

Exploration de l'utilisation des données ouvertes

La Base de données ouvertes sur les établissements d'enseignement (BDOEE)

Document de métadonnées : concepts, méthodologie et qualité des données

Version 2.1



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

Date de diffusion : 28 novembre 2022



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
GÉOCODAGE.....	5
IMPUTATION DES NIVEAUX DE LA CITE.....	5
IMPUTATION DES NOMS DE SUBDIVISION DE RECENSEMENT (SDR)	6
TYPE D'ÉTABLISSEMENT FOURNI DANS LES ENSEMBLES DE DONNÉES SOURCES	6
STANDARDISATION DES DONNÉES	7
<i>Analyse des adresses</i>	7
<i>Suppression des enregistrements en double</i>	7
6. DICTIONNAIRE DE DONNÉES	7
7. EXACTITUDE DES DONNÉES	11
8. CONTACTEZ-NOUS.....	11

Remerciements

Une première version de la base de données a été réalisée grâce au financement de Services aux Autochtones Canada (SAC) et de Relations Couronne-Autochtones et Affaires du Nord Canada (RCAANC). Cette version mise à jour, qui comprend les écoles de langue officielle en situation minoritaire, a été réalisée grâce au financement du Secrétariat du Conseil du Trésor du Canada (SCT) et en consultation avec Patrimoine canadien (PCH). Ces organisations nous ont fait part de leurs précieux commentaires, et nous les en remercions.

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 sur les établissements d'enseignement qui soit accessible, 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 sur les établissements d'enseignement ayant servi à la création d'une mise à jour de la deuxième version de la *Base de données ouvertes sur les établissements d'enseignement* (BDOEE), accessible en vertu de la *Licence du gouvernement ouvert – Canada*².

Dans sa version actuelle (version 2.1), la BDOEE contient 18 982 enregistrements individuels. Pour cette mise à jour de la base de données, des renseignements sur les écoles publiques des minorités de langues officielles (EMLO) ont été ajoutés à la version 2.0 existante de la BDOEE. Une EMLO s'entend d'une école anglophone au Québec ou d'une école francophone à l'extérieur du Québec. Au total, 967 enregistrements existants ont été désignés comme des enregistrements d'EMLO, et 38 nouveaux enregistrements ont été ajoutés à la version 2.1. Comme les données des EMLO ont été recueillies plus récemment que les données de la BDOEE, certains établissements dont l'adresse a changé l'ont fait mettre à jour. De plus, les coordonnées de latitude et de longitude des EMLO ont été mises à jour dans les enregistrements appariés de la BDOEE pour lesquels il manquait des données. On a ajouté des renseignements sur les RMR avec une jonction spatiale en utilisant le paquetage `sf`³ dans R pour tous les enregistrements comportant des données sur les coordonnées à des fins de concordance avec les EMLO. On prévoit mettre à jour périodiquement la base de données à mesure que de nouveaux ensembles de données ouvertes seront rendus disponibles. La BDOEE est fournie sous forme de fichier CSV (champs séparés par des virgules) compressé.

Cet ensemble de données figure parmi plusieurs ensembles de données créés dans le cadre de l'Environnement de couplage de données ouvertes (ECDO). L'ECDO est une initiative 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

De nombreuses sources de données ont été utilisées pour créer la BDOEE. Les fournisseurs de données, qui comprennent divers ordres de gouvernement, sont indiqués dans le matériel supplémentaire du tableau 1, y compris l'attribution à chaque source de données 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. En plus des bases de données faisant l'objet d'une licence ouverte, la BDOEE comprend également un ensemble de listes accessibles au public d'établissements d'enseignement dont l'inclusion a été autorisée par les fournisseurs de données.

