Nations Unies ECE/CES/2024/5

Conseil économique et social

Distr. générale
11 avril 2024
Français
Original : anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Soixante-douzième réunion plénière
Genève, 20 et 21 juin 2024
Point 3 de l'ordre du jour provisoire
Liaison de données issues de différents domaines et de différentes sources

Examen approfondi de la liaison de données issues de différents domaines et de différentes sources

Communication du Canada

Résumé

Le présent examen approfondi, prescrit par le Bureau de la Conférence des statisticiens européens (CSE), porte sur le rôle qui revient aux organismes nationaux de statistique dans la liaison de données issues de différents domaines et de différentes sources en vue de répondre aux besoins d'information et d'éclairer d'un point de vue statistique des questions multidimensionnelles. Cette note donne un aperçu d'une approche systématique de la liaison de données et de l'expérience des organismes nationaux de statistique dans ce domaine. On trouvera dans la dernière section un résumé des délibérations de la réunion de février 2024 du Bureau de la CSE.

La Conférence est invitée à approuver les résultats de l'examen approfondi.

I. Résumé analytique

1. En février 2023, le Bureau de la Conférence des statisticiens européens (CSE) est convenu de procéder à un examen approfondi de la « liaison de données issues de différents domaines et de différentes sources » sous la conduite du Canada. Cet examen soulève des questions sur la façon dont les organismes nationaux de statistique (ONS) peuvent mettre à profit ce qu'ils ont déjà accompli en matière de liaison de données, notamment en exploitant les données préexistantes, de nouvelles sources de données et les outils et techniques de la science des données afin d'améliorer l'efficacité des systèmes statistiques nationaux (SSN) et de fournir des informations de meilleure qualité sur les grandes orientations et les problèmes multidimensionnels. Les questions qui figurent en tête des priorités d'action mondiales et nationales, telles que la pandémie de COVID-19, l'urgence climatique et les crises de l'énergie et du coût de la vie, mettent en lumière l'interconnexion de l'économie, de la société et de l'environnement. Au demeurant, il est évident que les sociétés ne ressentent pas uniformément l'impact de ces questions et qu'une meilleure granularité de l'information est nécessaire pour tenir compte des inégalités. De ce fait, les décideurs adoptent une vision plus globale des problèmes afin de prendre en considération les corrélations entre différents domaines. Les organismes nationaux de statistique sont ainsi invités à produire des statistiques éclairant ces questions transversales et à fournir des informations plus détaillées dans de multiples domaines.

2. Le présent examen décrit une approche systématique permettant de lier des données en vue de répondre aux besoins de politique générale, d'analyse et d'ordre opérationnel des organismes nationaux de statistique et des différents services. Il est fondé sur des études de cas et une enquête portant sur l'aptitude de ces organismes à agir en tant que coordonnateurs des activités de liaison de données dans des écosystèmes de données de plus en plus complexes. L'attention est également appelée sur les travaux antérieurs de la CSE dans ce domaine.

3. Plusieurs thèmes ressortent des études de cas et de l'enquête. Pour les organismes nationaux de statistique, bon nombre de problèmes de données ont été résolus en établissant des liens entre celles-ci ; cependant, on voit clairement que la liaison de données est non seulement une solution (réactive) à des problèmes, mais aussi une activité (anticipative) axée sur des perspectives. Premièrement, en tant que solutions axées sur les problèmes, les données liées sont couramment utilisées pour remédier à des difficultés qui tiennent à la baisse des taux de réponse, aux données manquantes et à la qualité des données. Deuxièmement, en tant que solutions axées sur des perspectives, elles permettent d'améliorer l'efficacité des SSN en réduisant les coûts des enquêtes, la charge de travail des répondants et la redondance de données. Troisièmement, les données liées constituent un moyen économiquement avantageux de produire des données plus fréquentes et plus réactives, désagrégées au niveau de sous-populations et d'unités géographiques, et susceptibles de déceler des phénomènes sociaux et économiques multidimensionnels qui restent invisibles dans une source unique de données.

4. Des difficultés ont été constatées dans la mise en place d'un écosystème de données qui soit efficace et qui réponde aux besoins d'information de toutes les parties prenantes. Ces difficultés tenaient essentiellement à l'absence d'une structure qui soit chargée de coordonner les données du SSN et d'élaborer un système intégré de données liées. L'une des principales recommandations du présent examen approfondi est la nécessité d'une passerelle utilisateur qui permette d'accéder aux données liées et de fournir les services correspondants. En l'absence de coordination, il y a un risque élevé de fragmentation (manque d'accès), de duplication (redondance et charge de travail des répondants) et d'incohérence des données dans les SSN, ainsi qu'une moindre capacité à observer des phénomènes interdépendants.

5. L'exploitation de sources de données qui ne sont traditionnellement pas utilisées par les organismes nationaux de statistique (l'imagerie satellite et les données de lecteurs optiques, par exemple) offre également des possibilités de repousser les limites de ce qui peut être réalisé grâce à la liaison de données. Cela étant, de telles perspectives s'accompagnent de nouvelles difficultés liées à l'acquisition de ces sources de données, qui sont souvent détenues à titre privé, et dont l'utilisation doit être acceptée par le public. De nouvelles technologies et de nouveaux outils (tels que l'intelligence artificielle (IA) et l'apprentissage

automatique) peuvent aussi être exploités pour mieux tirer parti des liens entre différents domaines, mais les risques associés à ces technologies (d'ordre éthique, technologique, infrastructurel) doivent être pris en compte.

II. Introduction

6. Le Bureau de la Conférence examine régulièrement et de manière approfondie certains domaines statistiques. L'objet d'examens de ce type est d'améliorer la coordination des activités statistiques dans la région de la Commission économique pour l'Europe (CEE), de déceler les lacunes ou les chevauchements d'activité et de traiter de nouveaux enjeux. Ils portent essentiellement sur des questions stratégiques et font ressortir les préoccupations des services de statistique, qu'elles soient de nature conceptuelle ou qu'elles touchent à la coordination.

7. Le Bureau de la Conférence a choisi de procéder à un examen approfondi de la question de la liaison de données issues de différents domaines et de différentes sources à sa réunion de février 2023. Cet examen donne un aperçu des activités statistiques internationales menées dans le domaine considéré, en mettant en évidence les perspectives et les problèmes qui se présentent, et énonce des recommandations sur d'éventuelles mesures de suivi. Le présent examen est piloté par le Canada, avec des contributions de la Pologne et d'autres pays qui ont participé à une enquête pour en étayer les résultats.

8. La liaison de données n'est pas nouvelle et les ONS la pratiquent depuis de nombreuses années. Or le type et le nombre d'ensembles de données mis en relation évoluent, tout comme les domaines dans lesquels les données sont reliées. Cela permet de mieux comprendre des questions complexes qui nécessitent l'adoption de mesures dans de multiples domaines. Comme on le verra ci-dessous, plusieurs examens approfondis de la CSE ont traité la question de savoir **comment** lier des données, qu'il s'agisse des aspects techniques ou des problèmes de confidentialité et de contrôle. Le présent examen approfondi s'attache à présent à déterminer **pourquoi** relier les données et **où** le fait de repousser les frontières en matière de liaison de données suscite des difficultés et ouvre des perspectives. Cet examen a fait apparaître trois domaines à passer en revue à la réunion de la CSE en février :

9. **Premièrement, la liaison de données peut servir à créer un nouveau système permettant aux ONS de produire des statistiques officielles et de répondre aux besoins d'information.** Les SSN disposent d'une quantité considérable de données et les ONS ont la possibilité de les rassembler afin d'être moins tributaires des enquêtes pour les statistiques officielles. Les pratiques traditionnelles de collecte de données ont consisté initialement à réaliser des enquêtes, puis à associer les données à d'autres sources pour combler les lacunes ou améliorer la qualité des données. Ce processus de collecte d'informations peut être inversé pour commencer par l'établissement de liens entre les données, selon une approche consistant à « relier d'abord, collecter ensuite », qui tire parti dans un premier temps des données préexistantes et recourt ensuite à des enquêtes s'il y a lieu pour les compléter. Une telle démarche cadre avec le principe de la saisie unique suivant lequel des données identiques sont collectées une seule fois et réutilisées dans l'ensemble du SSN. Pour réduire le temps que les entreprises passent à répondre à des enquêtes, Statistique Canada utilise par exemple, dans la mesure du possible, les données administratives que les entreprises et les exploitations agricoles ont déjà fournies, notamment les déclarations fiscales et les états de paie des salariés. D'autres sources de données, dont la télédétection et la traçabilité, ont également été évaluées en vue de remplacer les enquêtes auprès des entreprises.

