

Exploration de l'utilisation des données ouvertes

La Base de données ouvertes d'adresses (BDOA)

Document de métadonnées : concepts, méthodologie et qualité des données

Version 1.0



Laboratoire d'exploration et d'intégration des données (LEID)
Centre des projets spéciaux sur les entreprises (CPSE)

Date de diffusion : 29 avril 2021



Statistics Canada
Statistique Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, consultez notre site Web au www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel au STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- | | |
|---|----------------|
| • Service de renseignements statistiques | 1-800-263-1136 |
| • Service national d'appareils de télécommunications pour les malentendants | 1-800-363-7629 |
| • Télécopieur | 1-514-283-9350 |

Programme des services de dépôt

- | | |
|-----------------------------|----------------|
| • Service de renseignements | 1-800-635-7943 |
| • Télécopieur | 1-800-565-7757 |

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Ces normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « [Normes de service à la clientèle](#) ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie, 2018

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'[entente de licence ouverte](#) de Statistique Canada.

This publication is also available in English.

Table des matières

1. APERÇU	3
2. SOURCES DE DONNÉES	3
3. PÉRIODE DE RÉFÉRENCE	4
4. POPULATION CIBLE.....	4
5. MÉTHODOLOGIE DE COMPILATION	4
6. DICTIONNAIRE DE DONNÉES	5
7. EXACTITUDE DES DONNÉES	8
8. REPRÉSENTATION GÉOGRAPHIQUE	9

Remerciements

Ce projet a pu profiter d'une collaboration avec OpenAddresses, surtout sur le code pour la compilation et le traitement des adresses. Nous leur sommes très reconnaissants pour le travail accompli et les conseils essentiels qu'ils nous ont donnés.

1. Aperçu

En vue d'explorer l'utilisation des données ouvertes pour produire les statistiques officielles et de soutenir la recherche géospatiale dans divers domaines, le Laboratoire d'exploration et d'intégration des données (LEID) a entrepris un projet en vue de créer une base de données d'adresses, harmonisée et fondée sur les données ouvertes ayant été publiées par plusieurs ordres de gouvernement au Canada¹. Le présent document décrit en détail le processus de collecte, de compilation et d'uniformisation des divers ensembles de données d'adresses ayant servi à la création de la *Base de données ouvertes d'adresses* (BDOA), accessible en vertu de la *Licence du gouvernement ouvert – Canada*².

Statistique Canada reconnaît la contribution des nombreuses administrations locales qui produisent des listes d'adresses publiques, qui sont la source de la Base de données ouvertes d'adresses (BDOA). Ces adresses seront également intégrées dans un nouveau Registre national des adresses (RNA) d'adresses résidentielles et non résidentielles, que Statistique Canada rendra accessible plus tard cette année. Compilé à partir d'une multitude de sources, le RNA sera une source exhaustive et normalisée d'adresses et de codes géographiques connexes accessibles au public. Il fait partie de la Stratégie de données pour la fonction publique fédérale.

Dans sa version actuelle (version 1.0), la BDOA contient plus que 10 millions d'enregistrements individuels. On prévoit mettre à jour périodiquement la base de données à mesure que de nouveaux ensembles de données ouvertes seront rendus disponibles, jusqu'à l'intégration complète dans un registre national d'adresses. La BDOA est fournie sous forme de fichier CSV (champs séparés par des virgules) compressé à l'échelle provinciale ou territoriale.

De plus, les codes de compilation et de traitement utilisés pour générer la BDOA sont consultables sur <https://github.com/CSBP-CPSE/Canadian-Open-Address-Point-Processing>. Cela permet d'effectuer des mises à jour automatiques des données et, de cette manière, d'actualiser en temps réel une base de données d'adresses municipales exhaustive, au fur et à mesure que les municipalités et les administrations locales mettent à jour les fichiers de données ouvertes.

Cet ensemble de données figure parmi les divers ensembles de données créés dans le cadre de l'Environnement de couplage de données ouvertes (ECDO). L'ECDO est une initiative exploratoire qui vise à accroître l'utilisation et l'harmonisation des données ouvertes provenant de sources faisant autorité en fournissant une série d'ensembles de données diffusés en vertu d'une licence unique, ainsi que du code source libre pour relier ces ensembles de données. On peut accéder aux ensembles de données et au code de l'ECDO sur le site Web de Statistique Canada à l'adresse suivante :

<https://www.statcan.gc.ca/fra/ecdo>

2. Sources de données

Partout au Canada, les administrations locales créent et tiennent à jour des adresses municipales. La BDOA dérive son enregistrement directement de ces sources sûres, qui ont rendu ces enregistrements publics en vertu d'une licence pour l'utilisation des données ouvertes qui est compatible avec la *Licence du gouvernement ouvert – Canada*. Ainsi, de nombreuses sources de données ont servi à créer la BDOA. La compilation a prolongé les travaux amorcés par l'organisation OpenAddresses, qui présente des agrégats de données d'adresses ouvertes de partout dans le monde sur sa page GitHub³. En tout, les données d'adresses provenant de 99 fournisseurs de données ont été utilisées (malgré un chevauchement géographique de certaines sources).

