

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Couplage efficace d'enregistrements pour les grands ensembles de données, selon les noms d'entreprise

par Hanan Ather, Serge Godbout et Dave MacNeil

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Couplage efficace d'enregistrements pour les grands ensembles de données, selon les noms d'entreprise

Hanan Ather, Serge Godbout et Dave MacNeil¹

Résumé

Le couplage d'enregistrements exact et efficace s'avère crucial pour veiller à ce que le Registre statistique des entreprises (RSE) de Statistique Canada soit exhaustif et actuel. Le couplage de listes externes d'entreprises au RSE selon le nom présente des défis sur le plan méthodologique et des calculs, surtout au fur et à mesure que les volumes de données augmentent. Le présent article décrit une méthodologie évolutive qui se fonde sur des techniques d'établissement de blocs pour limiter l'espace de recherche informatique, et intègre de multiples mesures de similarité, des distances d'édition et du chevauchement de n-grammes aux méthodes fondées sur la vectorisation utilisant Sentence-BERT (SBERT), afin de déceler les paires appariées probables. En jumelant des comparaisons simples au niveau des caractères à des méthodes de vectorisation sémantique plus avancées, l'approche peut s'adapter à diverses conventions nominales et différents degrés de complexité. Même si cela ne garantit pas une précision supérieure dans toutes les situations, cette méthode offre un équilibre pragmatique entre la faisabilité des calculs et la qualité du couplage.

Mots clés : Intégration de données; couplage d'enregistrements; mégadonnées; apprentissage profond.

1. Introduction

1.1 Renseignements de base et contexte

Au sein de Statistique Canada, le Registre statistique des entreprises (RSE) soutient un vaste éventail d'enquêtes économiques et de programmes d'analyse. Il sert de base de sondage des entreprises en exploitation au Canada qui fait autorité. Il est régulièrement mis à jour pour assurer la couverture et la représentativité. Pour maintenir la qualité du RSE, il faut intégrer de nouvelles sources de données qui décrivent de manière détaillée des changements dans l'environnement des entreprises, qu'il s'agisse de dossiers administratifs, de listes externes ou d'autres ensembles de données de référence. Parmi les principaux défis de cette intégration, il y a le fait de s'assurer que les entités commerciales provenant de sources disparates peuvent être couplées avec exactitude à leurs entrées correspondantes dans le RSE. Bien que les noms d'entreprises constituent une clé de couplage commune, les variations au chapitre de l'orthographe, les abréviations, la mise en forme et les erreurs typographiques limitent l'efficacité des approches d'appariement exact des chaînes.

1.2 Nécessité d'adopter des méthodes de couplage évolutives

Au fur et à mesure que le volume et la variété des données sources augmentent, les exigences en calcul associées au couplage d'enregistrements deviennent plus prononcées. Des méthodes simples peuvent suffire pour des ensembles de données de petite taille ou relativement uniformes. Cependant, le couplage à grande échelle, comprenant potentiellement des millions d'enregistrements, nécessite des méthodes plus avancées. Sans une conception minutieuse, le nombre de comparaisons par paire peut augmenter de façon quadratique, rendant le processus coûteux et exigeant en temps. Parmi les objectifs clés, il faut donc créer une stratégie qui saisit les variations complexes des

¹**Hanan Ather**, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6 (hanan.ather@statcan.gc.ca); **Serge Godbout** (serge.godbout@statcan.gc.ca), Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6; **Dave MacNeil** (dave.macneil@statcan.gc.ca), Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6.

noms, et qui le fait de manière efficace. Cela permet de veiller à ce que les ressources nécessaires respectent les limites pratiques.

1.3 Approche de couplage à étapes multiples

La méthodologie présentée dans le présent article repose sur un cadre à étapes multiples conçu pour établir un équilibre entre la faisabilité des calculs et la fiabilité des paires liées. La première étape repose sur une procédure d'établissement de blocs pour subdiviser les enregistrements en sous-ensembles plus petits qui ont certains attributs en commun, comme les identificateurs géographiques. En limitant des comparaisons subséquentes avec ces sous-ensembles, on réduit de manière considérable le nombre de paires candidates. À la deuxième étape, on applique un ensemble de mesures de similarité à ces paires candidates. Les méthodes fondées sur les caractères, comme la distance d'édition et les chevauchements de n-grammes, quantifient la ressemblance textuelle, tandis que les approches fondées sur la vectorisation tirent parti de Sentence-BERT (SBERT) pour saisir les relations sémantiques qui transcendent le simple appariement des chaînes. Enfin, des règles de décision se fondant sur des seuils déterminent quelles paires sont considérées comme liées, ce qui permet aux analystes d'adapter les paramètres pour mettre l'accent sur la précision ou le rappel en fonction des exigences de leur cas d'utilisation particulier.

