., la qualité des données de ces questions serait améliorée considérablement.

6. CONCLUSION

Les gestionnaires du recensement ont remanié certaines activités de collecte et de traitement, afin de répondre à la nécessité de centraliser et d'intégrer les opérations. L'opération de saisie des données se situe au cœur de cette restructuration. Du fait de la nouvelle méthode utilisée pour la tenue du recensement au Canada, il n'est plus possible d'avoir recours au système d'entrée directe des données conçu par l'Agence des douanes et du revenu du Canada, pour saisir les données. Une nouvelle méthode fondée sur la reconnaissance intelligente de caractères (RIC)

est obligatoire. Le système de saisie des données proposé par Lockheed Martin est celui que les gestionnaires du recensement ont retenu, par suite d'un concours.

Nous avons décrit les deux méthodes, de même que leurs forces et leurs faiblesses. Certains problèmes inhérents à l'EDD disparaîtront, comme l'omission de champs, l'ajout de champs et les erreurs de coordination de la saisie des numéros d'identification. Avec la RIC, quelques nouveaux problèmes se poseront, comme l'intégrité des questionnaires et les données inscrites à l'extérieur des zones de réponse. Nous avons montré le déséquilibre qui existe entre les anciens problèmes éliminés et les nouveaux problèmes qui se posent. Étant donné que les deux méthodes comportent le même niveau de précision, les différences ont davantage trait aux types d'erreurs et aux raisons qui motivent ces erreurs. On notera probablement un moins grand nombre d'erreurs dans les champs « faciles », et peut-être plus d'erreurs dans les champs « difficiles ». Nous pourrons évaluer ces changements dans le cadre de la répétition générale de 2004, puis prendre des mesures, au besoin, pour en réduire les répercussions pour le Recensement de 2006, grâce notamment à la mise au point des moteurs, à des données réelles ou à l'ajout de fonctions à l'opération de SI, afin de traiter les situations plus complexes.

Pour le moment, nous sommes confiants que la qualité des données découlant de la saisie ne sera pas compromise. Étant donné l'expérience de Statistique Canada avec l'EDD, et celle de Lockheed Martin, avec la RIC et les recensements des États-Unis et du Royaume-Uni, le système de saisie des données qui servira à traiter le Recensement de 2006 réduira au minimum les risques de problèmes de qualité. Dans une optique plus positive, nous nous attendons à ce que la qualité des données soit améliorée.

RÉFÉRENCES

Boudreau J.R., et L. Liu (2002), "The 2001 Direct Data Entry Evaluation Study", unpublished report, Ottawa, Canada: Statistics Canada.