Les fournisseurs de données, qui comprennent divers ordres de gouvernement, sont indiqués dans le Tableau supplémentaire 1, accompagnés d'hyperliens vers les sources de données originales. Les sources de

¹ Cela comprend les niveaux municipal, régional et provincial.

² Voir : <https://ouvert.canada.ca/fr/licence-du-gouvernement-ouvert-canada>.

³ Voir : <https://www.github.com/openaddresses/openaddresses>

données sont attribuées à leur fournisseur respectif, conformément aux exigences de la licence. S'il y a lieu, la version de la licence est également indiquée. Pour en savoir plus sur les licences individuelles, les utilisateurs peuvent consulter directement les portails de données ouvertes des fournisseurs de données en question.

3. Période de référence

Idéalement, la période de référence aurait été la période à laquelle fait référence les données d'adresses. Malheureusement, ces renseignements n'étaient pas toujours disponibles dans les portails de données ouvertes. La fréquence d'actualisation des bases de données originales varie, de même que d'une source à l'autre, certaines déclarant des mises à jour hebdomadaires et d'autres, des mises à jour semestrielles, annuelles ou irrégulières. Dans le Tableau supplémentaire 1, on utilise donc plutôt la date du téléchargement de chaque ensemble de données municipal ayant servi à la création de la BDOA. Les données ont été recueillies dans les portails de données ouvertes entre janvier et avril 2021. Il est important de rappeler aux utilisateurs que la date du téléchargement ne doit pas être interprétée comme étant la période de référence des données. Si un utilisateur a besoin de renseignements précis sur la période de référence des données, il doit communiquer avec le fournisseur de données approprié, indiqué dans le Tableau supplémentaire 1.

4. Population cible

La BDOA vise à constituer un répertoire exhaustif et harmonisé d'adresses municipales qui sont disponibles grâce à des sources des données ouvertes d'administrations locales de tout le Canada. Les adresses peuvent répertorier des immeubles résidentiels, des édifices commerciaux ou des établissements institutionnels, ou simplement des terrains. De plus, les adresses attribuées aux bâtiments et aux terrains pourraient être multiples. La BDOA comprend toutes les adresses municipales sans double compte qu'il aura été possible de compiler à partir de sources de données ouvertes d'administrations publiques locales et provinciales qui sont mentionnées dans le Tableau supplémentaire 1.

5. Méthodologie de compilation

La méthodologie de compilation de la BDOA est presque entièrement automatisée afin de permettre une mise à jour potentiellement fréquente de la base de données. À mesure que les administrations locales seront plus nombreuses à augmenter la fréquence de mise à jour de leurs bases de données ouvertes d'adresses municipales, la BDOA deviendra complète en temps presque réel⁴.

Le code ayant servi à la collecte et au prétraitement s'inspire d'une version modifiée du pipeline de traitement mis au point par OpenAddresses. Ce processus télécharge les fichiers de données individuels et les transforme dans le même ensemble normalisé de colonnes faisant appel à un dictionnaire de mise en correspondance décrit dans les fichiers entrants JSON; il comprend un traitement mineur, au besoin, comme la séparation d'adresses en numéros de voirie distincts et des champs de noms de rue, ou encore la combinaison de champs provenant de données originales, au besoin. Pour chaque source, ce processus produit un fichier CSV de données d'adresses normalisées.

Les utilisateurs doivent noter qu'à l'intérieur des 99 ensembles de données obtenus, chaque fournisseur de données a joint un ensemble de variables différent aux données d'adresses. Dans certains cas, les divers champs qui composent l'adresse (numéro de rue, numéro de voirie, etc.) ont été fournis sous forme déjà séparée, alors que dans d'autres, ils ont dû être analysés à partir de champs d'adresse plus complets. De même, certains cas présentaient des genres de rue et leur direction normalisés en abréviations courantes, alors que d'autres fournissaient une forme entièrement développée. Enfin, les fournisseurs remettaient aussi des données dans une variété de formats de fichiers, allant de simples fichiers à valeurs séparées par des virgules (CSV) à des formats de fichiers géographiques,

⁴ Le code de compilation est accessible à l'adresse suivante : <https://github.com/CSBP-CPSE/Canadian-Open-Address-Point-Processing>. Ce code est basé sur une version modifiée du pipeline de traitement développé par OpenAddresses.

comme shapefiles ou geojson, ou encore des données accessibles par programmation, à l'aide d'une interface de programmation.

