Utilisation des technologies émergeantes
Obtention des meilleurs résultats de la saisie de données mises sur papier

Andy Tye[1] & Mike Smethurst[2]
DRS Data Services Ltd.[3]

Introduction:

Ce document examine les techniques de "collecte/traitement de données" de recensement, qui ont fait leurs preuves au moyen de brochures/dépliants et considère aussi la toute dernière technologie émergeante qui est maintenant disponible pour l'organisation du recensement pour saisir les données de formulaires de recensement sur papier. Les auteurs sont tous deux des directeurs haut placés chez DRS Data Services Limited au RU. DRS est un des principaux fournisseurs internationaux de produits et services de scanning. La société a une expertise toute particulière en saisie de données de recensement et inscription d'électeurs.

Bien qu'on reconnaisse que les méthodes de collecte de données non basées sur papier sont actuellement utilisées dans le recensement, telles que l'utilisation de l'Aide Numérique Personnelle (PDA), de l'Internet, du Téléphone, etc., ce document vise spécifiquement la collecte et le traitement des données de recensement à partir de formulaires en papier.

Il y a quatre méthodes de collecte[4] traditionnelle de données à partir de brochures/papier (résumées ci-dessous). La nouvelle technologie émergeante est en mesure de combiner les caractéristiques clés de ces techniques traditionnelles. Ce document analyse certains projets sur lesquels ces combinaisons de techniques sont actuellement utilisées.

Ce document fait les assomptions suivantes: les formulaires sur papier sont conçus spécifiquement. Une procédure[5] de contrôle est en place pour recevoir les rapports venant du terrain et une procédure initiale de Qualité est opérée. Tous les lots doivent être accompagnés d'une feuille de contrôle entièrement vérifiable.

Les principales méthodes de saisie de données à partir de formulaires, que l'on peut considérer sont:

1. Entrée manuelle

Cette méthode nécessite un opérateur pour entrer les données directement dans l'ordinateur à partir du formulaire de Recensement. Dans des versions plus sophistiquées de cette approche, l'entrée peut être "assistée par ordinateur" dans laquelle l'opérateur sélectionne une réponse parmi les diverses options affichées sur l'écran. Les vitesses moyennes d'entrée de données varient de 5,000 à 10,000 frappes à l'heure par opérateur[5]. Basé sur le formulaire de recensement de deux pages en Papouasie Nouvelle Guinée en l'an 2000[6], ceci équivaut à peu prés à 10 - 20 formulaires par heure par opérateur.

Avantages: on peut employer du personnel local en grand nombre. Il y aura un grand nombre de PC disponibles pour d'autres utilisations à la fin du recensement. Le logiciel CSPro[7] d'un coût relativement bas ne nécessite pas de licence et permet l'entrée directe de données.

Inconvénients: Nécessite un grand nombre de personnel dont des opérateurs de PC ainsi que des responsables de la gestion d'IT. Nécessite des procédures de contrôle de Qualité tel que pour les Doubles entrées ou les Sondages. Motivation du personnel: comment l'encourager à maintenir une vitesse raisonnable d'entrée durant tout le projet. Logistique et gestion de tout le processus: dans le recensement les volumes sont vastes; dans certains recensements, il peut y avoir plusieurs centaines de tonnes de bulletins à traiter. Nécessite tout un espace physique pour l'installation des systèmes informatiques.

2. Entrée à partir d'images

Cette méthode implique initialement le scanning des formulaires de recensement en utilisant un scanner de document standard et ensuite des opérateurs de PC utilisant l'image produite par le scanner pour entrer alors les données. Les systèmes de contrôle de Qualité utilisés dans l'entrée manuelle peuvent encore être utilisés selon le concept nommé 'Seeding' par DRS. 'Seeding' est la façon dont une image est affichée pour un opérateur et les résultats qui en sont déjà connus et les données entrées par les opérateurs sont comparées avec celles attendues et sont évaluées en conséquence (augmentant l'exactitude potentielle du processus).