1.4 Objectifs et contributions

Ces travaux ont pour objectif de fournir une méthodologie soutenant l'intégration de données externes sur les entreprises au RSE à une échelle qui correspond aux volumes de données dans le monde réel. Au lieu de prétendre surpasser toutes les méthodes existantes en termes d'exactitude, le présent article met l'accent sur les compromis pratiques qui permettent d'obtenir des gains d'efficacité utiles. En jumelant des stratégies d'établissement de blocs à des mesures de similarité de plus en plus sophistiquées, il vise à faciliter l'utilisation de techniques fondées sur la vectorisation dans des contextes où des comparaisons naïves et exhaustives seraient impossibles. Cette approche propose aux analystes un processus de couplage configurable qui peut tenir compte de différentes conditions de données, contraintes en matière de ressources et attentes au chapitre de la qualité.

2. Contexte et travaux connexes

2.1 Le rôle du Registre statistique des entreprises (RSE)

Le Registre statistique des entreprises (RSE) de Statistique Canada sert d'infrastructure essentielle pour la production de statistiques économiques. Compilé à partir de multiples sources de données et mis régulièrement à jour, il fournit une base de sondage complète des entreprises partout au pays. De nombreux programmes qui comptent sur le RSE, y compris plus de deux cents enquêtes économiques, dépendent de la capacité à maintenir une représentation exacte et courante de la population d'entreprises. Pour garantir cette exactitude, il faut intégrer de nouveaux ensembles de données, préciser les enregistrements au fil du temps et rapprocher les renseignements provenant de fichiers administratifs disparates ou de registres externes (Oyarzun et Wile, 2016).

2.2 Établir un équilibre entre complexité, évolutivité et exactitude

Il est difficile de coupler des enregistrements externes au RSE par nom d'entreprise, en raison des variations au chapitre de la représentation et de la mise en forme. Cela va des simples différences d'orthographe aux divergences plus complexes, comme l'utilisation d'abréviations ou de descripteurs partiels de l'entreprise. L'appariement exact des chaînes suffit rarement, car les homologues légitimes peuvent ne pas avoir du texte identique en commun. Les approches conventionnelles de couplage, même si elles sont efficaces pour les ensembles de données plus petits ou plus homogènes, peuvent avoir de la difficulté à s'adapter efficacement aux fichiers volumineux, devenant exigeantes en termes de calcul et de plus en plus difficiles à gérer au fur et à mesure que les volumes de données augmentent. Lorsqu'il faut traiter des millions de paires liées potentielles, il est essentiel d'adopter des méthodes qui peuvent gérer la complexité, maintenir des niveaux raisonnables d'exactitude et assurer une exécution rapide. L'approche décrite dans le présent article vise à offrir une solution plus facile à gérer sur le plan des calculs pour coupler de grands ensembles de données grâce à des améliorations méthodologiques graduelles, en reconnaissant qu'elle n'est ni parfaite

ni nécessairement supérieure dans tous les cas, mais qu'elle peut améliorer l'efficacité et l'évolutivité en fonction de contraintes pratiques.

3. Méthodologie

3.1 Aperçu du cadre

Cette nouvelle méthode de couplage d'enregistrements novatrice tente de trouver des paires (x, y) , où $x \in X$ provient du RSE et $y \in Y$ provient d'un ensemble de données externe, qui renvoie à la même entité sous-jacente. On attribue à chaque paire un score de similarité $s(x, y) \in [0, 1]$, et on considère les paires qui dépassent un seuil t comme une paire liée. Il est souvent peu pratique de réaliser une comparaison directe de toutes les paires à grande échelle, d'où l'utilisation d'une étape de méthode des blocs qui divise les ensembles de données en sous-ensembles plus petits. On applique ensuite des mesures de similarité, allant des distances d'édition simples aux vectorisations de SBERT, permettant aux opérations les plus exigeantes au chapitre des calculs d'être réalisées où elles sont nécessaires.