En raison de l'inclusion de la variable EMLO dans la version 2.1 de la BDOEE, toutes les sources d'information sur les EMLO sont incluses dans le tableau 2 du matériel supplémentaire. Pour chaque province et territoire où de multiples sources de données sur le statut d'EMLO ont été trouvées, on a choisi une seule source de données primaire qui contenait le plus grand nombre d'enregistrements et d'attributs utiles comme les niveaux scolaires et l'information sur les adresses.

¹ Cela comprend les niveaux municipal, régional et provincial.

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

³ SF est un progiciel dans R pour la manipulation de données géospatiales : <https://r-spatial.github.io/sf/>.

En plus des sources primaires énumérées au tableau 2, on a effectué une validation en comparant les listes aux pages Web des conseils scolaires de langue officielle en situation minoritaire. Cette validation a permis d'ajouter un petit nombre d'établissements qui manquaient dans les sources de données d'origine. Les sources supplémentaires utilisées sont énumérées au tableau 3 du matériel supplémentaire.

3. Période de référence

Le matériel supplémentaire présente la fréquence de mise à jour ou la date à laquelle chaque ensemble de données a été mis à jour par le fournisseur (lorsque celle-ci est connue), ainsi que la date à laquelle chaque ensemble de données utilisé dans la BDOEE a été téléchargé. Les données ont été recueillies entre août 2019 et mars 2021 pour les données de la BDOEE, et de novembre 2021 à mars 2022 pour le statut d'EMLO. Il importe de rappeler aux utilisateurs que la date du téléchargement ne doit pas être interprétée comme étant la période de référence des données. Si l'utilisateur nécessite des renseignements précis sur la période de référence des données, il doit communiquer avec le fournisseur de données concerné.

4. Population cible

Un établissement d'enseignement est un lieu physique dont l'activité première consiste à donner un enseignement à un ensemble d'élèves ou de participants. Tous les établissements d'enseignement au Canada sont pris en compte dans cet ensemble de données. Cela inclut tous les niveaux d'éducation, les écoles privées et publiques sans exclusions quant au mode de financement, au type d'exploitant, au domaine, à la dénomination, au type d'élève, au lieu, etc.

Compte tenu de cette définition, la base de données couvre des établissements tels que les services d'éducation de la petite enfance, la maternelle, les établissements primaires, secondaires et postsecondaires, et des centres de formation professionnelle précis (comme les écoles de coiffure). La base de données n'inclut pas les établissements d'enseignement virtuels.

Pour le statut d'EMLO, la population cible est limitée aux écoles publiques de langue officielle en situation minoritaire de la maternelle à la 12^e année, ce qui peut comprendre à la fois des écoles traditionnelles et des écoles alternatives si elles sont contrôlées par des conseils ou des autorités scolaires de langue officielle en situation minoritaire.

Seule une modification minime des ensembles de données originaux a été réalisée. Au fur et à mesure que le travail sur la BDOEE expérimentale avancera, les définitions et les seuils évolueront. Il importe de rappeler aux utilisateurs que, dans la plupart des cas, il est possible d'obtenir directement les données non modifiées dans les portails de données ouvertes des divers fournisseurs de données.

5. Méthodologie de compilation

La première composante de traitement de la base de données comprenait le reformatage des données sources au format CSV et la mise en correspondance des attributs de l'ensemble de données original avec les noms des variables normalisées (colonnes). Un dictionnaire de données des variables utilisées est présenté à la section 6. Dictionnaire de données. Afin de compiler les données dans une seule base de données, les activités suivantes ont été effectuées :

- Les données d'adresse concaténées ont été analysées et séparées dans les composantes qui les correspondent (p. ex. unité, numéro et nom de la rue, nom de la ville, etc.) au moyen de libpostal⁴, une solution de traitement du langage naturel pour l'analyse des adresses.

⁴ Voir : <https://github.com/openvenues/libpostal>.

- Déduplication au moyen de la mise en correspondance floue et parfaite de chaînes de caractères. Cette étape est réalisée de manière prudente afin d'éviter les faux positifs (pour plus de détails, voir Standardisation des données).