10. **Deuxièmement, les besoins d'information complexes des décideurs sont un facteur déterminant dans la liaison de données issues de différents domaines et de différentes sources.** Les questions qui figurent en tête des priorités d'action mondiales et nationales, telles que l'urgence climatique et les objectifs de développement durable, montrent que l'économie, la société et l'environnement sont des domaines étroitement imbriqués et que les problèmes qui se posent dans ces domaines ne doivent pas être examinés séparément. Des données liées sont nécessaires pour comprendre la nature interdépendante de ces problèmes et l'ampleur de leur impact. Un récent travail de recherche canadien fondé sur des données liées a par exemple montré que la pandémie de COVID-19 avait entraîné des

changements dans l'organisation du travail qui ont des répercussions sur l'utilisation des transports en commun et les émissions de gaz à effet de serre. D'autres projets de liaison de données ont utilisé des sources de données disparates (services de santé, coroner, revenu, système judiciaire, etc.) pour fournir des informations détaillées et nuancées sur la crise des opioïdes au Canada qui étaient invisibles dans des sources uniques. Statistique Canada a également mis en place l'indice en temps réel des conditions d'affaires locales (ITRCAL), qui relie les données du registre des entreprises avec des données commerciales (Google Places, Yelp Fusion et Zomato) et des données sur la circulation routière pour créer des statistiques de fréquence hebdomadaire, voire plus rapprochées, et d'un degré élevé de granularité géographique (au niveau de la ville et du quartier), en vue d'observer les activités commerciales à la suite des perturbations causées par la pandémie de COVID-19 et pendant la phase de reprise. Autrement dit, sans données liées, il y a un risque élevé de fragmentation de l'information ou de compréhension incomplète de questions complexes.

11. **Troisièmement, la disponibilité accrue de nouveaux types de données et la technologie de l'IA offrent de nouvelles possibilités de lier des données à des échelles plus vastes que jamais et d'y accéder.** De nouvelles sources de données sont de plus en plus disponibles − et nécessaires − pour traiter des questions multidimensionnelles et fournir des données indisponibles dans les sources traditionnelles. Un récent rapport de la CEE a étudié comment le fait de relier des données issues de sources traditionnelles (enquêtes, par exemple) à de nouvelles sources (médias sociaux, données de téléphonie mobile, etc.) peut servir à élaborer de meilleures mesures de la migration et de la mobilité transfrontalière. Cependant, la liaison et l'analyse de ces nouveaux types de données se révèlent plus complexes sur le plan technique. Le Groupe de haut niveau sur la modernisation de la statistique officielle a déjà fourni un cadre pour la qualité des mégadonnées et des lignes directrices pour l'établissement de partenariats dans les projets de mégadonnées. L'IA facilite la collecte de données de multiples façons, en permettant de classer les données en catégories, de faire des prédictions sûres et efficaces à leur sujet, et d'accroître la valeur de l'analyse des données. Statistique Canada a par exemple recouru à l'apprentissage automatique pour résoudre les problèmes d'interopérabilité (différentes façons de classer les données) qui existent entre des types similaires de données administratives (données des coroners et médecins légistes, par exemple) provenant de différentes juridictions. Un des autres domaines étudiés par Statistique Canada est l'utilisation de liens préservant la confidentialité des données sensibles, l'objectif étant de pouvoir établir des liaisons et procéder à l'analyse des données liées dans le cadre d'un modèle fédéré, sans avoir à les déplacer d'un endroit à l'autre. Des travaux complémentaires doivent être consacrés à la façon de mettre à profit l'IA pour exploiter les mégadonnées en vue d'utilisations innovantes et d'une meilleure information. L'accès à ce type de données et l'acceptation par le public de liaisons à grande échelle, qui peuvent porter atteinte à la vie privée, posent également des problèmes d'intendance plus importants.

12. Les ONS ont encore des difficultés à surmonter en matière de liaison de données, qu'il s'agisse des questions techniques à régler pour en assurer la faisabilité ou des questions d'intendance concernant le contrôle à exercer et l'acceptation par le public. La principale difficulté identifiée dans le cadre du présent examen approfondi est la nécessité d'une passerelle permettant aux utilisateurs d'accéder aux données liées et aux services correspondants. Compte tenu du grand nombre de fournisseurs et d'utilisateurs de données au sein d'un SSN, il est logistiquement nécessaire qu'un département ou un bureau soit chargé de dresser un inventaire des données disponibles et de coordonner les activités consistant à les lier sur la base des principes FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables).

13. Le contenu du présent examen approfondi est organisé comme suit. La section III décrit les objectifs de l'examen et donne un aperçu d'une approche systématique de la liaison de données. La section IV présente des études de cas et les résultats d'une enquête menée auprès d'organismes nationaux de statistique sur leur rôle et leur expérience en matière de liaison de données issues de différents domaines et de différentes sources. La section V résume les travaux sur la liaison de données déjà réalisés par le Bureau de la Conférence. La section VI porte sur les problèmes et les enjeux qui ressortent de l'examen et la section VII conclut par des recommandations pour la suite des travaux.

III. Domaine statistique visé

14. Le présent examen approfondi se concentre sur le rôle des organismes nationaux de statistique en tant que coordonnateurs des activités de liaison de données dans les SSN. Dans cet examen, la liaison de données est définie comme le couplage ou la compilation de données issues de deux sources ou plus, notamment la liaison de sources administratives provenant de différents services gouvernementaux ou ministères, la liaison de données d'enquête et de données administratives, ou la liaison de données géospatiales et de données d'enquête[1]. La liaison de données ne se limite pas au couplage de registres, même s'il s'agit du type de liaison le plus courant. Elle peut également consister à compiler des informations issues de différents domaines, à établir des modèles et à élaborer des séries d'indicateurs. Comme le montrent les exemples de pays présentés dans la section suivante, la liaison de données est un vaste sujet qui comporte de nombreux aspects, notamment des questions propres à un domaine particulier qui pourraient être inscrites à l'ordre du jour de différents groupes thématiques, et des questions horizontales telles que la terminologie, la sensibilisation, l'acceptabilité sociale, les techniques et la communication.

15. La question est ici de savoir *pourquoi* les ONS interviennent dans la liaison de données issues de différents domaines et de différentes sources, sans nécessairement s'attacher à déterminer *comment* de telles liaisons sont opérées. Ce dernier sujet a été abordé dans des travaux antérieurs de la CSE sur les problèmes technologiques, méthodologiques, administratifs, juridiques et institutionnels que pose la liaison de données. Le présent examen approfondi a trois objectifs :

a) Premièrement, définir et promouvoir une approche systématique de la liaison de données issues de différents domaines et de différentes sources. Cette approche fait suite à une recommandation formulée dans le précédent examen approfondi de la CSE sur l'intégration des données concernant la nécessité de prévoir des procédures normalisées pour faciliter l'intégration des données. L'approche systématique formalise une série d'étapes similaires qui peuvent être mises à profit pour guider les projets d'intégration de données ;

b) Deuxièmement, décrire les rôles actuels et futurs des ONS en tant que coordonnateurs de la liaison de données au sein des SSN et en tant que fournisseurs de services fondés sur des données liées. La CSE a demandé qu'une enquête sur les activités de liaison des données des ONS soit réalisée dans le cadre du présent examen approfondi. Les résultats de l'enquête mettent en évidence des exemples d'utilisation de données liées pour répondre à des besoins de politique générale, d'analyse ou d'ordre opérationnel, ainsi que les enseignements tirés de ces expériences ;

c) Troisièmement, appeler l'attention sur les résultats pertinents de récents travaux de la CSE sur la liaison de données à grande échelle, tels qu'un examen approfondi de l'intégration des données (2017), un examen approfondi de l'éthique des données (2022), un rapport sur l'intendance des données (2023) et un cadre de gouvernance pour l'interopérabilité des données (travail en cours). Le présent examen approfondi aborde brièvement certains de ces projets (sect. V) pour servir de contexte aux questions qui ont été soulevées dans l'enquête et se prêtent à des recommandations.

16. L'examen s'appuie sur les informations tirées des études de cas, de l'enquête et des travaux antérieurs de la CSE en vue de montrer que les ONS sont bien placés pour jouer un rôle de premier plan dans la coordination des activités de liaison de données dans les SSN. Il se conclut par une série de recommandations sur les questions et difficultés à traiter pour que les ONS puissent fonctionner comme des portails d'accès aux données liées et aux services fondés sur celles-ci.