Les codes de compilation tiennent compte de ces différences et harmonisent les sources en un format normalisé. Ainsi, l'adoption ou la modification des normes de mise en forme des sources nécessiterait d'autres ajustements dans les codes de traitement.

Un autre traitement a été appliqué, en quatre étapes :

1. *Normalisation* : les adresses municipales ont été analysées et normalisées en champs de nom de rue, de genre de rue et de direction de la rue (p. ex. « RUE PRINCIPALE NORD » en « PRINCIPALE », « RUE », « N »). Une version modifiée de l'outil RASK (Road Attribute Search Key) dont Statistique Canada se sert pour normaliser les adresses tirées de sources administratives pour le couplage d'enregistrements a été utilisée dans ce processus. Toute source n'ayant pas de colonne d'adresse complète avait enregistré ce renseignement en liant le bureau, le numéro de rue et le nom complet de l'adresse. Dans le cas des sources qui ne comportaient pas de noms de ville, ce nom a été imputé du fichier source (p. ex. pour transformer « city_of_banff.csv » en « BANFF »). Les colonnes traitées et attribuées dans la base de données sont celles portant le suffixe « _pcs ».

2. *Épuration* : les enregistrements n'ayant pas de coordonnées ou de noms de rue ont été abandonnés. Toutes les coordonnées ont été tronquées à des lieux à 5 décimales (correspondant à la précision au mètre près). Le dédoublement des fichiers, effectué au niveau des sources originales, consistait à abandonner les enregistrements ayant des coordonnées, des bureaux, un numéro de voirie et un nom de rue normalisé identiques.

3. *Jointure spatiale* : tous les enregistrements ont fait l'objet d'une jointure spatiale avec le fichier de découpage géographique de la subdivision de recensement (SDR) de 2016 de Statistique Canada afin de leur attribuer une SDRidu, un nom de SDR et un PRidu. Un petit nombre d'enregistrements qu'il était impossible de situer dans des SDR a été abandonné.

4. *Fusion définitive* : toutes les sources de données ont été regroupées pour former un seul ensemble de données d'adresses pancanadien. Les doublons ont été abandonnés de nouveau, selon les mêmes critères qu'à l'étape 2. Puisque parfois, dans les données originales, une même adresse municipale peut avoir plusieurs coordonnées représentatives, il aura fallu calculer un identificateur de groupe en mettant en commun les entrées ayant une même SDRidu, un même numéro de voirie et des éléments d'adresse traités, de sorte que les entrées ayant un même identificateur de groupe correspondent à la même adresse municipale et puissent être traitées par un utilisateur final, au besoin.

À l'étape 4, un identificateur unique est calculé et attribué à chaque enregistrement. Cet identificateur unique est le résultat d'un hachage utilisant l'algorithme Blake2b de la bibliothèque hashlib de Python, généré à partir de la liaison des coordonnées et des champs d'adresse traités (le numéro de voirie, le bureau, et le nom de rue normalisé).). Cela signifie que pour les besoins de la BDOA, un seul enregistrement est défini uniquement par ses coordonnées et son adresse municipale, et qu'il ne comporte aucun autre champ, comme le fournisseur ou la ville.

Dans certains cas, il a été nécessaire de télécharger et de prétraiter les données avant de les faire passer dans le pipeline de collecte initial (par exemple, pour tenir compte des fichiers étant formatés de telle manière que le pipeline ne puisse les lire, des problèmes d'encodage ou, dans le cas de Montréal, pour diviser les plages d'adresses en rangées individuelles). Les scripts de prétraitement et leur description sont disponibles sur la page GitHub du projet.

6. Dictionnaire de données

Le dictionnaire de données ci-dessous décrit les variables contenues dans la BDOA exploratoire.

Variable – Latitude	
Nom	latitude
Format	Flottant
Source	Fournie telle quelle dans les données originales.
Description	Latitude en fractions décimales de degrés de l'adresse, tronquée à 5 décimales près.

Variable – Longitude	
Nom	longitude
Format	Flottant
Source	Fournie telle quelle dans les données originales.
Description	Longitude en fractions décimales de degrés de l'adresse, tronquée à 5 décimales près.

Variable - ID source	
Nom	id_source
Format	Alphanumeric
Source	Fournie telle quelle dans les données originales.
Description	Objet ou identificateur de champ unique attribué aux enregistrements lors de leur consignation dans les sources des données originales.

Variable - ID BDOA	
Nom	id
Format	Alphanumeric
Source	Générée à l'interne lors du traitement des données
Description	Identificateur unique attribué aux dérivés d'un condensé numérique calculé à partir des champs de coordonnées et d'adresse normalisés.

