Note de recherche — Research Note

La saisie et la vérification semi-automatisée des données quantitatives

Mark Olsen et Phyllis LeBlanc*

La saisie des données exige beaucoup de temps et de travail de la part des praticiens de l'histoire quantitative. Le recours à un lecteur optique de caractères et un logiciel écrit en SNOBOL4 et capable de vérifier l'exactitude des données recueillies peut faciliter leur tâche. Voilà ce que les auteurs entendent montrer en utilisant comme exemple leur traitement des rôles d'évaluation de Moncton (Nouveau-Brunswick) de 1919 à 1929.

Assessment rolls constitute an important source to social historians. Our analysis of the published Moncton assessment rolls for 1919 and 1929, using an optical character reader and a software package written in SNOBOL4 as a semi-automatic tool for data verification, is presented here in the interest of historians working in the field of quantification and social history.

La saisie et la vérification des données posent de grands défis à l'utilisation de l'informatique en recherche historique. L'enregistrement des données par des systèmes usuels, les cartes perforées ou le système de traitement de texte, est un travail lent qui n'est pas dénué d'erreurs. Cependant, des logiciels ont été conçus pour aider le chercheur à entrer ses données sur ordinateur plus rapidement et avec plus de précision[1]. Mais le recours au lecteur optique, qui fait une lecture directe des documents, est la façon la plus efficace de créer une banque de données. De nombreux progrès au niveau de la technologie des lecteurs optiques de caractères (Optical Character Readers) ont été réalisés depuis une dizaine d'années. Au cours de travaux récents nous avons constaté que la collecte des données quantitatives avec un lecteur et un logiciel de vérification est très efficace pour certains types de documents.

Les lecteurs optiques sont utilisés fréquemment pour lire directement des textes imprimés ou dactylographiés, tel le manuscrit de livre[2]. L'opération est réalisée en deux

* Mark Olsen est directeur adjoint du projet ARTFL à l'Université de Chicago et candidat au doctorat en histoire à l'Université d'Ottawa. Phyllis E. LeBlanc est professeure adjointe au département d'histoire de l'Université de Winnipeg.

1. M. Olsen a élaboré des logiciels pour la collecte des données chiffrées et textuelles. Voir « DATCOL4 : Interactive Data Collector for Historians », dans *Computers and the Humanities*, 19, 3 (1985), et « TXTCOL4 : String Data Collection Program », dans *Computers and the Social Sciences*, 1,2 (1985). Voir aussi M. Olsen, « Computer Keys Data Entry Emulator », à paraître dans *Computers and the Humanities* (1988).

2. Nous avons utilisé un « Kurzweil Data Entry Machine 4 000 ». Les lecteurs disponibles aux États-Unis sont présentés par Tom Stanton, dans « Peripheral Vision : A Guide to Optical Character Readers », paru dans *PC : The Independent Guide to IBM Personal Computers*, 4 :14 (July, 1985). Une explication très utile du fonctionnement du Kurzweil 4 000 a été présentée par Susan Hockey, « The Kurzweil Data Entry Machine », dans *Literary and Linguistic Computing*, 1 (1986).

étapes distinctes. En premier lieu, le lecteur fait une image digitalisée de chaque caractère et l'enregistre dans la mémoire de l'ordinateur. L'exemplaire du texte doit être clair et propre pour obtenir une image bien définie. En deuxième lieu, le logiciel qui contrôle le système doit définir quel caractère sera représenté en fonction de l'image créée. Les lecteurs puissants, tel le Kurzweil 4 000, sont capables d'apprendre les nouveaux caractères, ainsi que les caractères imprimés en jeux différents. L'image réelle du caractère est comparée aux images constitutives (Multiple Property Descriptors) réalisées par le système. Suivant un algorithme de comparaison, l'ordinateur choisit l'image constitutive la plus rapprochée de l'image actuelle. Le Kurzweil 4 000 peut lire entre 20 et 50 pages à l'heure. La vitesse de lecture et la précision de la copie électronique dépendent entièrement du nombre de caractères et du nombre de jeux, ainsi que de la qualité du document. Si la photocopie est de qualité supérieure, le lecteur lit le texte facilement en faisant un minimum d'erreurs.