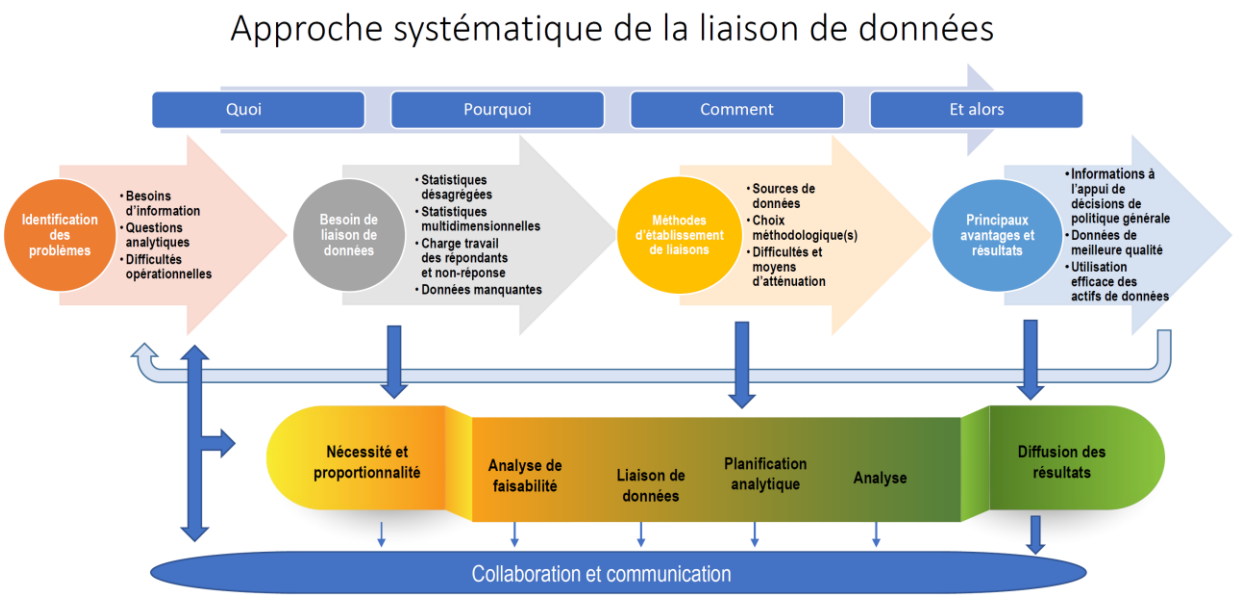
[1] Comme indiqué dans les observations de la CEE sur le présent examen (voir le document ECE/CES/BUR/2024/FEB/2/Add.1, par. 10), la terminologie est un aspect important car, dans certains cas, les mêmes termes peuvent avoir des sens différents selon les pays. Dans la plupart des pays, l'expression « données administratives », par exemple, s'entend uniquement des données conservées dans le secteur public pour des raisons opérationnelles ; dans quelques-uns, elle recouvre toutes les données collectées dans le cadre des activités d'une organisation, y compris les données privées.

Approche systématique

17. Cet examen approfondi souligne la nécessité d'une approche systématique de la liaison de données pour répondre à des besoins de politique générale, d'analyse ou d'ordre opérationnel. Le point de départ de la liaison des données devrait consister à préciser le besoin d'information, en observant les principes de nécessité et de proportionnalité (examinés ci-dessous) afin que les données liées répondent à la nécessité d'une prise de décision éclairée, tout en tenant compte de considérations éthiques (Rancourt, 2019). Un tel impératif cadre avec le constat selon lequel les statistiques sociales « devraient être construites comme un système, soudées par un cadre conceptuel qui prenne en considération les meilleures connaissances et données de la recherche sur les rapports de causalité, et les liens entre les politiques et les résultats » (Fellegi et Wolfson, 1999).

18. Une approche systématique de la liaison des données est guidée par des questions ou des besoins de politique générale, d'analyse et/ou d'ordre opérationnel, éclairée par une concertation avec les parties prenantes et des recherches fondées sur des données factuelles, et ciblée à toutes les étapes. Comme le montre la figure ci-dessous, une telle approche comporte une série d'étapes qui en garantissent une mise en œuvre efficace et l'intérêt pour les utilisateurs.

Approche systématique de la liaison des données



a) La première étape d'une approche systématique de la liaison des données est l'**identification des enjeux** qui consiste notamment à définir quels besoins d'information, questions analytiques ou problèmes opérationnels (charge de travail pour les répondants, données manquantes, etc.) doivent être pris en compte (**quoi**) ;

b) La deuxième étape est le descriptif de projet à étudier afin de déterminer **pourquoi** la liaison de données s'avère nécessaire, notamment en tant que solution efficace pour obtenir des données détaillées et/ou en améliorer la qualité. Cette étape suit les principes de nécessité et de proportionnalité. La *nécessité* est le principe qui a trait aux besoins d'informations, à savoir qui les requiert et les raisons pour lesquelles elles sont nécessaires. La **proportionnalité** est le principe selon lequel les données nécessaires (et pas plus que celles-ci) doivent être obtenues d'une manière qui cadre avec les avantages attendus du projet, tout en tenant compte de l'éthique, de la confidentialité et de la transparence des données. Cette étape s'appuie également sur la consultation des parties prenantes afin de garantir l'inclusion de variables essentielles et une couverture adéquate des sous-populations ou des unités géographiques ;

c) La troisième étape consiste à déterminer les **méthodes d'établissement de liaisons** ou **comment** les données liées seront utilisées pour répondre aux besoins en données spécifiés dans les première et deuxième étapes. Les différentes sources de données doivent être évaluées afin de sélectionner les sources les plus utiles en fonction des variables et des niveaux de désagrégation requis. Les principaux points de contrôle de cette étape sont l'identification des difficultés méthodologiques que soulève la liaison des données, les approches permettant d'atténuer ces difficultés et l'inspection des données pour s'assurer que la taille de l'échantillon et la qualité des données sont suffisantes, tout en respectant la confidentialité des données. Lorsque la liaison de données sert à répondre à un besoin d'information ou à une question d'ordre analytique, un plan est nécessaire pour guider l'analyse et aligner le projet sur les besoins des parties prenantes ;

d) La quatrième étape répond à la question « **et alors** ? », en présentant les principaux avantages et résultats ou la valeur ajoutée de la liaison des données (nouvelles informations numériques, meilleure qualité des données, réduction de la charge de travail des répondants, utilisation efficace des données dans le SSN, etc.). Une présentation précise des avantages de la liaison des données est essentielle pour que le public accepte de tels projets et continue d'y adhérer. Les produits fondés sur des données liées doivent être clairement interprétés et transformés en informations clefs pour éclairer la prise de décision et l'élaboration des politiques. Il convient à cet égard de passer en revue les points forts et les limites des données liées.

19. Une boucle de réaction renforce la corrélation entre les produits et les nouveaux besoins d'information qui peuvent apparaître. La collaboration à mettre d'emblée en place avec les parties prenantes (fournisseurs et utilisateurs de données) est un élément essentiel d'une approche systématique de la liaison de données, afin que les produits répondent aux besoins d'information d'une manière ciblée et rigoureuse et qu'ils favorisent une culture de partage des données dans le SSN.

20. L'approche et la méthode permettant de lier des données dépendent également du contexte propre à chaque pays sur le plan de la législation, du cadre institutionnel et de la protection de la vie privée. La section suivante présente des exemples par pays de liaisons de données qui répondent à différents besoins de politique générale, d'analyse ou d'ordre opérationnel. Ces exemples serviront également à identifier les problèmes posés par les différentes applications, approches ou méthodes adoptées par les organismes nationaux de statistique pour tirer un meilleur parti des informations statistiques.

IV. Contexte et pratiques des pays

21. En général, les pays dont le SSN est relativement centralisé sont mieux placés pour des activités statistiques de liaison de données que ceux qui ont des systèmes décentralisés. Les disparités en matière de données et de métadonnées, dues à l'absence de normes ou d'harmonisation, constituent un obstacle majeur à la liaison de données provenant de sources différentes. Un SSN ne peut fonctionner comme un écosystème sans **interopérabilité** ou sans la capacité logistique d'échanger et d'utiliser de façon cohérente des données issues de différentes sources. La liaison de données entre plusieurs sources peut en outre, dans le cas de systèmes décentralisés, être entravée par des obstacles juridiques au partage des données ou un manque d'adhésion aux initiatives horizontales, ainsi que des obstacles liés à la coordination.