Avantages: Comme pour l'entrée manuelle. Cependant cette approche présente la possibilité d'élargir le système pour faire face à des volumes record, en utilisant peut-être des agences spécialisées extraterritoriales pour opérer les entrées pour le compte de l'agence de recensement durant les périodes des volumes maximum. Une archive numérique de tous les formulaires de recensement remplis peut être conservée à la fin de l'opération si on le croit utile.

Inconvénients: Les entrées ne peuvent pas être exécutées avant que les formulaires aient été scannés. Il est nécessaire de posséder un réseau informatique relativement sophistiqué et un organigramme d'opération en place pour gérer le processus d'entrée de données.

3. Lecture de Signes Optiques (OMR)

Des formulaires spécialement conçus et imprimés sont utilisés, chaque formulaire ayant une case à cocher ou une réponse à bulle pour chaque question de recensement. Ils sont normalement scannés sur scanner OMR spéciaux qui

reconnaissent la signification d'un signe sur un formulaire donné et génèrent automatiquement et immédiatement des fichiers de sortie de données exactes.
Avantages: Grande exactitude et très grande vitesse. Coûts prévisibles et définis. Vitesses horaires de traitement de saisie de données, réalistes dans l'ensemble, de l'ordre de 4,000 formulaires OMR par heure peuvent être raisonnablement attendues dans un projet de recensement en direct, en utilisant les scanners spécialisés OMR de DRS.
Inconvénients: nécessite des formulaires spécialement imprimés et des scanners spéciaux, les réponses de case à cocher ne conviennent pas à tous types de questions. Les formulaires ne sont pas faciles à remplir par le public et nécessitent normalement des Enumérateurs formés pour les remplir.

4. Reconnaissance Intelligente de Caractère (ICR)
Les formulaires sont scannés et les images saisies. Les images saisies sont interprétées par le logiciel ICR qui est capable de reconnaître chiffres et lettres écrites dans les cases de réponses sur les formulaires.
Avantages: Les formulaires conçus pour le traitement d'ICR sont relativement faciles à remplir et des formulaires imprimés localement peuvent être utilisés. ICR fonctionne bien avec les caractères numériques
Inconvénients: ne fonctionne pas aussi bien avec les réponses en caractères alphabétiques qui peuvent nécessiter un grand nombre d'interventions manuelles pour en assurer l'exactitude. Le logiciel ICR est incapable de reconnaître toute écriture à la main et n'est pas toujours fiable dans son processus de reconnaissance, il est besoin d'une grande quantité d'interventions manuelles quand la reconnaissance tombe en dessous de certains standards prédéfinis de certitude. En conséquence, il n'est pas facile de prédire exactement la durée ou les coûts du processus de saisie de données. Le logiciel ICR et l'infrastructure informatique nécessaire peuvent être coûteux. Nécessité d'un personnel IT de haut calibre pour le soutien du système ICR.

5. Technologie émergeante, combinant OMR & ICR – quelquefois appelée IMR (Intelligent Mark Recognition)
Il s'agit d'une approche relativement nouvelle pour la saisie de données. Les données des cases à cocher de l'OMR sont reconnues immédiatement et exactement et saisies sur les scanners spéciaux IMR. Les données sont prêtes pour leur importation immédiate dans un logiciel tel que CSPro. En même temps, les images des formulaires sont saisies si nécessaire: par exemple quand une réponse OMR n'est pas appropriée ou quand le scanner OMR a souligné des erreurs de logique ou de validation sur le formulaire. Les Images Numériques de ces formulaires sont alors envoyées au logiciel ICR ou au logiciel d'entrée manuelle pour la saisie des champs pertinents.
Avantages: Combine les bénéfices de la technologie "traditionnelle" OMR avec le potentiel d'utiliser les toutes dernières techniques d'ICR. Une archive d'images de tous les formulaires de recensement scannés est créée automatiquement. On peut obtenir un résultat fiable et prévisible par un investissement donné dans ces techniques
Inconvénients: Cette technique nécessite encore des formulaires spécialement imprimés.