Nous utilisons dans l'exemple d'analyse présenté ici les rôles d'évaluation de la ville de Moncton des premières décennies du XX[e] siècle[3]. Nous recourons à ces documents dans le cadre d'une étude de la croissance de la communauté francophone à Moncton, dans le contexte du nationalisme acadien de la fin du XIX[e] et du début du XX[e] siècles, nationalisme empreint alors des valeurs rurales traditionnelles. Jusqu'ici, les historiens limitaient leurs recherches sur la communauté acadienne de Moncton aux sources qualitatives. Ils reflétaient ainsi les préjugés de l'idéologie contemporaine. Nous adoptons une approche plus systématique qui comprend une analyse des structures économiques et de la géographie urbaine, afin d'évaluer l'importance relative de la rhétorique nationaliste dans la formation d'une communauté francophone à Moncton pendant cette période[4]. D'une part, la présence active de la population francophone dans les structures économiques et sociales de Moncton montre un degré élevé d'enracinement dans la communauté. D'autre part, l'existence de ghettos francophones ou anglophones dans la ville est un facteur qui pourrait influencer le degré de pénétration de l'idéologie nationaliste dans la société francophone de Moncton.

Les rôles d'évaluation peuvent être fort utiles dans l'étude des hypothèses présentées. Ce genre de source est fréquemment employé par les historiens du secteur socio-économique[5]. Les rôles d'évaluation peuvent indiquer la richesse relative d'une population et, dans notre cas, ils nous permettent de préciser le statut socio-économique de chaque groupe ethnique de la ville de Moncton. De plus, ils peuvent être jumelés par l'informatique

3. Nous avons utilisé pour cette analyse deux documents : *Assessment List City of Moncton, 1919* (Archives publiques du Nouveau-Brunswick : MMN 6/2/1) et *Assessment List City of Moncton, 1929* (Archives publiques du Nouveau-Brunswick : MMN 6/2/1). La liste de 1919 est composée d'environ 5 800 lignes ou cas de données, et on retrouve environ 7 500 cas de données dans le document de 1929.

4. Une étude quantitative antérieure, à partir des annuaires de Moncton pour la période 1896 à 1938, nous a permis d'analyser les structures économiques et la géographie urbaine de Moncton. Nous avons tiré de cette étude des conclusions importantes sur l'influence de l'idéologie nationaliste de l'époque sur le groupe francophone urbain de Moncton. Voir P.-E. LeBlanc, « Économie et société en transition : un aperçu de Moncton, 1870-1938 », communication présentée au Congrès de l'Institut d'histoire de l'Amérique française, 17-18 octobre 1986; P.-E. LeBlanc et Mark Olsen, « Ethnic Geography and Industrialization in Moncton, 1890-1940 », communication au congrès *Atlantic Canada Workshop*, septembre 1986.

5. Il y a plusieurs cas d'études à partir des rôles d'évaluation. Voir l'étude de M.B. Katz, Michael J. Doucet et Mark J. Stern, *The Social Organization of Early Industrial Capitalism*, Harvard University Press, Cambridge, Mass., 1982; voir aussi David DeBrou, « Voting Behaviour of the Electorate of Haute-Ville de Québec, 1814-1836 : Ethnicity, Electoral Choice and Occupation », communication présentée à la réunion annuelle du Social Science History Association, Toronto, 25-28 octobre 1984.

à d'autres sources nominatives[6]. Malheureusement, les exemplaires des rôles d'évaluation des Archives provinciales du Nouveau-Brunswick à Fredericton sont noircis, ce qui a réduit la qualité des photocopies et l'efficacité du Kurzweil 4 000 (voir annexe 1).

Les rôles d'évaluation de la ville de Moncton pendant la période à l'étude nous offrent les renseignements suivants : le nom de famille de la personne, son prénom ou ses initiales, le numéro et le nom de la rue des résidents de Moncton et, dans le cas de non-résidents, l'adresse à l'extérieur de la ville. Pour l'année 1919, la valeur de la propriété immobilière et la valeur de la propriété personnelle sont inscrites, ainsi que le revenu net, le montant imposé et la taxe. Cependant, en 1929, le montant imposé a été remplacé par l'impôt personnel. Comme le démontre le tableau 1, les chiffres sont isolés les uns des autres par des lignes verticales qui délimitent aussi les colonnes vides, lors de l'absence des valeurs numériques. Dans les deux documents, les femmes sont généralement identifiées par l'appellation « Miss » ou « Mrs ».