Les fichiers et les champs de données originaux ont été convertis dans des formats et des champs normalisés à l'aide du logiciel personnalisé OpenTabulate⁵. Un nombre limité d'inscriptions ont été modifiées manuellement lorsqu'il était évident que l'analyse n'avait pas été réalisée correctement. Prenons l'exemple des adresses comportant des nombres avec un trait d'union comme « 1035-55 rue n° », qui peut avoir été interprété comme ayant le numéro « 1035-55 » et le nom de rue « rue no », plutôt que le numéro 1035 et le nom de rue « 55^e rue no ». Bien que des efforts aient été déployés pour assurer que les données soient correctes, il est possible que les scripts utilisés pour traiter et analyser les adresses aient entraîné par inadvertance d'autres erreurs non détectées. Si de telles erreurs sont détectées, elles seront corrigées dans les versions futures de la BDOEE.

En général, les données incluses dans la BDOEE sont les données accessibles dans les sources originales sans imputation. Le géocodage des entrées dont les coordonnées sont manquantes, et l'imputation des noms RSD et les niveaux de la CITE, décrite ci-après, fait exception à la règle.

Dans la version 2 de la BDOEE, l'identifiant unique est passé d'un nombre entier à un hachage calculé à partir du nom de l'établissement, de l'adresse et de l'identifiant de la source (si disponible) de l'enregistrement.

Géocodage

Les enregistrements qui ne comportaient pas de géocoordonnées provenant de la source ont été géocodés à l'aide du géocodeur ESRI ArcGIS Online (AGOL) et du géocodeur OpenStreetMap (Nominatim). Le géocodeur AGOL renvoie les coordonnées, ainsi qu'un score et un type de géocodage. Seuls les enregistrements dont le score est supérieur à 90 et dont le type d'adresse indique que les coordonnées sont soit une adresse, une sous-adresse, un point d'intérêt ou une intersection ont été retenus pour la base de données finale. Les enregistrements qui ne pouvaient pas être géocodés avec le niveau de précision décrit ci-dessus ont ensuite été transmis au géocodeur Nominatim. Les écoles ont été recherchées à l'aide du nom de l'école, de la ville et de la province, et ont été conservées si le nom de l'école obtenu correspondait de près au nom de l'école d'origine. La colonne Geo_Source indique si les coordonnées d'un enregistrement ont été fournies par la source originale ou si elles ont été géocodées.

Imputation des niveaux de la CITE

Les sources de données originales utilisent diverses normes, classifications et nomenclatures pour décrire le niveau d'éducation ou les années scolaires. La BDOEE utilise la Classification internationale type de l'éducation (CITE)⁶ pour fournir une définition normalisée du niveau d'éducation. Cela a requis la conversion des années scolaires ou du niveau d'éducation d'un établissement d'éducation à un niveau de CITE.

Les niveaux de la CITE ont été dérivés à partir des années scolaires indiquées dans les données du fournisseur, si des années sont accessibles. Autrement, le niveau d'éducation est converti en années scolaires, qui sont ensuite mises en correspondance avec les niveaux de la CITE. Les entrées dans les données d'origine qui ne contenaient pas d'informations sur le niveau d'éducation n'ont pas reçu d'attributs de CITE, alors, ces champs sont vides dans la BDOEE.

Le tableau 1 présente la mise en correspondance directe des niveaux de la CITE avec les années scolaires, et le tableau 2 présente les années scolaires comprises dans un niveau d'éducation par province et territoire. Il convient de souligner que la définition de la « maternelle » comme niveau d'éducation varie selon les sources de données, et que certaines de ces écoles offrent une éducation à la petite enfance. Pour éviter les faux positifs, des valeurs ne

⁵ Voir : <https://pypi.org/project/opentabulate/>.