Le succès d'une saisie de données de recensement à partir de papier dépend de façon substantielle de cinq facteurs cruciaux:

Facteur crucial de succès No. I: Equipement de scanning

Si le scanning de document va être exécuté, il est nécessaire de sélectionner le bon scanner pour l'opération. Les organisations de recensement doivent choisir un fabricant de scanners en mesure de proposer un scanner qui s'est révélé être entièrement à la hauteur de la tâche lors de projets similaires de recensement.

On doit considérer les impératifs suivants:

- Les scanners doivent être conçus en gardant à l'esprit un long cycle d'opération pour assurer une fiabilité générale et cohérence dans la saisie des données pendant des périodes de scanning ininterrompues ou prolongées.
- Si un formulaire OMR va être utilisé et s'il est prévu de traiter ce formulaire sur un scanner dédié OMR, alors ce scanner doit être construit en métal à usage industriel pour assurer l'exactitude de lecture durant le processus de scanning tout entier.
- Le scanner doit être capable de permettre l'extraction réussie de tout formulaire de recensement coincé; un acheminement de papier, ouvert, serait souhaitable.
- Le scanner doit être capable de détecter et prévenir l'alimentation de formulaires en double durant le scanning.
- Le scanner doit être capable d'être interconnecté pour s'intégrer avec les systèmes nécessaires de logiciel et les serveurs.
- Le scanner doit avoir une interface facile à utiliser et des consommables faciles à remplacer.
- On devra réfléchir longuement à la disponibilité et expérience des partenaires locaux pour apporter le soutien technique et service complets si nécessaire, dans l'idéal les ingénieurs locaux de soutien devraient avoir été formés au siège social du fabricant.
- Il est recommandé d'utiliser les scanners qui ont déjà été en fait utilisés dans un recensement précédent.

Il y a aussi des avantages à choisir un scanner qui peut:

- Acheminer les formulaires vers divers plateaux de sortie selon certaines règles prédéfinies
- Accomplir des débits à grande vitesse durant les chargements maximum dans le processus de scanning

Le meilleure pratique dicte qu'un plan d'urgence soit mis en place en cas de panne de scanner. Il est recommandé qu'on installe plus de scanners que le nombre requis. Cette approche présente deux avantages principaux:

- Tout scanner supplémentaire peut être utilisé durant le traitement des formulaires (ainsi le traitement du recensement peut se trouver plus avancé que prévu)
- Si quelque scanner tombe en panne, son effet en sera minime sur le planning (comme les remplacements sont déjà prêts et en état de marche)

Facteur Crucial de succès No 2: Conception du formulaire de recensement

Le temps et effort investis pour assurer que les formulaires de recensement sont remplis aussi exactement que possible et retournés dans le meilleur état possible paieront des dividendes significatifs durant l'opération de saisie de données.

- La conception du formulaire dépend de qui remplira le formulaire et aussi de la technique de saisie de données qu'on a l'intention d'utiliser. Les formulaires conçus pour être remplis par les chefs de famille, sont souvent plus longs en vertu de la nécessité d'y inclure des instructions (peut-être une brochure de recensement de 8 pages ou plus). Les formulaires qui seront remplis par des Enumérateurs peuvent être plus complexes et souvent beaucoup plus courts (peut-être même un seul côté d'une page A4 par ménage dans le cas de formulaires OMR avec des économies de coûts en conséquence).
- La qualité du papier doit se conformer et excéder des standards minimum; à la fois du point de vue du maniement du papier sur le territoire et aussi du point de vue des exigences des divers fabricants de scanners et des techniques de scanning.

A la fois les fabricants de scanners OMR et les fabricants de scanners d'Image recommandent que le papier se conforme à certains standards de performance prescrits par les organismes tels que ISO, BS, BSEN.

