The Canadian Peoples / Les populations canadiennes

Tableaux numérisés à partir des volumes publiés du recensement canadien

Version 2.0
Juin 2023

Geoff Cunfer, Laurent Richard and Marc St-Hilaire

Aperçu

L'un des livrables du projet The Canadian Peoples / Les populations canadiennes (TCP/LPC) est la création de fichiers numériques à partir de tableaux de données agrégées diffusées dans les volumes publiés. Assemblés à partir des données individuelles colligées dans l'ensemble du pays, le Bureau du recensement a publié des valeurs agrégées pour chaque Provinces, Territoires, Divisions (DR) et Subdivisions (SDR) au Canada. Disponibles seulement en version papier ou en version électronique (PDF) jusqu'à tout récemment, le projet TCP/LPC a sélectionné des tableaux et les a transformés pour une utilisation dans des logiciels de tableurs, de base de données, d'analyse statistique ou dans des Systèmes d'Information Géographique (SIG).

Afin de créer des cartes géographiques pour chacun des recensements à l'échelle des subdivisions (SDR), il a été nécessaire d'identifier une liste complète de ces entités pour l'ensemble du pays. Pour ce faire, un tableau publié par recensement a été numérisé en format MsExcel et a été retenu par le projet TCP/LPC comme le "Tableau Clef" établissant la liste définitive de SDR pour une année spécifique de recensement. Toutes les SDR mentionnées dans ce Tableau Clef ont un polygone correspondant dans les fichiers cartographiques créés par le projet.

Chaque SDR du Tableau Clef a été associée à un identifiant unique (TCPUID) qui regroupe 2 lettres identifiant la Province, 3 chiffres pour la Division (DR) et 3 autres chiffres pour la Subdivision (SDR). Par exemple, en 1891, le code ON112005 correspond à la subdivision du Village d'Hawkesbury, en fonction de la hiérarchie suivante : Ontario (ON), Division de recensement de Prescott (112) et cinquième Subdivision (005). Cet identifiant unique apparaît dans le fichier numérique du Tableau Clef et dans les couches cartographiques du SIG. Ce code peut donc être utilisé comme un champ d'appariement entre les polygones et les tableaux publiés convertis en feuilles de calcul. Veuillez noter que les TCPUID ne sont pas nécessairement les mêmes selon les années de recensement.

Ensuite, le projet TCP/LPC a créé des versions numériques interprétables de plusieurs autres tableaux publiés des divers recensements en procédant de la même manière et en créant des identifiants uniques semblables. Cependant, il est important de noter que la liste des SDR incluse dans ces autres tableaux publiés pourrait ne pas être identique à celle du Tableau Clef et des polygones créés, même si ces tableaux sont tirés d'une même année de recensement. Dans certains cas, le Bureau du recensement a établi des listes de subdivisions vraiment différentes

d'un tableau à l'autre. Un exemple extrême apparaît au recensement de 1911 où le Tableau 1 du Volume 1 présente des données pour 10 004 SDR alors que le Tableau 2 du Volume 2 présente des données pour seulement 3 532 SDR. Conséquemment, le Tableau Clef doit s'apparier sans heurts avec les polygones alors que ce n'est pas nécessairement le cas des autres tableaux publiés numérisés. Les usagers sont invités à vérifier eux-mêmes les divergences pouvant survenir lors d'appariements.

Méthodes

Afin de créer des tableaux chiffrés interprétables, l'équipe a téléchargé des images numériques (PDF) des tableaux publiés à partir de sites Web (https://archive.org/, https://publications.gc.ca) et a utilisé le logiciel de reconnaissance optique de caractères ABBYY Fine Reader 14. L'information transcrite a été révisée attentivement et des formules ont été utilisées pour valider les données. Pour chacun des champs attributaires, une comparaison a été effectuée entre la somme des SDR et les valeurs affichées par DR et par Provinces ainsi qu'entre le total et les variables représentant une partie de la totalisation. Par exemple, la population totale de l'Ontario telle que publiée dans le tableau original doit correspondre à l'addition de toutes les subdivisions de l'Ontario. De même, la population totale par SDR doit correspondre à l'addition des effectifs d'hommes et de femmes. En procédant ainsi, TCP/LPC a vérifié et corrigé les erreurs du processus de reconnaissance optique de caractères, et, dans certains cas, a signalé des erreurs probables contenues dans la source primaire. Lorsque des problèmes de cohérence des données ont été identifiés dans les tableaux originaux du recensement, l'équipe a tenté de les corriger autant que possible.