⁶ Voir : <https://doi.org/10.1787/9789264228368-en>.

sont pas attribuées dans la colonne CITE010 pour les établissements qui indiquent accueillir des élèves du préscolaire, décrit comme un niveau d'éducation (et non une année scolaire). Par exemple, les services de garde d'enfants en Alberta comprennent la maternelle et peuvent également inclure des services pour les enfants plus jeunes, mais ils n'ont été mis en correspondance qu'avec la CITE020. Malgré le fait que certains de ces établissements offrent une éducation à la petite enfance, la notion du préscolaire semble varier entre les fournisseurs de données et les écoles. Le tableau 2 en témoigne, le « préscolaire » étant associé à la maternelle lorsqu'il est converti en une année scolaire.

Tableau 1 : Variables du dictionnaire de données et niveaux de la CITE correspondants

Variable	Nom	Niveau de la CITE	Années scolaires
Éducation de la petite enfance	CITE010	010	Préscolaire
Maternelle	CITE020	020	Maternelle
Primaire	CITE1	1	1 à 6
Secondaire de premier cycle	CITE2	2	7 à 9
Secondaire de deuxième cycle	CITE3	3	10 à 12
Postsecondaire	CITE4+	4+	-

Tableau 2 : Définition de la conversion du niveau d'éducation en années scolaires selon la province/territoire

Province / territoire	Préscolaire / maternelle	Primaire	Secondaire de premier cycle	Secondaire de deuxième cycle
T.-N.-L., Î.-P.-É., N.-É., Alb., T.N.-O., Nt	Maternelle	1 à 6	7 à 9	10 à 12
N.-B.	Maternelle	1 à 5	6 à 8	9 à 12
Qc	Maternelle	1 à 6	7 à 11	
Ont.	Maternelle	1 à 8	9 à 12	
Man.	Maternelle	1 à 4	5 à 8	9 à 12
Sask.	Maternelle	1 à 5	6 à 9	10 à 12
C.-B., Yn	Maternelle	1 à 7	8 à 12	

Imputation des noms de subdivision de recensement (SDR)

Les noms de subdivision de recensement (SDR)⁷ proviennent des coordonnées géographiques, à savoir la latitude et la longitude. Les coordonnées sont attribuées aux SDR correspondantes en liant les points de coordonnées aux polygones de la SDR au moyen d'une opération de jointure spatiale en utilisant le paquet GeoPandas⁸ de Python.

Type d'établissement fourni dans les ensembles de données sources

Le type d'établissement fourni (p. ex. public, privé, confessionnel, etc.) a été utilisé tel qu'il fût indiqué dans l'ensemble de données source sans tentative d'interprétation, de nouvelle attribution ou de mise en correspondance avec une classification uniforme. Par rapport à l'utilisation de la CITE pour normaliser les niveaux d'éducation, il n'existe aucune norme liée au type d'établissement. Lorsque la source de données n'avait pas de colonne de type mais que la source de données elle-même correspondait à un type particulier (par exemple, un fichier d'écoles publiques ou un fichier d'écoles privées), le type d'établissement a été défini manuellement.

⁷ « Subdivision de recensement » est un terme générique qui désigne les municipalités (telles que définies par les lois provinciales ou territoriales) ou les régions considérées comme étant des équivalents municipaux à des fins statistiques. On peut obtenir une définition détaillée à l'adresse suivante : <https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/geo012-fra.cfm>.

⁸ GeoPandas est un progiciel de Python pour la manipulation de données géospatiales : <http://geopandas.org/index.html>.

Standardisation des données

En raison des différentes normes adoptées dans les données originales, les mesures prises pour normaliser les données peuvent donner lieu à des erreurs. Les principes clés de la méthodologie utilisée sont d'éviter les faux positifs et les modifications importantes des données. La méthodologie et les limites de chaque technique sont décrites ci-dessous. Les techniques de nettoyage banales, comme la suppression des espaces et de la ponctuation, ne sont pas décrites.