Le lecteur a réussi à déchiffrer les deux documents en moins de huit heures. Nous avons tenté d'inciter le lecteur à lire les lignes verticales comme délimitateurs des variables, mais la qualité inférieure des photocopies utilisées a contrecarré cet effort. Nous avons suggéré au lecteur, dans le cas du document de 1929, d'éliminer les caractères étroits dus à l'utilisation d'une photocopie de mauvaise qualité. Le système a éliminé en même temps quelques caractères, en particulier les « i » et parfois les « l ». Les fichiers créés par le lecteur ont été transmis à un micro-ordinateur où les données ont été corrigées et codifiées en vue d'une analyse statistique.

Bien que réalisée en une seule étape, la correction des données se fait sur trois plans : a) correction des fautes de lecture; b) définition de chaque ligne en variables définies; c) addition des valeurs qui représentent le sexe, l'ethnicité et le lieu de résidence de la personne (voir annexe 2). Plusieurs erreurs de la machine, telle l'omission de certains caractères, ont été répétées systématiquement. Elles ont pu être corrigées par la fonction « global search and replace » du traitement de texte. Mais la plupart des erreurs ont dû être corrigées manuellement, une à la fois. Nous avons divisé chaque ligne en sept champs ou variables. Chaque champ est délimité par une oblique. Un espace vide entre deux obliques indique un champ sans valeur. Trois nouvelles variables ont été ajoutées pour chaque cas étudié : le sexe, l'ethnicité et le secteur de résidence de la personne. Le sexe et l'ethnicité sont déterminés par le nom, le prénom et l'appellation, tandis que le secteur de résidence est déterminé par l'adresse en utilisant un livre de codes, créé lors d'une étude antérieure.

La correction et l'identification des données ont été les opérations les plus lentes de notre étude, mais elles étaient néanmoins nécessaires. Elles ont été difficiles du fait de la qualité inférieure de nos documents et à cause de la nécessité d'ajouter les codes qui représentent l'ethnicité, le sexe et l'adresse, que la lecture informatisée n'a pu fournir. Le processus de correction des données des deux documents a exigé environ une cinquantaine d'heures de travail. Une meilleure qualité des photocopies aurait considérablement réduit

6. Voir, entre autres, E. A. Wrigley, *Identifying People in the Past* (London, 1973), et Ian Winchester, « The Linkage of Historical Records by Man and Computer », dans *Journal of Interdisciplinary History* 1 (1970). David De Brou et Mark Olsen ont présenté un système permettant le jumelage automatique des populations canadiennes dans « The Guth Algorithm and the Nominal Record Linkage of Multi-Ethnic Populations », dans *Historical Methods* 1,19 (1986).

le temps nécessaire pour ce processus. Par contre, une collecte manuelle de 13 000 cas de données aurait pu prendre plus de 250 heures[7].

Une fois la correction manuelle des données effectuée, nous avons élaboré un logiciel qui a créé un nouveau fichier dans un format fixe que le « SPSS » peut analyser. Ecrit en SNOBOL4, le logiciel a mis chaque variable dans des colonnes fixes en utilisant une ligne par carte[8]. Avant de transférer chaque variable au nouveau fichier, le logiciel a effectué plusieurs examens de chaque nombre. Le premier examen devait vérifier si le nombre exact de variables délimitées par une oblique était bien inscrit. Le logiciel a examiné chaque nombre pour trouver les caractères alphanumériques, qu'il a remplacés par un chiffre qui lui ressemble visuellement, puisque cette erreur du lecteur n'a pas été corrigée lors du processus de vérification. La table de correction a remplacé, par exemple, « O » par « 0 », « B » par « 8 », « l » par « 1 » et « b » par « 6 ». Le système a aussi éliminé d'autres caractères, tels les virgules et les points dans chaque nombre. Enfin, si la variable a été indiquée par un chiffre, le logiciel a fait une comparaison entre la valeur de ce chiffre et celle d'une échelle prédéterminée pour chaque variable. Chaque fois que le chiffre donné ne rencontrait pas les valeurs assignées par l'échelle, le logiciel a rejeté le cas. Chaque cas non corrigé automatiquement par l'ordinateur a été classé dans un fichier spécial, accompagné d'une description de l'erreur, afin de permettre une vérification manuelle. Ce logiciel, conçu en quelques heures à peine, a non seulement effectué plusieurs corrections, mais il a trouvé des erreurs qu'il n'a pu cependant corriger[9].