Les tableaux assemblés par le Bureau du recensement ont été publiés dans un format adapté aux volumes imprimés. Les tableaux originaux s'étalent parfois sur des pages en vis-à-vis, qui s'étendent sur plus de deux pages, incluant de longues entêtes de colonnes et présentant des données concernant diverses échelles géographiques. Une telle structure n'est pas nécessairement adaptée à celle de données préparées pour la lecture et l'analyse informatique. Voilà pourquoi TCP/LPC n'a pas seulement numérisé et corrigé les tableaux mais les a aussi reconfigurés dans diverses structures conçues pour être utilisées avec une variété de logiciels informatiques. Chacun des tableaux présentés dans cette collection est disponible en quatre formats distincts : "OCR", "CD", "CSD", et "Pub Tab", tel que décrits ci-dessous.

Format OCR

ex. 1891_V1T2_OCR_202306.xlsx

La version OCR est une feuille de calcul représentant les résultats du processus de reconnaissance optique de caractères effectué à l'aide du logiciel ayant converti les valeurs affichées dans les images des tableaux publiés en valeurs numériques. Cette version constitue donc une copie de la source primaire en format numérique interprétable. Elle inclut le texte complet des entêtes et les noms des lieux comme ils apparaissent dans le volume publié. Les erreurs présentent dans le document original sont dupliquées dans cette version OCR (en supplément des erreurs introduites par le logiciel de reconnaissance optique de caractères et non détectées initialement), mais, les données sont désormais dans un format numérique interprétable.

Formats CD et CSD ex. 1891_V1T2_CD_202306.xlsx
1891_V1T2_CSD_202306.xlsx

Ces deux versions séparent les Divisions (DR-CD) et les Subdivisions (SDR-CSD), des échelles géographiques entremêlées dans les volumes publiés originaux. Un chiffrier inclus seulement les DR et un autre seulement les SDR. Ces données ont été corrigées et harmonisées; elles sont quelque peu différentes de la source primaire originale. Les erreurs statistiques contenues dans les tableaux publiés ont été corrigées lorsque possible. Les longs entêtes de colonnes ont été remplacés par des variables aux noms plus courts et des champs standardisés d'identification ont été ajoutés.

Pour chacune des années de recensement considérées, un fichier particulier de documentation est disponible à propos des noms de variables utilisées dans les formats CD et CSD. La disponibilité des variables à travers l'ensemble des années de recensement du projet est présentée dans un « fichier maître » nommé TCP_CANADA_CD-CSD_Mastvar.xlsx.

Format Pub Tab ex. 1891_V1T2_PUB_202306.xlsx

Ce format de chiffrier calque la présentation des données publiées, incluant la mixité des échelles géographiques (Provinces, Divisions, Subdivisions). Toutefois, les noms des variables ont été raccourcis et les noms des entités géographiques ont été harmonisées de manière systématique (points cardinaux, nom des quartiers associés aux cités appropriées, etc.).

Ce format contient aussi des données originales corrigées autant que possible. Le champs « Notes » explique ces corrections. Si le champs Notes commence par le terme « TCP », cela indique généralement qu'un changement a été apporté aux données originales. Si le contenu du champs « Notes » ne débute pas par le terme « TCP », cela signifie que le commentaire est une transcription d'une information provenant du Bureau de recensement se trouvant dans le volume imprimé original. Le format Pub Tab correspond à celui créé par le projet *Canadian Century Research Infrastructure* / Infrastructure de recherche sur le Canada au 20e siècle (CCRI/IRCS) (https://ccri.library.ualberta.ca/).