Analyse des adresses

L'analyseur d'adresses libpostal, une solution libre de traitement du langage naturel permettant d'analyser les adresses, est utilisé pour séparer les chaînes d'adresse concaténées en chaînes correspondant aux variables d'adresse, comme le nom de rue et le numéro de rue. À l'occasion, les adresses ne seront pas séparées correctement en raison du formatage non conventionnel de l'adresse originale. Il est possible que des inscriptions ayant été analysées de façon erronée n'aient pas été détectées, malgré les efforts déployés pour les relever et les corriger dans la base de données finale. Les inscriptions dont le numéro d'immeuble est composé de deux nombres séparés par un trait d'union ou une espace font exceptions. Ces inscriptions indiquent habituellement que l'analyseur d'adresses a mal analysé une adresse, par exemple, dans l'inscription « 123 100 ave », « 123 100 » est considéré comme le numéro d'immeuble et « ave », comme le nom de rue ou alors une unité n'est pas identifiée correctement (comme dans l'entrée « 3-100 rue principale »). Ces nombres sont automatiquement séparés, et, si le nom de rue est une variante du mot « rue » ou « avenue », le nombre de droite est considéré comme le nom de rue.

Pour les inscriptions d'EMLO où seule une adresse de case postale a été fournie, les adresses ont été supprimées et remplacées par les adresses de voirie, qui ont été trouvées au moyen de recherches manuelles sur Internet.

Finalement, une quantité limitée d'inscriptions n'ayant pas été analysées correctement ont été relevées lors d'une vérification manuelle, puis corrigées.

Suppression des enregistrements en double

La suppression des doublons a été effectuée à l'aide du paquet Record Linkage Toolkit⁹ en Python, où les distances de Levenshtein et de Cosine ont été calculées sur les champs de nom et d'adresse pour les installations au sein de la même SDR. Les paires d'enregistrements dont la métrique de similarité des chaînes de caractères était supérieure à 0,9 ont été signalées pour inspection et supprimées s'il s'agissait de doublons.

Pour les inscriptions d'EMLO, on a inspecté manuellement les paires d'enregistrements pour déterminer si les appariements indiquaient de vrais ou de faux doublons. En effectuant des recherches sur Internet pour comparer les noms et les adresses entre les paires appariées et, dans certains cas, en vérifiant la réalité de terrain au moyen de sites cartographiques, on a établi que la plupart des paires d'enregistrements étaient de faux doublons. En outre, on a constaté que plusieurs paires appartenaient à la même école, mais couvraient des années scolaires différentes — elles ont été indiquées séparément. En fin de compte, seules les inscriptions qui semblaient être des doublons évidents (noms et adresses très semblables et renseignements égaux sur les années scolaires) ainsi que les établissements dont les noms et les adresses correspondaient parfaitement ont été choisis en vue d'être supprimés.

6. Dictionnaire de données

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

Variable – Numéro d'enregistrement	
Nom	Index
Format	Chaîne de caractères
Source	Générée à l'interne lors du traitement des données.
Description	Numéro d'enregistrement unique généré automatiquement lors du traitement des données.

⁹ Voir : <https://recordlinkage.readthedocs.io/en/latest/about.html>

Variable – Source ID	
Name	Source_ID
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales
Description	L'identifiant unique de l'enregistrement tel qu'il figure dans la source de données originale, si disponible.

Variable – Nom de l'établissement	
Nom	Nom_Établissement
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Nom de l'établissement.

Variable – Type d'établissement	
Nom	Type_Établissement
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Type d'établissement (p. ex. public, privé, gouvernemental, etc.).

Variable – Nom de l'autorité	
Nom	Nom_Autorité
Format	Chaîne de caractères
Source	Fournie telle quelle dans les données originales.
Description	Nom de l'autorité.

Variable – Éducation de la petite enfance	
Nom	CITE010
Format	Booléen
Source	Fournie telle quelle dans les données originales ou imputées à partir des données sur les années scolaires.
Description	Accueil des élèves de la petite enfance telle que la définit le niveau de la CITE au tableau 1.