Ces standards définissent les caractéristiques suivantes:

- Grammage, Epaisseur, Rugosité Bendsten, Friction statique, Rigidité, Résistance à l'air, Résistance à déchirure interne, Endurance aux pliures, Réflecteur de Lumière, Opacité, Sens du grain, Plis & Faux Plis, Humidité

DRS fournit en principe un papier conçu en utilisant un mélange spécifique de ces standards, basé sur le standard généralement disponible APAC[8] de CBS2

Il serait hautement recommandé d'engager une discussion sur la qualité du papier avec le fournisseur du scanner et le fournisseur de papier pour identifier les standards les plus appropriés avant que la production ne commence. L'environnement dans lequel le formulaire sera utilisé peut dicter la spécification de chacun de ces standards. Par exemple si un formulaire va être utilisé dans un environnement humide, le client devra considérer l'impact et l'effet que ceci peut avoir sur tout type de papier sélectionné pour assurer une exécution exacte et un bon scanning. Les exigences de qualité de papier pour les scanners OMR sont plus rigoureux que pour les scanners d'image, à la fois sur les caractéristiques de papier ci-dessus mais aussi sur le type d'encre qui peut être utilisé et l'exactitude du procédé d'impression.

Les meilleurs résultats utilisant la technologie OMR pour le recensement ont été obtenus quand on a utilisé les formulaires importés fournis directement par le fabricant de scanner. Le fabricant de scanner accepte alors toute responsabilité et garantit que leur formulaire pourra être scanné et reconnu sur les scanners OMR.

Il est avantageux quand on choisit les techniques OMR et ICR de faire imprimer les formulaires avec des codes barre uniques et séquentiels. Un code barre séquentiel permet de stocker toute l'information sur le formulaire et de la référencer contre le numéro individuel de code barre de ces formulaires. L'impression de ces codes barre séquentiels n'est pas une technique simple; elle nécessite des imprimeurs possédant un haut niveau de compétences. Il est recommandé que les imprimeurs soient qualifiés au standard ISO 9001:2000 pour l'Impression spécialisée de formulaires.

Facteur Crucial de Succès No 3: Formation & Soutien

Quelle que soit la technique utilisée pour la saisie des données à partir de formulaires, les techniques par elles-mêmes ne peuvent transformer de "mauvais formulaires" en "bons formulaires". Si les formulaires ont été mal remplis, alors les coûts monteront en flèche et les retards se produiront et on stockera des résultats inexacts. Evidemment la meilleure solution est de faire remplir les formulaires correctement en premier lieu:

- La formation des Enumérateurs est fondamentale pour réussir. Utiliser les bons crayons, hachurer les bulles correctement et garder les formulaires au sec. Traiter les formulaires avec respect
- Motiver les Enumérateurs – pas seulement avec de l'argent mais aussi avec d'autres valeurs plus élevées.

Le soutien local par l'intermédiaire du fournisseur et/ou ses partenaires formés et accrédités durant le traitement des formulaires de recensement aidera à obtenir les meilleurs résultats. L'expérience a montré que la formation correcte des opérateurs de scanner améliore sensiblement le débit de formulaires dans le processus de l'organigramme d'opération.

Composant 4: Matériel Informatique

Une évaluation complète du matériel informatique actuel disponible à l'organisation du recensement doit être faîte tôt dans le processus de sélection de technique de saisie de données. Des approches différentes de saisie de données nécessitent des serveurs de PC et des exigences de mémorisation différentes, et il est peu probable que les PC actuels de l'organisation du recensement soient totalement appropriés aux exigences des toutes dernières techniques de saisie de données à grande vitesse. Il faudra inévitablement quelque investissement dans ce domaine quelle que soit la technique utilisée. Toutefois, comparativement, le matériel informatique et le stockage de données sont maintenant beaucoup moins chers comparés aux années précédentes. L'investissement en logiciel à la fois le logiciel de saisie de données et aussi le logiciel de stockage et de récupération des données. On devra considérer le montant de données à stocker et on devra employer un 'backup' de sauvegarde adéquat dans tous les systèmes de serveur pour protéger les données et les images. On devra aussi tenir compte de la continuité de l'électricité locale quand on spécifiera du matériel informatique neuf et les alimentations électriques 'ininterrompables' (UPS) sont potentiellement la meilleure option.

