Notes sur la numérisation de ma thèse de Ph.D., « The Merchants and Négociants of Montreal, 1750-1775: A Study in Socio-Economic History », Michigan State University, 1974.

José E. Igartua

Mai 2016

Un collègue m'a demandé il y a quelque temps si j'avais une version PDF de ma thèse de Ph.D. Je n'en avais pas. Il a suggéré que cela pourrait être utile. Je m'y suis donc mis, selon la démarche suivante.

Ma thèse avait été tapée en 1974 sur une machine à écrire électronique IBM. Elle permettait des tailles de caractères variées et faisait l'espacement proportionnel même de polices Courier 12. Il y avait donc des appels de note surélevés dans le texte et dans les notes de bas de page, du texte souligné, et du texte en exposant, comme dans « $C^{11}A$ ».

À partir d'un exemplaire en feuilles détachées, j'ai fait la saisie numérique et la reconnaissance optique du texte pour la création d'un fichier PDF contenant le texte et l'image de chaque page. Ce fichier peut donc faire l'objet de recherche de texte.

Le travail fut effectué sur un ordinateur HP Envy modèle 700-249 doté d'un processeur i7-4770 cadencé à 3,4 GHz, avec 12 Go de mémoire vive, sous Windows 10 édition Famille.

J'ai utilisé le logiciel ABBYY FineReader 12 Professional en version française. J'avais déjà l'expérience de la version 6 de ce logiciel qui avait été fournie avec mon scanner, un Epson Perfection V700 PHOTO. Le couvercle du scanner peut être enlevé, ce qui accélère le placement des pages sur le scanner. Mais la qualité de l'image produite peut nuire à la reconnaissance de caractères (voir plus bas). Sur plusieurs pages, une bande noire verticale dans la marge de droite et horizontale en bas de page est apparue. Au final, donc, mieux eût valu de ne pas retirer le couvercle du scanner !

J'ai choisi l'option de sauvegarde en format PDF/A et l'option de mise en page Copie exacte, ce qui permet de conserver une image exacte et non modifiable de la page saisie.

J'ai d'abord utilisé l'anglais comme langue, puis j'ai choisi la sélection automatique de langue, qui passe de l'anglais au français selon le texte, ce qui a grandement amélioré la reconnaissance des caractères français. Le logiciel permet une sélection automatique des langues désirées ; le nombre de langues disponibles est considérable : dans les centaines !

Dans les options, sous Document, j'ai choisi le Mode couleur noir et blanc, et sous l'onglet Lire, j'ai choisi Lecture approfondie (par défaut).

Les paramètres par défaut du scanner étaient de 300 ppp et la Luminosité à Manuel. J'ai essayé le réglage à 600 ppp sans que cela améliore la reconnaissance de caractères, sauf tel qu'indiqué plus bas.

Le logiciel peut produire un document FineReader, ce qui est en fait un dossier contenant l'image et le texte de chaque page, ainsi que les paramètres de saisie. Cela permet donc de revenir facilement sur la reconnaissance du texte, ce que j'ai dû faire pour le texte du début jusqu'au chapitre 2 inclusivement après avoir constaté des problèmes de reconnaissance des caractères qui m'avaient tout d'abord échappé. On peut sauver la version PDF du texte ou examiner une page numérisée dans un logiciel de visionnement de fichiers PDF, ce qui permet facilement de vérifier si les corrections apportées au texte numérisé sont reconnues par le lecteur de fichier PDF.

J'ai trouvé le logiciel FineReader rapide et facile à utiliser, sans doute parce que j'avais déjà un peu d'expérience dans la numérisation de documents. **Cependant, il est vite apparu qu'il fallait relire attentivement tout le texte.** D'abord, parce que le logiciel repère et surligne en vert (par défaut) les caractères qu'il n'est pas sûr d'avoir reconnu (par exemple, des « é » lus comme « e »), mais dont la reconnaissance est souvent correcte (par exemple, la ponctuation, comme le point suivi de la virgule dans « Ibid., »). Ensuite, parce que certains caractères sont mal lus (ou carrément sautés) sans que le logiciel s'en aperçoive. Les cas les plus fréquents furent les suivants :

- Appels de note sautés, mal lus, ou contenant un espace entre les chiffes
- Caractères accentués pas toujours reconnus
- Texte souligné pas toujours rendu comme souligné

Afin de réduire les erreurs de reconnaissance de texte, j'ai choisi, dans les Options, sous l'onglet Document, « Machine à écrire » comme type de document, plutôt que l'option par défaut, « Auto ». Cela a semblé améliorer la reconnaissance des caractères.