Variable – Maternelle	
Nom	CITE020
Format	Booléen
Source	Fournie telle quelle dans les données originales ou imputées à partir des données sur les années scolaires.
Description	Accueil des élèves de maternelle telle que la définit le niveau de la CITE au tableau 1.

Variable – Primaire	
Nom	CITE1
Format	Booléen
Source	Fournie telle quelle dans les données originales ou imputées à partir des données sur les années scolaires.
Description	Accueil des élèves du primaire tel que le définit le niveau de la CITE au tableau 1.

Variable – Secondaire de premier cycle	
Nom	CITE2
Format	Booléen
Source	Fournie telle quelle dans les données originales ou imputées à partir des données sur les années scolaires.
Description	Accueil des élèves au premier cycle du secondaire tel que le définit le niveau de la CITE au tableau 1.

Variable – Secondaire de deuxième cycle	
Nom	CITE3
Format	Booléen
Source	Fournie telle quelle dans les données originales ou imputées à partir des données sur les années scolaires.
Description	Accueil des élèves au deuxième cycle du secondaire tel que le définit le niveau de la CITE au tableau 1.

Variable -- Postsecondaire	
Nom	CITE4Plus
Format	Booléen
Source	Fournie telle quelle dans les données originales ou imputées à partir des données sur les années scolaires.
Description	Accueil des élèves de niveau postsecondaire tel que le définit le niveau de la CITE au tableau 1.

Variable – Désignation d'école de minorité de langue officielle	
Nom	Statut_EMLO
Format	Booléen
Source	Enregistrements appariés avec une base de données des écoles publiques des minorités de langues officielles de la maternelle à la 12 ^e année.
Description	Une école de minorité de langue officielle est une école anglophone au Québec ou une école francophone dans les autres provinces et territoires. Une valeur de 1 indique que l'enregistrement est un enregistrement d'EMLO.

Variables de lieu

Variable – Adresse complète	
Nom	Adr_Complète
Format	Chaîne de caractères
Source	Une combinaison de composants d'adresses ou fournis tels quels.
Description	Adresse complète de l'établissement.

Variable – Unité	
Nom	Unité
Format	Chaîne de caractères
Source	Analysée à partir de la chaîne de l'adresse complète ou fournie telle quelle.
Description	Numéro du local.

Variable – Numéro de la rue	
Nom	Numéro_Rue
Format	Chaîne de caractères
Source	Analysée à partir de la chaîne de l'adresse complète ou fournie telle quelle.
Description	Numéro d'immeuble.

Variable – Nom de la rue	
Nom	Nom_Rue
Format	Chaîne de caractères
Source	Analysée à partir de la chaîne de l'adresse complète ou fournie telle quelle.
Description	Nom de la rue (type et direction).

Variable – Ville	
Nom	Ville
Format	Chaîne de caractères
Source	Analysée à partir de la chaîne de l'adresse complète ou fournie telle quelle.
Description	Nom de la municipalité.