Il est évident que pour la saisie des données, un système semi-automatique de ce genre est nettement supérieur aux moyens traditionnels. Nos recherches ont montré que l'on peut épargner entre 50 et 70 % du temps exigé par les méthodes non informatisées. Quelques institutions canadiennes d'enseignement supérieur ont déjà fait l'acquisition d'un lecteur optique de caractères. D'autres suivront, car cette technologie offre aux historiens et aux chercheurs la possibilité de créer de grandes banques de données à partir de sources imprimées plus rapidement et à un moindre coût que les méthodes usuelles. Ces améliorations accroissent énormément l'attrait des sources quantifiables pour les historiens de la période contemporaine.

Les lecteurs optiques conservent cependant certaines lacunes dont il faut tenir compte. Ils ne peuvent lire les documents manuscrits, tels les registres paroissiaux. De plus, l'efficacité du lecteur est reliée directement à la qualité du document, qu'il s'agisse d'un original ou d'une photocopie. Malgré le fait que les lecteurs peuvent lire des données en effectuant un minimum d'erreurs, il reste indispensable de vérifier la lecture. Enfin, l'informatique reste le moyen le plus sûr et le plus efficace pour trouver les erreurs et les corriger.

---

7. Nos calculs, basés sur un rendement de 50 cas à l'heure (13 000 / 50 = 260 heures), ne tiennent pas compte du temps exigé pour la correction des erreurs. L'estimation de 50 cas à l'heure nous paraît raisonnable pour la saisie des données avec un logiciel de collecte.

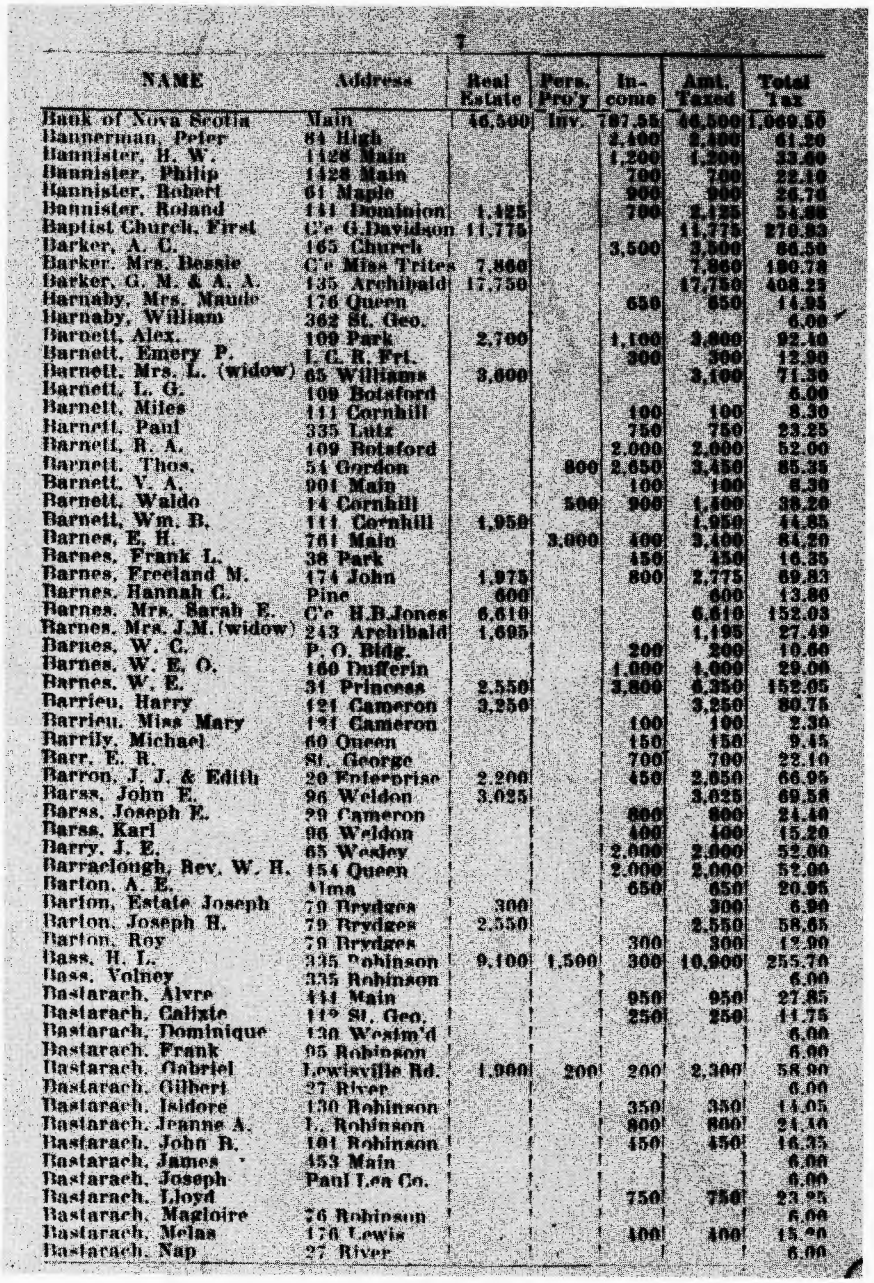
8. SNOBOL4 est un langage informatique créé spécialement pour la manipulation des données textuelles ou, selon le langage technique, le « string processing ». Voir R.E. Griswold *et al.*, *The SNOBOL4 Programming Language* (Englewood Cliffs, NJ, 1971), et Susan Hockey, *SNOBOL Programming for the Humanities* (Oxford, 1985). Le langage SNOBOL4 est disponible pour les micro-ordinateurs IBM. Voir M. Olsen, « SNOBOL4+ » dans *Computers and the Humanities* 2,19 (1985). D'autres langages informatiques, tels BASIC et TURBO PASCAL, peuvent aussi accommoder la manipulation de données textuelles.