J'ai aussi expérimenté le mode d'apprentissage dans l'option Lire. Cela est déconseillé par les auteurs du logiciel parce que cela ralentit considérablement la reconnaissance de texte. J'ai utilisé l'option « Utiliser les gabarits intégrés [des polices de caractères] et les gabarits utilisateur ». Lorsque le logiciel a un doute sur un caractère, il fournit une image cadrée du caractère à reconnaître. Il faut parfois modifier le cadrage et taper le caractère à reconnaître dans la case appropriée, puis cliquer sur « Apprendre ». Après deux pages de lecture de tels caractères, j'ai désactivé l'apprentissage. Mais l'apprentissage a sans doute fonctionné, car le nombre de caractères marqués comme incertains par le logiciel a considérablement diminué.

FineReader soumet le texte numérisé à une vérification par dictionnaire. Cela m'a permis de repérer un certain nombre de coquilles qui étaient passées inaperçues autant par mon comité de thèse que par moi-même. Une leçon d'humilité ! Les coquilles repérées ont été laissées telles quelles dans la version PDF, afin de

conserver le caractère de copie exacte par rapport au document original. Voir *Coquilles conservées dans la version PDF*.

Le texte original contient quelques mots anglais abréviés, comme « Merch^{ts}. » (p. 250) où le point est sous la lettre finale de l'abréviation qui est en exposant. Pour la reconnaissance de mot en PDF, il faut tout rapporter sur la même ligne, l'image de la page demeurant une copie exacte de la page numérisée.

Le caractère « £ » est représenté dans le texte original par un « L » avec un « - » superposé, car la machine à écrire utilisée pour taper la thèse n'avait le symbole de la livre sterling dans son jeu de caractère. Le texte de la version numérisée remplace le « L » avec un « - » superposé par le « £ ».

Je n'ai eu qu'une seule difficulté avec le logiciel. Les annexes de la thèse contiennent des pages où le texte est disposé en format paysage plutôt qu'en format portrait. On peut faire pivoter l'image pour la mettre dans sa disposition originale, mais on perd alors la reconnaissance des caractères. J'ai donc laissé le logiciel faire pivoter l'image de la page avant de faire la reconnaissance des caractères, qui sont alors reconnus correctement.

Le graphique de la p. 309 a présenté une difficulté particulière. Tout d'abord, il est disposé en format paysage. Deuxièmement, la reconnaissance de caractères a très mal fonctionné. La solution a été de désactiver la reconnaissance de caractères et de sélectionner le graphique comme zone d'image. La p. 309 a alors été parfaitement rendue.

Les tableaux 1-B et 2-A, 3-A, et 3-B p. 320-324, ont été numérisés à 600 ppp à cause de la petite taille des caractères. L'alignement des entêtes de colonnes a dû être ajusté à la main. Notons que ces tableaux ont été tapés originellement dans une police différente du reste du document, et que quelques accents ont été ajoutés à la main sur le texte original parce que cette police ne contenait pas de caractères accentués.

Sur quelques pages, des petites bandes noires sont visibles sur l'image PDF. J'en ai supprimé quelques-unes au moyen de l'outil Modifier l'image. Mais une fois l'image modifiée, il faut refaire la reconnaissance des caractères et des zones d'image pour les graphiques. J'ai donc préféré laisser cette bande sur les pages où elle apparaît, car elle ne nuit ni à la lecture de l'image ni à la reconnaissance des caractères.

Dans la bibliographie, la convention typographique pour indiquer que l'auteur est le même que pour le titre précédent, était le « _____ . » Dans quelques cas, la reconnaissance des caractères ne l'a pas bien reproduit, et je l'ai rajoutée à la main, avec quelquefois un décalage de la marge gauche. Cela n'a aucune incidence sur la recherche de mots dans le document PDF.

Je n'ai pas encore essayé l'option de sauvegardé le texte numérisé en format Word. Je pressens que les appels de note de bas de page et les numéros des notes de bas de page ne seront pas liés. Mais comme mon objectif premier était de constituer une copie PDF du texte, ma démarche est pour l'instant satisfaisante.