Variable – Province/territoire	
Nom	Prov_Terr
Format	Chaîne de caractères
Source	Convertie en un code de deux lettres (approuvé à l'échelle internationale) après analyse à partir de la chaîne de l'adresse complète ou indiquée par le fournisseur.
Description	Nom de la province ou du territoire.
Variable – Code postale	
Nom	Code_Postale
Format	Chaîne de caractères
Source	Analysée à partir de la chaîne de l'adresse complète ou fournie telle quelle.
Description	Code postale.
Variable – Identificateur unique de province	
Nom	PRIDU
Format	Nombre entier
Source	Converti du code de province.
Description	Identificateur unique de la province.
Variable – Nom de SDR	
Nom	SDR_Nom
Format	Chaîne de caractères
Source	Imputée à partir des coordonnées géographiques et des noms de ville au moyen de Geosuite 2016.
Description	Nom de la subdivision de recensement.
Variable – Identificateur unique de la SDR	
Nom	SDRIDU
Format	Chaîne de caractères
Source	Imputée à partir des coordonnées géographiques ou du nom de la SDR au moyen de GeoSuite 2016.
Description	Identificateur unique de la subdivision de recensement.
Variable – Longitude	
Nom	Longitude
Format	Flottant
Source	Fournie telle quelle dans les données originales.
Description	Longitude.
Variable – Latitude	
Nom	Latitude
Format	Flottant
Source	Fournie telle quelle dans les données originales.
Description	Latitude.
Variable – Source géocoordonnées	
Name	Geo_Source
Format	Chaîne de caractères
Source	Créé sur la base des origines des géocoordonnées.
Description	Une indication pour savoir si la latitude et la longitude ont été fournies dans la source originale, ou si elles ont été géocodées pour la BDOEE.
Variable – Fournisseur de données	
Nom	Fournisseur
Format	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.

Variable – Nom de la RMR	
Nom	RMR_Nom
Format	Chaîne de caractères
Source	Imputée à partir des fichiers des limites du Recensement de 2021 d'après l'emplacement spatial.
Description	Nom de la région métropolitaine de recensement.

Variable – Identificateur unique de la RMR	
Nom	RMRIDU
Format	Chaîne de caractères
Source	Imputée à partir des fichiers des limites du Recensement de 2021 d'après l'emplacement spatial.
Description	Identificateur unique de la région métropolitaine de recensement.

7. Exactitude des données

Toutes les données relatives aux établissements d'enseignement figurant dans la BDOEE ont été collectées à partir de sources de données gouvernementales, soit à partir de portails de données ouverts, soit de pages web publiques. 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.

Quelques exceptions s'appliquent aux inscriptions d'EMLO. Certaines inscriptions qui ne figuraient pas dans les sources de données d'origine ont été ajoutées après avoir été comparées aux pages Web des conseils scolaires de langue officielle en situation minoritaire. Lorsqu'il manquait des renseignements sur les écoles, comme l'adresse ou le conseil scolaire, les données ont été complétées au moyen de recherches manuelles.

L'imputation des niveaux de la CITE est réalisée de manière prudente afin d'éviter les faux positifs. En conséquence, les pourcentages des niveaux de la CITE ayant des valeurs non vides diffèrent selon le niveau. Des méthodes de traitement du langage naturel sont utilisées pour effectuer l'analyse et la séparation des chaînes de caractères d'adresse en variables d'adresse, comme le numéro et le code postal. Les méthodes sont reconnues pour offrir un grand rendement et une grande exactitude, mais, comme pour toutes les méthodes d'apprentissage statistique, elles présentent également des limites. Un mauvais formatage ou un formatage non conventionnel des adresses peut entraîner une analyse erronée. À cette étape, il n'y a eu aucune autre tentative d'intégration à d'autres sources d'adresses; ainsi, bien qu'on s'attende généralement à ce que les enregistrements d'adresse soient corrects, des erreurs résiduelles peuvent être présentes dans la version actuelle de la base de données.

Enfin, il convient de souligner que le type d'établissement, qui distingue les établissements publics, privés et d'autres types d'établissements, a des interprétations différentes selon la province et le fournisseur de données. À titre d'exemple, les écoles religieuses peuvent être financées par l'État dans une juridiction, mais pas dans une autre.

8. Contactez-nous

Les projets de Statistique Canada sur les données ouvertes sont conçus pour être améliorés de façon continue. Pour fournir des informations sur les ajouts, les mises à jour, les corrections ou les omissions, ou pour plus d'informations, veuillez nous contacter à l'adresse suivante : statcan.lode-ecdo.statcan@statcan.gc.ca. Veuillez inclure le titre de la base de données ouvertes dans le sujet du courriel.