Composant 5: Logiciel & Organigramme d'Opération

L'organigramme pour l'opération de traitement de données sera adapté aux circonstances individuelles d'une administration de recensement. En conséquence les fournisseurs doivent être flexibles dans leur approche plutôt que rigides dans leur solution. Il est recommandé que durant la période de recensement-pilote, on procède à un bilan pour obtenir les meilleurs résultats de l'organigramme d'opération. Celle-ci peut inclure:

- Bouger le stockage plus prés de l'aire de traitement de données
- Exécuter des travaux d'audit et des systèmes/logiciel de rapportage qui suivent les lots tout au long du processus
- Désigner ou changer les rôles et fonctions du personnel
- Réexaminer la sécurité des données à chaque stade du processus
- Ajouter ou ajuster les procédés de contrôle de qualité

Le logiciel sélectionné pour le traitement de recensement doit être compatible avec l'ensemble des compétences techniques utilisées à l'intérieur de l'Administration de recensement. Il est crucial d'essayer d'aligner les compétences locales et le logiciel aux procédés nécessaires pour tirer le maximum du système.

Une bonne compréhension locale du logiciel de saisie de données et des systèmes d'opération sur lequel il est utilisé accroîtra le niveau de succès et permettra une résolution prompte de tout défi pendant l'exercice de traitement de données. L'utilisation de progiciels standards qui ont fait leurs preuves sur le marché du recensement réduira les risques, donnera confiance aux utilisateurs et potentiellement permettra d'utiliser efficacement l'information et les expériences accumulées à l'occasion d'autres exercices de traitement de recensement.

Résumé

La décision finale du choix de technique de saisie de données sera un compromis qui visera à équilibrer les éléments de décision souvent en compétition. Un équilibre doit être réalisé qui aura tenu compte de tous les cinq facteurs cruciaux de succès examinés et qui reflètera aussi les facteurs suivants:

- Budgets disponibles et financement
- Traditions culturelles et méthodes utilisées précédemment
- La Géographie locale
- Vitesse totale nécessaire de l'opération de traitement de données
- L'économie locale, l'infrastructure locale et les capacités logistiques
- L'ensemble des compétences du personnel local et leur expérience.
- L'infrastructure en place actuellement à l'intérieur de l'Administration du recensement
- La bonne volonté et faisabilité quant à l'utilisation de formulaires importés/la disponibilité de formulaires de haute qualité imprimés localement.
- Le besoin d'utiliser des conceptions de formulaire qui peuvent être compris par la population locale
- Le niveau d'éducation des Enumérateurs ou de la population. Ceci aidera à décider la mise en page potentielle des formulaires et ensuite la meilleure méthode de saisie de données.
- Facteurs historiques et culturels.
- Information et expérience gagnées pendant tout exercice pilote
- Les expériences d'autres administrations de recensement et d'organismes membres géographiques

Si une décision est prise d'utiliser des scanners ordinaires d'image pour le traitement de données de recensement, quand l'intention est de saisir les données en utilisant le logiciel ICR ou les techniques d'entrée à partir d'image, alors ces formulaires peuvent normalement être imprimés localement. Quand on fait imprimer les formulaires ICR localement, il pourrait paraître y avoir une économie de coût au premier abord comparée aux coûts d'importer des formulaires spécialement imprimés – *toutefois attention en faisant ce calcul*. L'économie immédiate de coûts en utilisant des formulaires produits localement dans la solution ICR/entrée risque d'entrainer des coûts totaux beaucoup plus élevés

dans les autres stades du recensement- par ex. coûts de logiciel beaucoup plus élevés et potentiellement un temps beaucoup plus long dans la saisie de données.