3.2 Méthode des blocs pour assurer l'efficacité des calculs

La méthode des blocs réduit l'espace de recherche avant le calcul des mesures de similarité. Une fonction B attribue des enregistrements à des blocs qui se fondent sur des attributs, comme l'information géographique. On compare uniquement les paires (x, y) qui correspondent au même bloc, transformant un processus potentiellement quadratique en somme gérable d'opérations au sein du bloc. Bien que la méthode des blocs puisse manquer certaines paires liées éloignées, elle améliore considérablement l'évolutivité et permet l'utilisation subséquente de méthodes de similarité plus complexes.

3.3 Mesures de similarité fondées sur les chaînes

Les mesures au niveau des caractères et des sous-chaînes continuent d'être des composantes essentielles du pipeline de couplage. Elles fournissent un niveau de comparaison textuelle de référence à un coût de calcul relativement faible.

3.3.1 Distance d'édition (distance de Levenshtein)

La distance d'édition de Levenshtein $d_{LEV}(x, y)$ mesure le nombre minimum de modifications à un caractère requises pour transformer une chaîne en une autre (Levenshtein, 1966). Le score de similarité est extrait comme suit :

$$s_{LEV}(x, y) = 1 - \frac{d_{LEV}(x, y)}{\max(|x|, |y|)}$$

Cette mesure traite des variations orthographiques et des erreurs typographiques simples. Cependant, elle est sensible aux changements dans l'ordre des mots et aux différences structurelles plus importantes.

3.3.2 N-gramme

Les méthodes n-gramme représentent des chaînes sous forme d'ensembles de sous-chaînes de n se chevauchant. Des méthodes fondées sur des ensembles sont également utilisées pour quantifier numériquement les similarités entre deux chaînes. On considère les deux chaînes comme des ensembles de jetons, et la similarité de deux chaînes x et y est une mesure de leur « chevauchement ». Laissons T_x être l'ensemble de jetons produit par la chaîne x , et T_y être l'ensemble de jetons produit par la chaîne y . On définit le chevauchement des deux ensembles à l'aide de

$O(x, y) = |T_x \cap T_y|$. Par exemple, prenons $x = \text{'STATCAN'}$ et $y = \text{'STATISTICS CANADA'}$; $T_x =$

$\{\text{'ST'}, \text{'TA'}, \text{'AT'}, \text{'TC'}, \text{'CA'}, \text{'AN'}\}$ et $T_y =$

$\{\text{'ST'}, \text{'TA'}, \text{'AT'}, \text{'TI'}, \text{'IS'}, \text{'ST'}, \text{'TI'}, \text{'IC'}, \text{'CS'}, \text{'S'}, \text{'C'}, \text{'CA'}, \text{'AN'}, \text{'NA'}, \text{'AD'}, \text{'DA'}\}$. Alors,

$$O(x, y) = |T_x \cap T_y| = \{\text{'ST'}, \text{'TA'}, \text{'AT'}, \text{'CA'}, \text{'AN'}\}$$

Et respectivement.

$$S_{NGRAM}(x, y) = \max\left(0, \frac{O(x, y)}{\text{Longueur de la chaîne}}\right)$$

3.4 Mesures de similarité fondées sur la vectorisation

Dans le traitement du langage naturel (NLP), les vectorisations sont des représentations numériques de texte qui facilitent la compréhension par le programme. Le texte est codé dans un espace vectoriel multidimensionnel, où chaque dimension saisit des propriétés linguistiques ou contextuelles particulières. Dans cet espace vectoriel, la position d'un mot est déterminée par ses valeurs numériques dans de multiples dimensions. Les mots ayant des significations ou des contextes semblables sont cartographiés plus près les uns des autres, permettant au modèle de saisir des relations sémantiques entre eux. Les données de sortie du modèle de vectorisation correspondent à un lieu particulier dans cet espace multidimensionnel, représentant la signification sémantique sous-jacente du texte. Les vectorisations sont cruciales pour saisir la signification sémantique, comprenant le contexte et l'intention du texte. Les modèles de transformateurs, tirant parti des vectorisations, sont mieux à même de reconnaître que ces variations renvoient à la même entité, même lorsque l'ordre des mots est modifié, en saisissant une compréhension sémantique plus approfondie.

3.4.1 Sentence-BERT (SBERT)

Le modèle SBERT est une variante du modèle BERT (représentations de l'encodeur bidirectionnel à partir de transformateurs) optimisé précisément pour produire des vectorisations de phrases ayant une signification sémantique (Reimers et Gurevych, 2019). Au lieu de reposer uniquement sur un objectif de modélisation du langage masqué, SBERT utilise une architecture de réseau siamois dans laquelle deux réseaux de transformateurs se partagent les poids et sont entraînés à réaliser une tâche de classification ou de régression de paires de phrases. Cet entraînement encourage des phrases similaires (en ce qui concerne la signification) à avoir des vectorisations rapprochées dans l'espace vectoriel.