Information importante concernant les tableaux numérisés TCP/LPC

Le processus de correction des erreurs et l'attribution de noms de variables ont été effectués indépendamment pour le format Pub Tab et les formats CD-CSD. Cela signifie que ces deux ensembles de formats utilisent des noms de variables différents et peuvent, dans des cas rarissimes, présenter des décisions différentes quant aux corrections d'erreurs.

Chacun des tableaux publiés numérisés est disponible en quatre formats à l'exception de ceux de 1911 et 1921 qui existent seulement en versions Pub Tab.

L'équipe en charge du volet géohistorique du projet TCP/LPC rend disponible un tableau numérisé concernant l'Île-du-Prince-Édouard (ÎPE) en 1871, bien que l'ÎPE a rejoint le *Dominion of Canada* seulement en 1873. Les usagers sont priés de noter que les variables disponibles pour l'ÎPE sont passablement différentes de celles transcrites pour le Canada. Même si ces deux tableaux réfèrent au "Volume 1, Table 1," ils sont publiés dans deux publications différentes et qui n'ont quasi rien en commun.

Les données sur les cultures et le bétail diffusées dans le chiffrier 1851_V2T6 contiennent plusieurs défis, principalement au sujet des unités de mesure des terres et des

récoltes. Il s'agit d'un problème bien connu (R.M. McInnis, "*Some Pitfalls in the 1851-1852 Census of Agriculture of Lower Canada*", *Histoire sociale/Social history*, Vol. XIV, No 27, pp. 219-231, mai 1981). Ces valeurs ont été transcrites sans modification; aucune tentative n'a été effectuée afin de solutionner ces problèmes d'unités de mesure.

Parfois, la numérotation des tableaux a posé des défis, particulièrement pour les recensements de 1851 et 1861. Pour ces années-là, les tableaux portant sur l'Ontario et le Québec ont été numérotés séparément, même s'ils contiennent des variables identiques. Occasionnellement, le Bureau de recensement utilise le terme « appendice » au lieu de « tableau ». Nous avons tenté de régulariser les appellations pour l'ensemble des années de recensement. Nous avons aussi attribué des codes uniformes pour les Provinces (ON, QC) en 1851 et 1861, même si les colonies étaient nommées « Haut-Canada / Canada Ouest » et « Bas-Canada / Canada Est » dans ces recensements pré-confédération.

Tableaux numérisés disponibles

Rec.	Volume	Tableau	Sujets principaux	Clef	Pub Tab	OCR, CD, CSD
1921	1	16	Lieu de naissance	O	X	
1921	1	27	Origine ethnique		X	
1921	1	38	Religion		X	
1921	3	3	Logement		X	
1911	1	1	Population	O	X	
1911	1	2	État matrimonial		X	
1911	2	2	Religion	O	X	
1911	2	7	Origine ethnique		X	
1911	2	28	Alphabétisation		X	
1901	1	7	Population	O	X	X
1891	1	2	Population et logement		X	X
1891	1	3	État matrimonial	O	X	X
1891	2	16	Occupation des terres		X	X
1891	4	2	Récolte		X	X
1891	4	3	Bétail		X	X
1881	1	1	Population	O	X	X
1881	3	24	Récolte		X	X
1881	3	27	Pêcheries		X	X
1871	1	1	Population	O	X	X
1871	1	1 (ÎPE)	Divers	O	X	X
1871	3	23	Récolte		X	X
1861	1	1 et 2	Origine ethnique	O	X	X
1861	1	5 et 6	Âge		X	X
1861	2	11 et 12	Récolte et bétail		X	X
1851	1	1	Origine ethnique		X	X
1851	1	3	Âge	O	X	X
1851	2	6	Récolte et bétail		X	X
1851	2	7	Moulins et manufactures		X	X

Globalement, les 28 tableaux numériques contiennent plus de 1 000 variables et plus de 75 000 rangées de données constituant une matrice d'environ 1 873 200 cellules.