22. Les pays qui ont des systèmes statistiques fondés sur les registres sont ceux qui présentent le plus haut degré de centralisation et les pays scandinaves, dotés de tels systèmes, en ont tiré parti pour devenir des pionniers au niveau mondial en matière de liaison de données. Dans le cas où des registres sont utilisés, les données administratives provenant de différents domaines (santé, éducation, population active, etc.) sont intégrées dans un système statistique. Ces systèmes rationalisent le processus d'intégration des données en ayant satisfait à des conditions préalables essentielles telles que la base juridique requise, l'acceptation par le public de la liaison de données de masse (licence sociale) et la mise en place d'un système d'identification unifié pour intégrer les données recueillies au niveau de

l'unité auprès de diverses sources[2]. Pour les pays dépourvus de registres nationaux, les obstacles juridictionnels à l'accès aux données sont plus nombreux et il est techniquement nécessaire de prévoir un environnement approprié pour rassembler les données.

23. L'un des inconvénients des systèmes centralisés est que l'ONS a tendance à être plus éloigné des débats de politique générale dans certains domaines et peut ne pas avoir les compétences nécessaires pour livrer des statistiques et des produits analytiques pertinents pour l'utilisateur (Edmunds, 2005). Conformément à l'approche systématique, des consultations doivent être menées en continu pour fixer les priorités et veiller à ce que les résultats de la liaison de données répondent aux besoins d'information des services qui composent le système statistique.

24. La liaison de données est essentielle pour la production de statistiques dans tous les pays. Au Royaume-Uni de Grande-Bretagne et d'Irlande du Nord, par exemple, elle revêt une grande importance dans la production statistique de l'ensemble de l'administration publique et dans la coopération avec les universités et d'autres organismes de recherche. On trouvera ci-après des études de cas concernant la liaison de données au Canada, en Pologne et dans la Fédération de Russie qui démontrent en quoi les ONS peuvent jouer un rôle de premier plan dans un écosystème de données complexe. Les résultats d'une enquête menée auprès des ONS sur ce sujet sont présentés dans la section suivante.

A. Canada

25. Statistique Canada joue un rôle de coordination en liant les données administratives issues de sources extérieures à cet organisme. Ce rôle consiste à recenser les besoins d'information, à consulter les fournisseurs et les utilisateurs de données, à effectuer des analyses de faisabilité, à harmoniser les données et à en garantir la qualité, ainsi qu'à développer l'infrastructure informatique et les pratiques institutionnelles pour préserver la confidentialité des données des répondants. L'Environnement de couplage de données sociales (SDLE), l'Environnement de fichiers couplables (EFC) des entreprises et l'Environnement de couplage de données ouvertes (ECDO) sont des exemples d'environnements se prêtant à la liaison de données à grande échelle élaborés par Statistique Canada. Un « environnement de couplage » est une infrastructure de données sécurisée et un système de traitement reliant des enregistrements dépersonnalisés au niveau de l'unité à partir de plusieurs fichiers de données administratives dans un but particulier − il ne s'agit pas d'un fichier de données unique et intégré, car la liaison de sources et de variables spécifiques est effectuée uniquement en fonction des besoins pour les projets approuvés.

26. Outre les « environnements de couplage » de caractère général, Statistique Canada a également mis au point des plateformes analytiques susceptibles de relier différentes sources et différents domaines pour répondre à des besoins d'information complexes. Il convient de mentionner par exemple la Base de données sur la dynamique canadienne entre employeurs et employés (BDCEE), les Cohortes santé et environnement du recensement canadien (CSERCan) et la Plateforme longitudinale entre l'éducation et le marché du travail (PLEMT). La BDCEE offre un exemple de ce type de liaisons de données qui procurent des avantages dans un écosystème de données complexe.

27. La BDCEE fournit des données appariées sur les employés et les employeurs du marché du travail canadien pour permettre des recherches sur divers sujets, notamment les performances du marché du travail et la mobilité de la main-d'œuvre, l'organisation par branche d'activité, le développement économique et la croissance. Il ne s'agit pas d'un jeu de données unique et intégré − la structure des données est un ensemble de fichiers pouvant être corrélés sur la base d'identifiants individuels et professionnels uniques. Ces fichiers proviennent de données administratives de Statistique Canada ainsi que des services du revenu, de l'emploi et du développement social, et de l'immigration.

[2] Pour plus d'informations sur ce type de système statistique, voir le rapport de la CEE intitulé *Register-based statistics in the Nordic Countries* (Statistiques fondées sur les registres dans les pays nordiques).

28. Le BDCEE est conçue pour protéger les données conformément aux dispositions relatives à la confidentialité de la loi sur la statistique, à la politique de Statistique Canada sur l'utilisation de données administratives et à la directive sur le couplage de microdonnées, et adhère également aux principes de nécessité et de proportionnalité définis ci-dessus. Tous les enregistrements au niveau de l'unité sont dépersonnalisés et stockés sur un serveur sécurisé de Statistique Canada. Les données de la BDCEE ne sont pas accessibles au public en raison de leur caractère confidentiel. Pour en faciliter l'accessibilité, une version partielle de la BDCEE (appelée Microdonnées analytiques sur les entreprises et les employés) est disponible dans les Centres de données de recherche de Statistique Canada, qui sont des environnements physiques sécurisés où les utilisateurs de données accrédités affiliés à l'organisation hôte, ayant des projets analytiques préapprouvés et considérés comme des employés de Statistique Canada, peuvent accéder aux données.

29. La BDCEE a créé à l'intention des spécialistes de l'analyse des politiques et des chercheurs des possibilités de combler les lacunes en matière d'information à l'appui d'objectifs de politique générale. Les peuples autochtones et les membres des minorités visibles[3] sont par exemple des groupes désignés dans la loi sur l'équité en matière d'emploi du Canada, mais il n'y avait guère d'informations permettant de suivre leurs progrès en matière de propriété d'entreprises avant la mise en place de la BDCEE. Un tel manque d'informations, considéré par les parties prenantes comme faisant obstacle à l'élaboration de programmes et à l'appui des entreprises aux groupes visés par l'équité en matière d'emploi (Gueye, Lafrance-Cooke, et Oyarzun, 2022), risquait de compromettre la réalisation des objectifs d'inclusion et de diversité. La BDCEE (augmentée d'une liaison supplémentaire avec les données de recensement) a permis de combler les lacunes de l'information sur les propriétaires d'entreprises autochtones et noirs au Canada, en dressant un portrait de leur présence dans l'économie canadienne, de leurs caractéristiques sociodémographiques et des résultats de leurs entreprises au fil du temps à des fins de comparaison.

30. La BDCEE a aussi donné à Statistique Canada la possibilité de produire des informations multidimensionnelles sur l'emploi des Canadiens dans le contexte de l'évolution du marché du travail. Les données de la BDCEE ont fourni des informations sur l'impact de l'essor du secteur des ressources sur l'économie nationale du Canada, montrant comment la migration et les trajets pendulaires à longue distance ont réparti les avantages et les coûts d'un tel boom sur une vaste région géographique (Green *et al*, 2019). L'un des avantages des données fiscales liées est qu'elles présentent une mesure plus précise des salaires des travailleurs pendulaires et non pendulaires et permettent donc de mieux comprendre comment l'essor des régions riches en ressources a des retombées sur les salaires sur le marché du travail d'autres régions. D'autres études de Statistique Canada se sont appuyées sur la BDCEE pour mesurer l'« économie des petits boulots » au Canada et les caractéristiques des travailleurs à la tâche, ce qui permet de mieux comprendre le développement des modalités de travail non traditionnelles (Jeon, Liu et Ostrovsky, 2021) et la façon dont les caractéristiques des entreprises influent sur l'intégration économique des immigrés (Ci et Hou, 2017).

B. Pologne

31. Le Système intégré de métadonnées (IMS) mis au point par Statistique Pologne est un environnement informatique conçu pour générer un registre statistique et exploiter des données issues de différents domaines et de différentes sources. Il permet d'étudier et d'analyser des phénomènes qui ne peuvent être observés séparément dans aucune des sources de données introduites dans le système. Celui-ci se compose de trois listes, à savoir une liste de la population, une liste des bâtiments et des appartements, et une liste des entreprises. Pour la liste de la population, sept sources de données sont utilisées et l'un des principaux

[3] Cette expression utilisée dans le présent document correspond à la catégorie démographique officielle définie par la loi sur l'équité en matière d'emploi, qui est actuellement en cours d'examen (voir Consultation sur l'examen de la Loi sur l'équité en matière d'emploi). Statistique Canada a également mené des consultations afin de déterminer la terminologie appropriée pour le concept de minorité visible (voir Mobilisation consultative sur le concept de minorité visible).

processus consiste à collecter et à fusionner les enregistrements à l'aide des numéros d'identification personnels nationaux provenant de différentes sources de données administratives. Outre le cadre principal des personnes, le système dispose de blocs thématiques distincts (liés au cadre de la population de personnes), qui contiennent des informations de fond sur l'ensemble de la population ou une sous-population particulière, permettant de répondre à divers besoins d'information. Le Système intégré de métadonnées peut être utilisé pour des comparaisons annuelles ou l'analyse de l'évolution de phénomènes démographiques, sociaux et économiques.