Si l'on prend un nom d'entreprise x , on commence par réaliser un traitement préalable (p. ex. conversion des lettres en majuscules, normalisation des espaces blancs) avant de transmettre le tout au modèle SBERT. L'encodeur SBERT établit une correspondance entre x et un vecteur $v(x) \in R^d$, où d est la dimension de vectorisation (généralement des centaines). Les vectorisations $v(x)$ et $v(y)$ devraient être rapprochées dans l'espace R^d , tenant compte de la signification similaire.

3.4.2 Similarité cosinus pour vectorisations

Pour mesurer la similarité entre deux noms vectorisés, $v(x)$ et $v(y)$, on utilise la similarité cosinus. Par similarité cosinus, on entend

$$s_{cos}(x, y) = \frac{v(x) \cdot v(y)}{\|v(x)\| \|v(y)\|}$$

Le tableau 3.4.2.1-1 montre comment différentes méthodes de notation traitent l'ordre des mots lorsqu'on compare des chaînes, en utilisant « HANAN ATHER TRUCKING INC » comme référence canonique.

Tableau 3.4.2.1-1
Comparaison des différents scores

Nom au RE	Score de Levenshtein	Score 1-gramme	Score 2-grammes	Score 3-grammes	Modèle de base	Modèle adapté
HANAN TAHER TRUCKING	0,75	0,83	0,67	0,58	0,91	0,94
TRUCKING INC HANAN ATHER	0,00	1,00	0,88	0,79	0,93	0,96
ATHER TRUCKING INC	0,75	0,75	0,71	0,67	0,73	0,49
GODBOUT TRUCKING INC	0,58	0,58	0,50	0,46	0,58	0,35
HANAN ATHER PHARMACY INC	0,67	0,67	0,58	0,50	0,59	0,56
ATHER INC	0,38	0,38	0,33	0,25	0,48	0,29

Les approches traditionnelles, comme le score de Levenshtein, sont très sensibles aux changements au chapitre de l'ordre des mots, ce qui se traduit par des scores de similarité significativement plus faibles (p. ex. « TRUCKING INC HANAN ATHER » obtient un score de 0,00). En revanche, les méthodes n-gramme offrent une plus grande robustesse en évaluant les sous-chaînes qui se chevauchent, les valeurs n les plus élevées (par exemple, 2-grammes et 3-grammes) montrant une résilience accrue. Notamment, les modèles BERT adaptés sont plus performants que toutes les autres approches, car ils saisissent des relations sémantiques plus approfondies, atteignant les scores de similarité les plus élevées (0,94 et 0,96) même lorsque l'ordre des mots est modifié.

4. Déroulement des opérations du programme de couplage d'enregistrements

Le déroulement des opérations du programme de couplage d'enregistrements est structuré et compte de multiples étapes pour maximiser l'efficacité, l'évolutivité et l'exactitude. Le processus comporte trois étapes principales, c'est-à-dire l'étape préalable au couplage, le couplage et l'étape après le couplage. Chaque étape comprend des composantes modulaires qui facilitent les transformations systématiques, les comparaisons et la vérification de la qualité.

4.1 Étape préalable au couplage : préparation et normalisation des données

Le processus commence à l'étape préalable au couplage, dans le cadre de laquelle les données sources, y compris les données du RSE et les ensembles de données externes, sont importées. Ces enregistrements font l'objet d'un traitement de texte normalisé, afin d'atténuer les incohérences superficielles qui pourraient avoir une incidence sur la notation de la similarité. Parmi les procédures de normalisation, il y a la conversion du texte en majuscules, la suppression des caractères spéciaux et des accents, l'enlèvement des espaces blancs excédentaires et la compression des espaces. Cela permet de s'assurer que les différences au chapitre de l'orthographe, de la ponctuation ou de la mise en page n'ont pas d'incidence indue sur les comparaisons en aval. Une fois les données normalisées, le programme divise les enregistrements en sous-ensembles gérables à l'aide d'une stratégie d'établissement des blocs. La méthode des blocs réduit la complexité des calculs en regroupant les enregistrements en fonction d'attributs prédéfinis, comme les identificateurs géographiques (p. ex. les codes postaux, les provinces et autres). Par exemple, le programme cherche d'abord des paires liées dans de petits blocs restreints, comme des codes postaux identiques, avant d'élargir graduellement la recherche à des blocs de plus grande taille, comme des enregistrements liés correspondant aux trois premiers caractères du code postal ou au niveau provincial. Pour tenir compte du compromis entre les recherches élargies et l'augmentation du nombre de faux positifs, des pénalités sont appliquées aux scores à chaque tour successif.