9. Élaborer un logiciel comme celui-ci ne pose aucune difficulté. Malheureusement, on ne peut en généraliser l'utilisation pour la vérification des données quantitatives par d'autres chercheurs, parce que leurs fichiers de données comportent des structures différentes. Il semble qu'il est plus simple de créer un logiciel pour chaque banque de données.

Malgré ces limites, la technologie des lecteurs optiques, bien connue des chercheurs en langues et en littérature, est aussi très utile aux historiens du secteur socio-économique de la période contemporaine. Cette technologie peut être particulièrement utile pour la saisie de textes imprimés, tels les recensements publiés.

Annexe 1

Fac-similé d'une page de l'exemplaire utilisé du rôle d'évaluation de la ville de Moncton en 1929



7

NAME	Address	Real Estate	Pers. Pro'y	In-come	Amt. Taxed	Total Tax
Bank of Nova Scotia	Main	46,500	Inv.	787.55	46,500	1,069.50
Bannerman, Peter	84 High			2,400	2,400	61.20
Bannister, H. W.	1428 Main			1,200	1,200	33.60
Bannister, Philip	1428 Main			700	700	22.10
Bannister, Robert	61 Maple			900	900	26.70
Bannister, Roland	141 Dominion	1,425		700	2,125	54.88
Baptist Church, First	C'e G.Davidson	11,775			11,775	270.83
Barker, A. C.	165 Church			3,500	3,500	66.50
Barker, Mrs. Bessie	C'e Miss Trites	7,860			7,860	190.78
Barker, G. M. & A. A.	135 Archibald	17,750			17,750	408.25
Barnaby, Mrs. Maude	176 Queen			650	650	14.95
Barnaby, William	362 St. Geo.					6.00
Barnett, Alex.	109 Park	2,700		1,100	3,800	92.10
Barnett, Emery P.	L.C.R. Frt.			300	300	12.90
Barnett, Mrs. L. (widow)	65 Williams	3,000			3,100	71.30
Barnett, L. G.	109 Botsford					6.00
Barnett, Miles	111 Cornhill			100	100	8.30
Barnett, Paul	335 Lutz			750	750	23.25
Barnett, R. A.	109 Botsford			2,000	2,000	52.00
Barnett, Thos.	54 Gordon		800	2,650	3,450	85.35
Barnett, V. A.	901 Main			100	100	8.30
Barnett, Waldo	14 Cornhill		500	900	1,400	38.20
Barnett, Wm. B.	111 Cornhill	1,950			1,950	44.85
Barnes, E. H.	761 Main		3,000	400	3,400	84.20
Barnes, Frank L.	38 Park			450	450	16.35
Barnes, Freeland M.	174 John	1,975		800	2,775	69.83
Barnes, Hannah C.	Pine	600			600	13.80
Barnes, Mrs. Sarah E.	C'e H.B.Jones	6,610			6,610	152.03
Barnes, Mrs. J.M. (widow)	243 Archibald	1,695			1,195	27.49
Barnes, W. C.	P. O. Bldg.			200	200	10.60
Barnes, W. E. O.	160 Dufferin			1,000	1,000	29.00
Barnes, W. E.	31 Princess	2,550		3,800	6,350	152.05
Barrieu, Harry	121 Cameron	3,250			3,250	80.75
Barrieu, Miss Mary	121 Cameron			100	100	2.30
Barrily, Michael	60 Queen			150	150	9.45