32. L'IMS englobe :

a) Le système de traitement des registres administratifs (SPRA) dans lequel les données provenant des registres officiels, des systèmes d'information de l'administration publique et des systèmes d'information non publics sont préparées, transformées, validées, corrigées et intégrées ;

b) Le système de qualité des variables (VQS), qui est utilisé pour faciliter l'analyse plus poussée de la qualité des variables administratives et contrôler les variations des métadonnées des registres ;

c) Le système des opérations statistiques (SOS), permettant de créer un registre de population en intégrant des sources administratives sélectionnées ;

d) Les ensembles de données de domaine (DDS), considérés comme des blocs d'informations thématiques uniques, dont la portée subjective est définie par les rapports du système des opérations statistiques et couvre un sujet particulier qui constitue la base de l'observation des phénomènes sociaux, économiques et spatiaux.

33. Statistique Pologne a mis à profit le Système intégré de métadonnées pour combler d'importantes lacunes sensibles au facteur temps en matière d'information. Pendant la pandémie de COVID-19, le problème a été de déterminer les ressources en personnel médical disponibles en Pologne. Les enquêtes statistiques réalisées étaient fondées sur les rapports des établissements médicaux. Les données obtenues étant agrégées, tout membre du personnel médical ou infirmier travaillant sur différents sites apparaissait plusieurs fois dans ces données. L'autre phénomène faussant les données est le fait qu'une partie du personnel médical vivant à proximité de la frontière nationale exerçait parfois son activité dans un pays voisin. En outre, certains médecins, et plus souvent des infirmiers et infirmières titulaires d'une autorisation d'exercer, travaillaient ailleurs que dans des installations médicales. Grâce à l'intégration de données provenant de multiples sources (registre des médecins, registre du personnel infirmier, système de sécurité sociale, registre des établissements médicaux, etc.), il a été possible d'effectuer des calculs très précis et quasiment en temps réel. Cela a permis de déterminer combien de médecins et d'infirmiers vivent en Pologne, s'ils ont une autorisation d'exercer, s'ils travaillent auprès de patients et dans quel type d'établissement ils exercent. Il a également été possible de présenter les caractéristiques démographiques de ces personnes et leur répartition territoriale.

C. Fédération de Russie

34. Le Service de statistique de la Fédération de Russie (Rosstat) a élaboré un système d'information d'État appelé « plateforme analytique numérique pour la fourniture de données statistiques » (la plateforme). Celle-ci applique les principes suivants : a) transparence dans la collecte et le traitement des informations ; b) utilisation de méthodes harmonisées de collecte et de traitement des données statistiques ; c) fiabilité et cohérence des informations du système ; d) fourniture en un seul point de données à usages multiples. L'activité de la plateforme suppose une interaction avec le système d'information et de calcul de Rosstat ainsi qu'avec des systèmes d'information externes. La capacité de télécharger, de traiter et d'analyser des données (notamment des informations statistiques officielles, des données administratives et des données statistiques primaires) provenant de diverses sources est intégrée dans les sous-systèmes de la plateforme. Les fonctions de la plateforme sont les suivantes :

a) **Calcul et analyse d'indicateurs pour la réalisation des projets prioritaires nationaux de la Fédération de Russie.** Cette fonction a pour objet d'évaluer l'efficacité des projets ainsi que les objectifs de développement national du pays. Les ensembles de données utilisés pour le calcul des indicateurs sont configurés automatiquement à partir de systèmes d'information externes. Le système fournit des algorithmes de calcul des indicateurs fondés sur des méthodes approuvées. Il est procédé à une analyse corrélative pour évaluer le degré de corrélation entre les composantes de l'indicateur et leur effet sur l'indicateur proprement dit ;

b) **Sous-système de méthodes numériques**, pour automatiser les processus d'élaboration, de coordination et d'approbation des méthodes de calcul des indicateurs et les convertir en documents électroniques structurés. Cette fonction a pour objet d'unifier et de systématiser les critères de calcul des indicateurs applicables aux projets nationaux et fédéraux, en garantissant l'utilisation d'approches généralisées axées sur la reproductibilité et la traçabilité des données. Le sous-système est conçu pour : i) rationaliser le processus d'harmonisation des données ; ii) servir de point unique de collecte, de calcul et de communication des valeurs des indicateurs ; iii) évaluer la progression vers les valeurs de l'indicateur, en tenant compte des facteurs qui influent sur celles-ci ; iv) synchroniser les processus de prise de décision dans la gestion de projet et la production statistique ;

c) **Module technologique pour l'évaluation rapide des revenus de la population par tranche de 10 % (décile)**. Le module vise à améliorer la création et la diffusion d'informations statistiques et à élaborer des mesures de réduction de la pauvreté. Le système accumule des données issues de différentes sources (microdonnées d'enquête, tableaux de sortie des formulaires officiels d'observation statistique, données législatives et réglementaires, données administratives et autres) ;

d) **Système de microsimulation des prestations sociales**. Cette microsimulation prédit l'évolution d'indicateurs clefs selon divers scénarios d'évolution des revenus de la population et du seuil de pauvreté. Le système détermine également les mesures d'aide sociale et les dépenses budgétaires les plus efficaces pour obtenir des résultats optimaux. Il est fondé sur l'observation d'un échantillon de données relatives aux revenus de la population et permet de créer différents scénarios d'évolution du revenu individuel en fonction du minimum vital et des variations du seuil de pauvreté. L'interface client est conçue de façon à permettre à l'utilisateur de simuler les revenus provenant des prestations d'assurance sociale et d'autres composantes du revenu telles que l'emploi ou les revenus de la propriété. Le résultat de la simulation est un ensemble de rapports analytiques interactifs composés d'indicateurs calculés (évolution du niveau de pauvreté, des dépenses budgétaires et du revenu par habitant). Il y a également une option de simulation et de prévision fondée sur des données issues de différents domaines de la statistique.

D. Résultats de l'enquête

35. Dans le cadre du présent examen approfondi, une enquête a été réalisée auprès d'ONS qui s'étaient auparavant déclarées disposées à faire part d'informations sur leur expérience en matière de liaison de données issues de différents domaines et de différentes sources. L'enquête comprenait des questions sur : a) le type de SSN ; b) les tâches actuelles ayant trait à la liaison de données ; c) les tâches futures prévues en la matière ; d) des exemples de recours à la liaison de données pour les besoins d'information ; e) les protocoles, outils et infrastructures permettant de faciliter la liaison de données ; f) les enseignements à retenir de la liaison de données ; g) les possibilités de collaboration internationale. Des réponses ont été reçues du Canada, de l'Estonie, de la Hongrie, de l'Italie, de la Lettonie, du Mexique et des Pays-Bas. Ces réponses sont résumées ci-dessous et présentées en détail dans le document ECE/CES/2024/INF.1.

36. **Type de SSN (contexte national).** Le contexte national détermine dans une large mesure l'aptitude des ONS à établir des liaisons entre des données issues de différents domaines et de différentes sources de données. L'enquête invitait les répondants à fournir des informations sur le type de système statistique mis en place dans leur pays, notamment des précisions sur le degré de centralisation de ce système et le rôle de l'ONS dans la

coordination et la production de statistiques officielles. La centralisation des statistiques officielles est chose courante dans les SSN des pays étudiés. La plupart des ONS qui ont participé à l'enquête ont légalement accès aux données administratives de l'ensemble du SSN et coordonnent les activités statistiques. Cependant, même dans les systèmes centralisés, les données et les responsabilités en matière de statistiques officielles dans certains domaines sont dispersées entre divers départements et niveaux de l'administration.

37. **Tâches actuelles des ONS**. Il était demandé dans le questionnaire de décrire les tâches et les responsabilités actuelles de l'ONS concernant la liaison de données administratives issues de différentes sources. Les ONS visés par l'enquête ont généralement un rôle de coordination pour ce qui est de relier des données provenant de différents départements, et l'utilisation de ces données se limite à un usage statistique et est régie par des lois qui garantissent la confidentialité des données. La plupart des ONS ont mis en place l'infrastructure informatique nécessaire à la liaison de données et respectent les lois sur les statistiques afin de garantir la confidentialité des données et de s'assurer qu'elles sont exploitées uniquement à des fins statistiques et ne font pas l'objet d'un usage abusif.