4.2 Étape du couplage : appariement des paires candidates et établissement du score

À l'étape du couplage, on détermine, dans chaque bloc, les paires candidates et on calcule les cotes de similarité à l'aide de multiples algorithmes. Les méthodes au niveau des caractères, comme la distance d'édition de Levenshtein, saisissent les variations orthographiques mineures, tandis que les méthodes de chevauchement de n-gramme comparent des ensembles de sous-chaînes contiguës pour déceler les similarités textuelles, quel que soit l'ordre des mots. La conception modulaire permet au programme de calculer ces cotes de manière efficace, en établissant un profil complet pour chaque paire candidate.

4.3 Étape après le couplage : évaluation de la qualité et données de sortie

À l'étape après le couplage, le système évalue les scores de similarité par rapport à des seuils préalablement définis, afin de classer les paires en deux catégories, paires liées ou paires non-liées. Les paires liées font l'objet d'autres évaluations de qualité, qui peuvent inclure une validation croisée avec des attributs supplémentaires (comme des numéros de téléphone, des codes d'industrie), des rajustements à des systèmes de pondération ou une vérification manuelle en présence de cas ambigus. Les résultats sont ensuite réunis dans un tableau définitif de données de sortie, qui comprend les paires liées, les scores qui leur sont associés et les **notes de qualité** tenant compte des niveaux de confiance. Ces indicateurs de qualité permettent aux analystes d'interpréter la fiabilité du couplage et d'adapter leurs décisions en fonction des exigences particulières de leurs applications.

Tout au long du déroulement des opérations, on accorde une attention particulière à l'efficacité des calculs et à la transparence méthodologique. La nature modulaire du système facilite la précision continue, ce qui permet d'intégrer de manière transparente de nouveaux algorithmes de correspondance de chaînes, des modèles de vectorisation ou des stratégies d'établissement de blocs adaptatives au fur et à mesure qu'ils apparaissent. En établissant un équilibre entre l'évolutivité, la précision et la souplesse, ce déroulement des opérations fournit une base solide pour le couplage d'enregistrements, capable d'évoluer en fonction des progrès au chapitre du traitement du langage naturel et des méthodologies statistiques.

5. Défis et leçons retenues

La mise en œuvre du processus de couplage d'enregistrements a révélé plusieurs défis considérables et leçons retenues, qui fournissent des données utiles pour les projets futurs. Ces défis concernent la sélection des variables de la méthode des blocs, l'importance des approches adaptées, les variations des exigences temporelles et l'équilibre entre la qualité du couplage et le taux de couplage global.

5.1 Répercussions des variables de la méthode des blocs

Les variables de la méthode des blocs jouent un rôle essentiel quand vient le temps de déterminer l'efficacité de calcul du processus de couplage. Le choix de variables de la méthode des blocs appropriées influence considérablement le temps de calcul, car elles réduisent l'espace de recherche en regroupant les enregistrements en sous-ensembles. Des variables de la méthode des blocs mal choisies peuvent entraîner des coûts de calcul excessifs ou un manque de paires liées. Il est donc nécessaire d'examiner attentivement les critères de la méthode des blocs et d'en faire l'essai, comme les attributs géographiques ou les identificateurs clés, afin d'optimiser le compromis entre la rapidité et l'exhaustivité.

5.2 Approches adaptées à chaque projet

Il n'existe pas de méthode unique pour le couplage d'enregistrements. Chaque tâche de couplage exige une approche individualisée adaptée aux sources de données particulières, aux objectifs du projet et aux variables visées. La majorité de l'adaptation se produit au **début** du processus, c'est-à-dire au cours de la préparation des données et de l'application de la méthode des blocs, et à la **fin**, étape lors de laquelle les résultats peuvent nécessiter un examen ou une validation manuels. Cette souplesse permet de garantir que la méthodologie va de pair avec les défis uniques que posent les projets individuels.

5.3 Variabilité au chapitre du temps d'achèvement du couplage

Le temps requis pour réaliser le processus de couplage varie considérablement en fonction de certains facteurs, comme la taille de l'ensemble de données, la complexité et le nombre de variables utilisées pour la comparaison. Des projets de plus petite taille peuvent être réalisés en moins d'une heure, tandis que les grands ensembles de données complexes peuvent nécessiter plusieurs jours. Il est crucial de comprendre ces variations de temps pour assurer la planification et l'affectation des ressources, tout particulièrement lorsqu'il s'agit de tâches d'intégration de données à grande échelle.