Une solution OMR a le grand avantage de présenter un coût de saisie de données clairement défini et calculable. En comparaison, les coûts d'une solution de saisie de données ICR (même si les formulaires pourraient être moins chers à imprimer initialement), peuvent être beaucoup plus difficiles à estimer exactement à l'avance du recensement. Néanmoins la facilité de remplissage des formulaires ICR font des techniques ICR une option de plus en plus attrayante pour les Administrations de recensement. Cependant pour toute Administration de recensement pensant à implémenter les techniques de saisie de données ICR, il faudra inévitablement accepter une augmentation des effectifs possédant des compétences poussées en IT comparé au nombre de personnel nécessaire pour gérer les autres techniques. Il est intéressant de noter que le Bureau Australien de statistiques[9] a annoncé publiquement qu'il a plus de 400 effectifs IT. Inévitablement, les Administrations de recensement avec moins de ressources IT peuvent prendre des décisions différentes sur les méthodes de saisie de données à partir de leurs formulaires par rapport à ces Administrations de recensement qui ont plus de ressources à leur disposition.

DRS Data Services Limited a l'expertise de toutes les cinq méthodes de saisie de données qui ont été examinées dans ce document. Chaque projet de recensement auquel DRS participe est évalué sur ses propres mérites et la société peut conseiller et recommander la meilleure méthode de saisie de données pour chaque projet, basé sur les particulières circonstances locales actuelles. Après une revue en profondeur de la technologie, les deux projets de recensement les plus récents auxquels DRS a participé, utilisent l'approche combinée:

L'Agence Centrale de Statistiques (CSA) d'Ethiopie a accordé un contrat à DRS en 2006 pour fournir une solution combinée de traitement de données de recensement OMR/Image en 2007 (utilisant la technique "IMR") Ils ont entrepris un recensement-pilote en 2006 avec une technologie combinée. CSA a été très satisfait des résultats de l'exercice pilote et ultérieurement vont utiliser ce système combiné pour le recensement national en 2007. Ils ont choisi spécifiquement d'utiliser ce système combiné car ils ont rencontré beaucoup de difficultés dans l'entrée manuelle des données de leur recensement national précédent. Il a été décidé après l'exercice pilote de rationaliser certaines des règles de validation professionnelle pour les formulaires pendant le traitement. Ceci a simplifié le processus de validation et de vérification des données à la saisie.

DRS a aussi été attribué le projet de traitement de données de recensement-pilote au Soudan utilisant une technologie similaire, qui sera achevé à la mi-année 2007 dans l'intention de traiter le recensement national 2007/08.

---

[1] Auteur – Directeur International, DRS Data Services Ltd, UK.

[2] Co Auteur – Directeur International Régional, DRS Data Services Ltd, UK

[3] DRS Data Services Limited est le principal fournisseur international de produits et services de scanning. La société a une expertise particulière dans les domaines tels que recensements, inscriptions d'élections et solutions d'examens. DRS a plus de 35 ans d'expérience à la fois au RU et Internationalement

[4] Définition dans le rapport UNSD '*Principles & Recommendations for Population & Housing Censuses (Series M, No.67/Rev 1,1998)'*

[5] Description de procédés recommandés dans le rapport UNSD *Principles & Recommendations for Vital Statistics Systems - Control of receipt of statistical reports (Series M, No.19/Rev 2, 2001)*

[6] UNSTATS - *http://unstats.un.org/unsd/demographic/sources/census/censusquest.htm#P*

[7] Donné par le Bureau de Statistiques US *CSPro (Census and Survey Processing System) is a public-domain software package for entering, editing, tabulating and mapping census and survey data*

[8] APACS est une association commerciale pour les paiements et le secteur bancaire au RU – *www.apacs.gov.uk*

[9] 14ème conférence du Commonwealth Britannique sur la gestion des statistiques en Afrique du Sud 2005 –' *Recent Advances in the Use and Management of Technology at the Australian Bureau of Statistics'*

.