Barr, E. R.	St. George			700	700	22.10
Barron, J. J. & Edith	20 Enterprise	2,200		450	2,650	66.95
Barss, John E.	96 Weldon	3,025			3,025	69.58
Barss, Joseph E.	29 Cameron			800	800	24.40
Barss, Karl	96 Weldon			400	400	15.20
Barry, J. E.	65 Wesley			2,000	2,000	52.00
Barraclough, Rev. W. H.	154 Queen			2,000	2,000	52.00
Barton, A. E.	Alma			650	650	20.95
Barton, Estate Joseph	79 Brydges	300			300	6.90
Barton, Joseph H.	79 Brydges	2,550			2,550	58.65
Barton, Roy	79 Brydges			300	300	12.90
Bass, H. L.	335 Robinson	9,100	1,500	300	10,900	255.70
Bass, Volney	335 Robinson					6.00
Bastarach, Alvre	141 Main			950	950	27.85
Bastarach, Calixte	119 St. Geo.			250	250	11.75
Bastarach, Dominique	130 Westm'd					6.00
Bastarach, Frank	95 Robinson					6.00
Bastarach, Gabriel	Lewisville Rd.	1,900	200	200	2,300	58.90
Bastarach, Gilbert	27 River					6.00
Bastarach, Isidore	130 Robinson			350	350	14.05
Bastarach, Jeanne A.	L. Robinson			800	800	24.40
Bastarach, John B.	101 Robinson			450	450	16.35
Bastarach, James	453 Main					6.00
Bastarach, Joseph	Paul Lea Co.					6.00
Bastarach, Lloyd				750	750	23.25
Bastarach, Magloire	76 Robinson					6.00
Bastarach, Melas	176 Lewis			400	400	15.20
Bastarach, Nap	27 River					6.00

Annexe 2

Ville de Moncton, rôle d'évaluation de 1919

Exemples des données corrigées et séparées avant vérification par ordinateur. Une oblique sépare les champs les uns des autres.

Bank of Nova Scotia/ Main/ 46500 / / 787.55 / 46500 /1069.50/ / /04
Bannerman, Peter/ 84 High/ / / 2400/ 2 400/61.20/M/A/13
Bannister, H. W. / 428 Main/ / / 1200 /1200/ 33.60/M/A/05
Bannister, Philip/ 1428 Main/ / /700/ 700/ 22.10/M/A/08
Bannister, Robert/ 61 Maple/ / /900/ 900/ 26.70/M/A/01
Bannister, Roland/ 141 Dominion/ 1425/ /700/2,125/ 54.88/M/A/12
Baptist Church,/ FirstC e G.Davidson/ 11775/ / /11,775/270.83/ / /00
Barker A. C./ 165 Church/ / /3500/ 3,500/ 86.50/M/A/14
Barker Mrs Bessie/ C e Miss Trites/7860/ / /7,860/180 78/F/A/00
Barker G. M. & A. A./ 135 Archibald/ 17750/ / /17750/408.25/ /A/09
Barnaby, Mrs Maude/ 176 Queen/ / /650 /650/ 16.95/F/A/06
Barnaby, William/ 362 St. Geo/ / / / /6.00/M/A/23
Barnett, Alex./ 109 Park / 2700/ / 1100/ 3800/ 92.40/M/A/18
Barnett, Emery P./ I. C. R Frt/ / / 300/ 300/ 12.90/M/A/00
Barnett, Mrs L. (widow)/ 65 Williams /3600 / / / 3100/71.30/F/A/12
Barnett, L G/ 109 Botiford/ / / / /6.00/M/A/10
Barnett, Miles/ 111 Cornhill/ / /100 / 100/ 8.30/M/A/13
Barnett, Paul/ 335 Lutz/750/750 / / / 23.25/M/A/34
Barnett, R. A. / 109 Botsford/2 000/ /2,000/ / 52.00/M/A/16