38. **Tâches futures des ONS.** Le questionnaire portait également sur les tâches supplémentaires que l'ONS pourrait assumer dès à présent ou à l'avenir et qui lui permettraient de se repositionner pour passer d'un rôle de fournisseur de données à celui de producteur d'indicateurs statistiques pertinents et d'informations multidimensionnelles. Il a été indiqué que les ONS pouvaient jouer un rôle plus important en coordonnant les activités de liaison de données dans le cadre du SSN. Un rôle accru favoriserait l'utilisation efficace des ressources (actifs de données, capital humain et infrastructure informatique) et l'échange d'informations dans le système statistique. Le thème dominant dans les réponses est que les ONS sont bien placés pour servir de passerelles d'accès aux données liées. L'ONS peut se positionner en tant que « centre national d'analyse et de compétences chargé de la liaison des données administratives et de l'accès à ces données » dans un environnement sécurisé (Lettonie).

39. **Liaison de données pour les besoins d'information.** Le questionnaire demandait de fournir des exemples de besoins d'information qu'il avait été possible de satisfaire dans le pays en liant des données. La plupart des ONS ayant participé à l'enquête ont eu recours à la liaison de données pour répondre à des besoins découlant de problèmes opérationnels (réduction de la charge de travail des répondants, diminution des coûts inhérents aux enquêtes, amélioration de la précision des statistiques, moyens de remédier aux erreurs de couverture et à la baisse des taux de réponse, etc.). Les liaisons axées sur les perspectives visaient à produire plus rapidement des statistiques et des informations sur des phénomènes qui ne peuvent être observés à l'aide d'une seule source de données. Comme l'ont fait observer certains ONS, la liaison de données « est le service le plus utile qui soit à notre disposition − du fait de sa rapidité et de sa souplesse par rapport aux statistiques officielles » (Estonie) et « fournit de multiples solutions au problème de la publication de statistiques démographiques instantanées et précises » (Hongrie). « L'exemple le plus important de liaison de données est probablement le système de bases de données sociales » (Pays-Bas). Celui-est axé sur les perspectives dans la mesure où les statistiques du recensement sont fondées sur des données déjà disponibles dans le système.

40. **Moyens de faciliter la liaison de données**. Il a été demandé dans le cadre de l'enquête s'il existait des protocoles, des outils ou des infrastructures spécifiques (dans le domaine juridique, sur le plan informatique ou en matière de traitement des données) que l'ONS avait mis au point pour faciliter la liaison de données issues de différents domaines et de différentes sources. Les ONS interrogés ont : a) pris des mesures pour favoriser l'interopérabilité des données administratives dans leurs SSN ; b) développé l'infrastructure informatique nécessaire pour relier et partager les données ; c) élaboré des procédures permettant de lier les données administratives pour réduire les redondances et d'autres facteurs d'inefficacité dans leurs systèmes statistiques. Parmi les mesures prises, il convient de mentionner la mise en place d'environnements propices à la liaison de données (infrastructure de données et systèmes de traitement sécurisés pour la liaison d'enregistrements dépersonnalisés), de logiciels libres et de solutions écosystémiques pour le partage des données. L'infrastructure informatique des ONS a également été aménagée de façon à permettre l'application du

principe de la saisie unique, qui vise à collecter les mêmes données une seule fois pour éviter les doublons et réduire la charge de travail des répondants.

41. **Enseignements à retenir.** Il était demandé dans l'enquête de fournir des précisions sur les enseignements tirés de la liaison de données aux fins des besoins d'information et de l'interopérabilité. Deux thèmes en sont ressortis, à savoir la nécessité d'avoir des identifiants uniques pour lier des données et de rationaliser les accords de partage des données entre les SSN.

a) **Utilité des identifiants uniques.** Un code d'identification commun ou un registre unique pour l'ensemble des données administratives est généralement avantageux pour la production de statistiques officielles et favorise les liaisons de données qui permettent de créer des produits innovants répondant aux exigences spécifiques des processus décisionnels et de fournir des données plus détaillées et plus fréquentes que celles que procurent les enquêtes. Faute d'identifiant unique, les liaisons probabilistes peuvent être une solution, mais des travaux complémentaires sont à prévoir pour mettre à l'essai ces méthodes ;

b) **Nécessité d'améliorer les protocoles de partage des données**. Des ONS ont fait remarquer qu'« il reste excessivement laborieux et difficile de solliciter des sources administratives » (Italie) et qu'il y a un « risque de perte du flux de données » (Mexique) vu la nécessité d'établir des accords de partage de données avec des tiers. Les cadres législatifs actuels sont insuffisants pour accéder rapidement aux données et permettre l'accès à celles du secteur privé, ce qui est nécessaire « pour obtenir des données de meilleure qualité et plus diversifiées » (Estonie). Il faut renforcer les accords de partage de données pour formaliser les relations entre l'ONS et d'autres départements et pour faire accepter par la société la liaison de données administratives et l'utilisation de données provenant du secteur privé. Pour atténuer les préoccupations relatives à la confidentialité et coupler des microdonnées provenant de différentes sources, la liaison d'enregistrements préservant la vie privée peut permettre d'effectuer des analyses sur des ensembles de données sensibles sans avoir à déplacer les données hors des institutions qui en ont la garde (Canada).

42. **Possibilités de collaboration internationale.** Il a été demandé dans l'enquête s'il y avait à l'échelle internationale une initiative ou un mécanisme de collaboration que les ONS puissent mettre à profit dès à présent ou envisager pour l'avenir en vue de pouvoir compiler plus efficacement des informations provenant de sources multiples. Les répondants ont recensé plusieurs activités internationales en cours sur les aspects techniques et les enjeux de la liaison de données, qui représentent d'importantes initiatives. Le Groupe de l'application de la science des données et des méthodes modernes et le Groupe d'appui à la mise en œuvre des normes, constitués sous l'égide du Groupe de haut niveau sur la modernisation de la statistique officielle, ont été considérés comme des dispositifs majeurs de collaboration internationale.

V. Travaux connexes réalisés dans le cadre de la Conférence des statisticiens européens et du Groupe de haut niveau sur la modernisation des statistiques officielles

43. La CSE et le Groupe de haut niveau sur la modernisation de la statistique officielle ont mené plusieurs activités sur le thème de la liaison de données. On trouvera ci-après un bref aperçu de ces activités. Les informations présentées dans cette section permettent de situer dans leur contexte les questions soulevées dans le présent examen approfondi et d'éviter tout double emploi avec les travaux antérieurs sur ce sujet.

44. La CSE a procédé à un examen approfondi de l'expérience tirée du projet 2016 du Groupe de haut niveau sur la modernisation de la statistique officielle. Le projet avait pour but d'acquérir une expérience qui permettrait « d'élaborer des recommandations et des directives générales aux fins de l'intégration des données, ainsi qu'un cadre de qualité ». L'examen approfondi de l'intégration des données, présenté à la réunion de la CSE de février 2017, fournissait une description d'ensemble des types d'intégration de données les plus

courants et des expériences menées pour chacun d'eux au niveau national. L'examen portait également sur les enjeux de l'intégration des données, parmi lesquels :

a) **Questions juridiques et institutionnelles** − à savoir la législation sur la confidentialité, la présentation des activités pour en garantir l'acceptation par le public, la collaboration avec les fournisseurs de données et le contrôle des projets d'intégration de données ;

b) **Questions de gestion** − c'est-à-dire les ressources humaines et l'infrastructure informatique indispensables à l'intégration des données, ainsi que les protocoles d'organisation nécessaires pour atténuer les risques inhérents à celle-ci ;

c) **Questions méthodologiques** − autrement dit, les problèmes à surmonter tels que l'absence d'identifiant unifié, les différences entre les concepts et les classifications utilisés pour définir et organiser les données, les données manquantes et les erreurs de couverture.

45. L'une des importantes recommandations de l'examen approfondi était que « l'utilisation de procédures normalisées qui sont communes à différents types d'intégration des données faciliterait grandement une telle intégration ». L'examen a permis de dresser une liste de contrôle des éléments qu'une procédure normalisée d'intégration des données pourrait inclure. L'approche systématique décrite dans la section III du présent examen s'appuie sur cette recommandation.