5.4 Établir un équilibre entre la qualité et le taux d'appariement

Parmi les défis les plus persistants en matière de couplage d'enregistrements, il y a le fait d'établir un équilibre entre la qualité des paires liées et le taux d'appariement global. Si l'optimisation de paires liées de haute qualité garantit l'exactitude, elle peut se faire au prix de l'omission de certaines paires valides. À l'inverse, le fait de chercher à maximiser le taux d'appariement global augmente le risque de faux positifs. Pour établir le bon équilibre, il faut caler soigneusement les seuils, les systèmes de pondération et les mesures de similarité afin de répondre aux besoins particuliers de chaque projet.

5.5 Qualité des données et optimisation du processus

La qualité et l'exhaustivité des ensembles de données ont une grande incidence sur la réussite du processus de couplage. Des données incomplètes ou incohérentes peuvent nuire à l'exactitude et à la fiabilité des paires liées

(Archer, 1995), ce qui souligne l'importance d'un traitement préalable et d'un contrôle de la qualité rigoureux. En outre, l'optimisation du processus, c'est-à-dire l'adaptation de la méthodologie pour relever les défis propres au projet, est essentielle pour assurer un couplage efficace et exact. Chaque projet exige une amélioration itérative des paramètres et des méthodes afin d'obtenir un rendement optimal.

6. Travaux futurs et conclusion

6.1 Travaux futurs

Bien que la présente méthodologie fasse état d'améliorations en termes d'efficacité par rapport aux procédures antérieures de couplage fondées sur les noms, il est encore possible d'apporter des améliorations. Parmi les solutions prometteuses, il y a le fait d'approfondir l'inclusion des vectorisations fondées sur les transformateurs. Bien que l'approche actuelle repose sur des espaces vectoriels sémantiques pour tenir compte de la complexité textuelle, les travaux futurs pourront mettre à l'essai des modèles spécialisés préentraînés axés sur des corpus commerciaux et économiques. Les vectorisations adaptées au domaine pourraient saisir plus efficacement les termes, les acronymes et les modèles de dénomination propres à l'industrie, augmentant ainsi à la fois la précision et le rappel.

Parmi les autres orientations importantes, il y a l'amélioration des algorithmes sous-jacents qui régissent les décisions au chapitre de l'établissement de blocs, de notation et d'appariement. La mise en œuvre de stratégies d'établissement de blocs plus adaptatives, par exemple, permettrait d'adapter dynamiquement les critères de partitionnement en fonction de l'évolution des données entrantes, en optimisant l'efficacité et l'exactitude au fur et à mesure que les ensembles de données prennent de l'expansion. L'intégration de techniques d'apprentissage actif et d'approches semi-supervisées pourrait permettre d'améliorer encore plus la sélection des seuils et les évaluations de la qualité, en réduisant la dépendance à l'égard des seuils arbitraires et en renforçant l'harmonisation entre les résultats du couplage et les normes de qualité définies par l'utilisateur. En outre, la migration du programme vers Python et l'intégration de nouvelles bibliothèques et de nouveaux cadres permettront d'élargir la gamme des outils de NLP accessibles, ce qui permettra finalement de réaliser des mises à jour plus régulièrement et de procéder à des expérimentations sans faille.

6.2 Conclusion

En conclusion, la méthodologie et le système décrits dans le présent article représentent une amélioration considérable de la capacité à coupler de manière exacte et efficace des ensembles de données externes au RSE par nom d'entreprise. Alors que de nouvelles améliorations sont apportées, le programme devrait encore gagner en exactitude, en adaptabilité et en transparence, renforçant ainsi l'infrastructure de données de base qui soutient les statistiques économiques et les efforts de recherche au Canada.

Bibliographie

- Oyarzun, J., et Wile, L. (2016), "An Overview of Business Record Linkage at Statistics Canada: How to Create an Effective Business Register", Statistique Canada.
- Reimers, N., et Gurevych, I. (2019), "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 3982-3992.
- Archer, D. (1995), "Maintenance of Business Registers", in B. G. Cox et al. (eds.) *Business Survey Methods*, New York: Wiley, p. 85-100.
- Levenshtein, V. I. (1966), "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", *Soviet Physics Doklady*, 10(8), p. 707-710.