Variable - ID group	
Nom	id_group
Format	Alphanumeric
Source	Générée à l'interne lors du traitement des données
Description	Identificateur de champ attribué aux enregistrements qui ont des renseignements communs quant à l'adresse (numéro de voirie, nom de rue, genre de rue, direction de la rue), mais des coordonnées géographiques différentes.

Variable – Numéro de rue	
Nom	numero_rue
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Numéro de voirie de l'adresse, qu'il soit fourni ou analysé à partir de l'adresse complète.

Variable – Nom complet de la rue	
Nom	rue
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Nom de rue de l'adresse, y compris le genre de rue et la direction de la rue, le cas échéant, qu'il soit fourni ou analysé à partir de l'adresse complète.

Variable – Nom de la rue	
Nom	nom_rue
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Nom de rue de l'adresse, sans genre ni direction, comme prévu.

Variable – Genre de rue	
Nom	type_rue
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Genre de rue de l'adresse, comme prévu

Variable – Direction de rue	
Nom	dir_rue
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Direction de la rue de l'adresse, comme prévu.

Variable – Unité	
Nom	unite
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Numéro du local, qu'il soit fourni ou analysé à partir de l'adresse complète.

Variable – Municipalité	
Nom	ville
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Nom de la municipalité.

Variable – Code postal	
Nom	code_postal
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Code postal de l'adresse

Variable – Adresse complète	
Nom	adr_complete
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales ou imputée
Description	Adresse complète, qu'elle soit fournie ou créée par liaison d'autres champs.

Variable – Municipalité traitée	
Nom	ville_pcs
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Nom de la municipalité, tiré du nom de fichier de la source originale, au besoin.

Variable – Nom de la rue traitée	
Nom	nom_rue_pcs
Format	Chaîne de caractères
Source	Générée à l'interne lors du traitement des données
Description	Nom de rue normalisé de l'adresse, sans genre ni direction.

Variable – Type de la rue traitée	
Nom	type_rue_pcs
Format	Chaîne de caractères
Source	Générée à l'interne lors du traitement des données
Description	Genre de rue normalisé de l'adresse.

Variable – Direction de la rue traitée	
Nom	dir_rue_pcs
Format	Chaîne de caractères
Source	Générée à l'interne lors du traitement des données
Description	Direction normalisée de la rue de l'adresse.

Variable – Identificateur unique de la subdivision de recensement	
Nom	sdridu
Format	Nombre entier
Source	Limites des subdivisions de recensement du Canada, 2016 (Produit GeoSuite de Statistique Canada)
Description	Identificateur unique de la subdivision de recensement.

Variable – Nom de la subdivision de recensement	
Nom	sdrnom
Format	Chaîne de caractères
Source	Limites de la subdivision canadienne de recensement, 2016 (Produit GeoSuite de Statistique Canada)
Description	Nom de la subdivision de recensement.

Variable – Identificateur unique de province	
Nom	pridu
Format	Nombre entier
Source	Limites de la subdivision canadienne de recensement, 2016 (Produit GeoSuite de Statistique Canada)
Description	Identificateur unique de la province.

Variable – Fournisseur de données	
Nom	fournisseur
Format	Texte (chaîne de caractères)
Source	Créée à partir des origines de l'ensemble de données ayant servi d'intrant.
Description	Nom de la municipalité, de la région ou de la province/territoire ayant fourni l'ensemble de données.

7. Exactitude des données

Toutes les adresses ont été collectées à partir de sources de données gouvernementales. En général, les ensembles de données obtenus ont été laissés tels quels, à l'exception d'un traitement d'uniformisation des sources afin de constituer une seule base de données.

Durant la phase du traitement des ensembles de données afin de créer la BDOA, plusieurs étapes ont été suivies pour accroître l'uniformité des données de sortie notamment la normalisation des genres de rue et le dédoublement des entrées. Il se pourrait que le processus utilisé pour normaliser les adresses ait inséré quelques erreurs, mais ces dernières devraient être infimes. De même, il est possible qu'il reste des entrées en double dans la base de données. La colonne de l'adresse complète est aussi fournie sans avoir subi de normalisation.

La BDOA expérimentale ne contient que des données ouvertes gouvernementales qui sont disponibles au moment de la compilation, et ne doit donc pas être interprétée comme un répertoire exhaustif et objectif de toutes les adresses du Canada. Présentement, la BDOA ne couvre pas tout le Canada. La base de données contient encore des espaces vides, car il y a des régions pour lesquelles on n'a pas pu trouver de données ouvertes gouvernementales sur les adresses. Certaines de ces lacunes pourraient être comblées à mesure que les administrations locales publient davantage d'adresses civiques sous forme de données ouvertes.

8. Représentation géographique

La base de données ouvertes d'adresses est consultable sur le site Web de Statistique Canada, avec les coordonnées présentées en latitudes et en longitudes obtenues à l'aide de l'ellipsoïde WGS84 standard.