46. Une équipe spéciale composée d'experts des ONS a élaboré la publication intitulée *Guidance on Data Integration for Measuring Migration* (document d'orientation sur l'intégration des données pour mesurer les migrations). Les décideurs, les chercheurs et d'autres parties prenantes ont besoin de données sur les migrants (nombre, chiffres des entrées et des sorties, caractéristiques, intégration dans la société). Ces données doivent être complètes, précises et fréquemment mises à jour. Il n'y a pas de source unique pouvant fournir de telles données sur les migrations, mais en combinant plusieurs sources, il serait sans doute possible de produire les informations dont les utilisateurs ont besoin. La publication donne un aperçu de la façon dont des données ont été intégrées pour produire des statistiques migratoires, au vu d'une enquête menée auprès des fournisseurs de données sur les migrations dans plus d'une cinquantaine de pays. Treize études de cas fournissent des renseignements plus détaillés sur l'intégration des données dans divers contextes nationaux. La publication énonce des principes relatifs à de bonnes pratiques proposées en matière d'intégration des données pour mesurer les migrations et présente des méthodes permettant de combiner les sources de données administratives, statistiques et autres pour produire des statistiques migratoires.

47. Le cadre de gouvernance pour l'interopérabilité des données (Data Governance Framework for Interoperability) (DAFI) est le fruit d'un projet du Groupe de haut niveau sur la modernisation de la statistique officielle mené en 2022-2023. L'interopérabilité des données s'entend de la capacité d'échanger et d'utiliser les informations sans communication préalable ou avec un minimum de communication. Elle sert également de base à des flux continus de données entre les sources et à la transformation de données cloisonnées « en un réseau connecté d'ensembles de données et de métadonnées harmonisées ». Le cadre de gouvernance pour l'interopérabilité des données décrit les éléments de base nécessaires à la création et à la gestion d'une plateforme de données, de métadonnées et de systèmes interopérables. Le présent examen étant axé sur la liaison de données issues de différentes sources, les recommandations du DAFI s'avèrent tout à fait pertinentes.

48. En 2023, le Bureau de la Conférence a procédé à un examen approfondi de l'éthique des données. L'un des messages clefs de cet examen est que les données liées nécessitent de plus vastes considérations éthiques que les données traditionnelles. Dans le cadre traditionnel, les ONS privilégient surtout l'éthique des affaires et la sécurité des données. L'intégration des données doit faire l'objet d'une meilleure vision de l'éthique des données, qui se concentre sur l'acceptation par le public en sus des questions de confidentialité et de sécurité des données. Dans les enquêtes et les recensements, le processus de collecte des données est transparent car les personnes interrogées savent quelles informations sont recueillies à leur sujet et à quelles fins elles seront utilisées. C'est moins le cas pour les données administratives et les liaisons entre différentes sources de données. Le public peut

ne pas comprendre l'intérêt et la portée des données liées ou ne pas consentir à leur utilisation. Un organisme national de statistique doit donc réfléchir à ce qu'il convient de faire − et pas seulement à ce qu'il est possible de faire − pour garantir l'acceptation par le public des données liées : il lui faudra pour cela faire ressortir les avantages qu'elles procurent et être prêt à annuler des projets lorsque les utilisations futures des données (ou leur mauvais usage éventuel) ne sont pas connues.

VI. Problèmes et enjeux

49. Les organismes nationaux de statistique qui ont participé au présent examen ont tous déclaré avoir un rôle de coordination dans la production de statistiques officielles et un accès légal aux données administratives provenant d'autres administrations publiques. Cependant, les SSN sont des écosystèmes de données complexes comportant de nombreux fournisseurs et utilisateurs de données, ce qui pose des problèmes d'accès aux données et d'efficacité dans l'ensemble du SSN. On trouvera ci-après un résumé des thèmes apparus dans les études de cas et l'enquête menée auprès des ONS concernant leur expérience en matière de liaison de données issues de différents domaines et différentes sources.

A. Rôles actuels et futurs des organismes nationaux de statistique

50. **Les activités actuelles de liaison de données des ONS visent à mieux tirer parti des données préexistantes du SSN pour produire des informations à jour et exactes.** L'augmentation du coût des enquêtes et la diminution des taux de réponse posent un problème à bon nombre d'organismes nationaux de statistique, alors que la demande de réponses rapides et de données désagrégées et multidimensionnelles ne cesse de croître. Les organismes nationaux de statistique étudiés ici ont décrit les raisons pratiques de la liaison de données, qui confirment les avantages de l'intégration des données mentionnés dans le précédent examen approfondi de la CSE sur ce sujet :

a) Les liaisons de données axées sur des problèmes qui sont couramment pratiquées sont fondées sur le principe de la saisie unique selon lequel les données administratives sont intégrées dans les enquêtes pour remplacer des variables (voire l'enquête proprement dite, comme dans le cas de recensements fondés sur des registres) de façon à résoudre le problème de la duplication de données dans l'ensemble du SSN, d'où une réduction des coûts des enquêtes et de la charge de travail des répondants. La liaison de données administratives a également servi à créer des bases d'échantillonnage pour les enquêtes et à remédier aux problèmes de qualité des données dus à l'augmentation des taux de non-réponse et des erreurs de couverture ;

b) Les liaisons axées sur les perspectives sont utilisées pour :

i) Exploiter les flux continus de données administratives comme une ressource permettant de produire des estimations plus fréquentes, de suivre les tendances de plus près et de réagir face à des crises telles que la pandémie de COVID-19 ;

ii) Tirer parti d'un grand volume de données administratives pour améliorer la couverture de populations peu nombreuses et difficiles à atteindre, répondre aux besoins de statistiques désagrégées et observer des phénomènes restant invisibles dans des sources uniques de données.

51. **Les organismes nationaux de statistique sont bien placés pour jouer le rôle de passerelles d'accès aux données liées et aux services fondés sur ces données.** Les personnes interrogées dans l'enquête s'accordent à dire que les ONS devraient se voir confier un rôle de coordination bien défini en raison de la complexité de la liaison des données dans les SSN, composés de nombreuses entités et de quantités croissantes de données administratives. Il est devenu peu pratique d'établir des accords de partage de données sur une base bilatérale et les ressources humaines et informatiques requises pour lier correctement les données sont trop coûteuses pour en faire de multiples reproductions dans l'ensemble du SSN. Les ONS disposent de compétences leur permettant de mettre en place l'infrastructure, les méthodes, les procédures et les protocoles à prévoir pour la liaison de

données, tout en garantissant les principes FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables) et l'éthique des données.

B. Problèmes se posant aux organismes nationaux de statistique

52. **L'absence de rôle clairement défini et de procédure rationalisée de partage des données fait obstacle à la liaison de données issues de différents domaines et de différentes sources.** Le fait de disposer d'un système statistique centralisé et d'une législation permettant aux ONS d'accéder aux données administratives de l'ensemble du SSN n'est pas une condition suffisante pour que les ONS deviennent des portails d'accès aux données liées et aux services fondés sur celles-ci.

a) Si les ONS qui ont répondu à l'enquête dans le cadre du présent examen ont en général légalement accès aux données administratives, ils exercent un moindre contrôle sur la façon dont ces données sont collectées et sur la structure des données, ce dont l'interopérabilité et la qualité des données peuvent pâtir ;

b) L'absence d'identifiant unique a été fréquemment mentionnée comme un obstacle à la liaison de données issues de sources externes, et le recours au couplage probabiliste a été considéré comme une mesure palliative susceptible de dégrader la qualité du couplage de données ;

c) Un processus commun de transfert, de transformation et de liaison des données est nécessaire pour éviter de répéter ce travail pour chaque base de données du système ;

d) Des normes communes applicables à la collecte des données, à la structure des données, aux méthodes, aux identifiants, au vocabulaire et aux définitions font parfois défaut, mais elles sont essentielles à l'interopérabilité et à la réduction du temps et des ressources nécessaires à la liaison de données ;

e) Les modalités d'obtention de données administratives auprès d'autres services sont souvent peu rapides et fastidieuses.

53. Les risques associés à des protocoles de partage de données déficients sont l'allongement du temps de mise en œuvre et la perte de flux de données. Des normes communes relatives aux données sont nécessaires pour que celles-ci se prêtent davantage à des utilisations correspondant aux besoins opérationnels actuels et futurs, ainsi qu'à des fins autres que celles pour lesquelles elles ont été initialement collectées. L'amélioration des normes et l'harmonisation des données (notamment des métadonnées) issues de différents domaines et de différentes sources facilitent la liaison de données et aident les utilisateurs qui peuvent être experts dans leur propre domaine, mais pas dans d'autres.

54. **Il existe une demande de partage des données liées avec les entreprises et les organisations non gouvernementales, ce qui soulève des problèmes supplémentaires.** Le partage des données avec des parties extérieures au secteur public représente en particulier une nouvelle frontière. Les entreprises sollicitent de plus en plus de données et de services connexes auprès des ONS, mais le partage des données avec les entreprises soulève de nouvelles questions liées à l'acceptation par le public, à l'éthique des données et à la protection de la vie privée, qui doivent être résolues. Dans bon nombre de cas, il est interdit de partager des données avec le secteur privé ou des organisations non gouvernementales. Compte tenu de cette asymétrie, les organismes nationaux de statistique doivent trouver un juste équilibre dans la liaison de données et trouver des moyens d'améliorer le partage des données dans le cadre d'un processus plus large d'optimisation et de modernisation de la production statistique.

55. **Il est de plus en plus admis que les mégadonnées et les données privées sont nécessaires pour répondre aux besoins d'information.** Il faut que les organismes nationaux de statistique puissent accéder aux données produites par les entreprises et les technologies de l'information et des communications (« moissonnage de données », par exemple) pour que les statistiques officielles soient mieux à même de répondre aux questions économiques se posant dans le secteur privé et de suivre les questions sociales qui ne sont pas couvertes par les données administratives ou les données d'enquête. La CSE a réalisé un

examen approfondi de la collaboration avec les fournisseurs de données du secteur privé, qui met en évidence les enjeux et les enseignements dégagés jusqu'ici par les organismes nationaux de statistique dans ce domaine.

56. **L'utilisation de données existantes et nouvelles provenant de multiples sources pour élaborer des indicateurs peut entraîner une prolifération d'indicateurs.** Les ONS peuvent se charger de normaliser ou d'harmoniser les indicateurs pour garantir leur cohérence et leur comparabilité entre les différents cadres d'indicateurs. Le fait que les organismes nationaux de statistique assument ce rôle permet également de mettre au point des indicateurs à partir des mêmes sous-ensembles de la population, ce qui n'est pas toujours possible avec des données non liées.

57. Ainsi qu'il ressort des enjeux décrits ci-dessus, outre les changements culturels que les ONS ont dû opérer, des réponses de la part des détenteurs de données administratives et d'autres sources de données sont également nécessaires et importantes pour permettre à ces organismes de se repositionner afin de passer du rôle de fournisseurs de données à celui de producteurs d'indicateurs et d'informations statistiques utiles.

VII. Conclusions et recommandations

58. Le présent examen approfondi a donné un aperçu de la façon dont les organismes nationaux de statistique recourent à la liaison de données pour résoudre des problèmes opérationnels (par exemple, la baisse des taux de réponse) et répondre à des besoins d'information, tout en remédiant aux difficultés soulevées par le partage des données. L'examen a également envisagé la façon dont ces organismes peuvent à l'avenir se repositionner et endosser, en sus de la fonction de fournisseur de statistiques officielles, un rôle complémentaire de portail d'accès aux données liées et de fournisseur d'informations sur des phénomènes multidimensionnels. L'examen s'est appuyé sur des études de cas, une enquête auprès des organismes nationaux de statistique et des travaux antérieurs de la CSE sur l'intégration des données et des sujets connexes.

59. Les recommandations ci-après peuvent être formulées :

a) **Les projets de liaison de données doivent faire l'objet d'une approche systématique.** La liaison de données est une méthode clef à laquelle il est possible de recourir pour répondre à des besoins d'information complexes, mais qui nécessite une approche systématique pour en garantir l'application efficace et la pertinence sur le plan des politiques. Cette approche doit être : i) guidée par des questions ou des besoins de politique générale, d'analyse et/ou d'ordre opérationnel ; ii) éclairée par les travaux antérieurs ; iii) ciblée à toutes les étapes ; qui peuvent être axées sur les problèmes (réactives) ou sur les perspectives (anticipatives) ; et iv) assortie d'une concertation permanente avec les parties prenantes dans l'optique d'un environnement de coopération sur le partage et la normalisation des données et de l'intérêt que des données liées présentent pour l'utilisateur ;

b) Dans le cadre d'une approche systématique de ce type, il faudrait d'abord évaluer les sources de données existantes et d'autres sources pour sélectionner les plus utiles en fonction des variables et des niveaux de désagrégation requis. **Un changement d'attitude quant à la façon de collecter les données est également nécessaire.** Les organismes nationaux de statistique devraient dans un premier temps étudier les options à envisager pour exploiter les sources de données existantes, puis évaluer la possibilité de mettre à profit d'autres sources de données complémentaires issues de différentes sources ou de différents domaines ;

c) **Il faudrait officialiser le rôle de coordination revenant aux organismes nationaux de statistique dans les activités de liaison de données.** L'accès légal aux données administratives de l'ensemble du SSN n'élimine pas les obstacles à l'accessibilité des données provenant de différentes sources, ce qui a des répercussions sur les services que les organismes nationaux de statistique peuvent fournir et sur les moyens de satisfaire les besoins d'information du SSN dans son ensemble. Le fait de doter les organismes nationaux de statistique de la capacité de fonctionner comme des passerelles permettant aux utilisateurs d'accéder à des données liées présente plusieurs avantages. Ce portail peut servir de plaque

tournante pour l'accès aux données liées dans l'ensemble du SSN, en évitant ainsi la répétition de liaisons. Une telle fonction permet de garantir le principe de la saisie unique en utilisant efficacement les données du SSN, de contribuer à la cohérence des liaisons et de réduire la prolifération de cadres d'indicateurs disparates ;

d) **L'élaboration d'une feuille de route est à envisager en vue d'aider les organismes nationaux de statistique à relier des données issues de différents domaines pour communiquer des informations de meilleure qualité aux décideurs qui cherchent à traiter les problèmes multidimensionnels de la société.** Cette feuille de route devrait fournir : 1) des orientations relatives à la gouvernance à mettre en place pour procéder sur une plus grande échelle à la liaison de données issues de différents domaines ; 2) des exemples des données de ce type qui ne sont traditionnellement pas prises en compte et qu'il pourrait être utile de lier pour mieux éclairer l'élaboration des politiques ; 3) des exemples nationaux concrets de liaisons de données multidimensionnelles qui ont été réalisées dans les secteurs de l'action des pouvoirs publics, pour mettre en évidence les liens établis entre différents domaines ;

e) **Les travaux antérieurs de la CSE sur l'intégration des données et des questions complémentaires peuvent être mis à profit pour l'élaboration d'une telle feuille de route.** Les résultats de l'enquête menée aux fins du présent examen approfondi ont mis en évidence les problèmes techniques d'interopérabilité et les problèmes d'intendance des données propres aux accords de partage des données et à la nécessité de faire accepter des données liées par le public. Les personnes interrogées dans le cadre de l'enquête menée pour cet examen ont considéré que, au sein du Groupe de haut niveau sur la modernisation de la statistique officielle, le Groupe de l'application de la science des données et des méthodes modernes et le Groupe de l'appui à la mise en œuvre des normes représentaient d'importantes instances de collaboration internationale sur les aspects techniques de la liaison de données issues de différents domaines et de différentes sources. Les travaux antérieurs de la CSE sur l'interopérabilité et sur l'éthique et l'intendance des données fournissent également des orientations pour rationaliser le partage des données dans les SSN et garantir l'acceptation par le public des projets d'intégration des données. Plutôt que de créer un nouveau groupe pour procéder à l'élaboration d'une feuille de route, il serait sans doute possible de confier cette tâche aux groupes existants.

VIII. Délibérations du Bureau de la Conférence des statisticiens européens

60. Le Bureau a procédé en février 2024 à un examen approfondi de la liaison de données issues de différents domaines et de différentes sources, sur la base d'un document établi par le Canada et d'observations de la CEE. Les observations ci-après ont été formulées :

a) Le document donne une bonne vue d'ensemble des questions ayant trait à la liaison de données, l'accent étant mis sur les questions de stratégie et de gestion, la nécessité de lier des données et les problèmes et perspectives connexes. La proposition consistant à élaborer une feuille de route constituerait une avancée constructive dans les travaux entrepris sur ce sujet au niveau international ;

b) La liaison de données joue un rôle clef dans le repositionnement des organismes nationaux de statistique, qui passent du statut de fournisseurs de données à celui de producteurs d'indicateurs et d'analyses statistiques face au besoin croissant d'informations statistiques multidimensionnelles. Une telle question d'ordre existentiel nécessite un changement culturel de la part des organismes nationaux de statistique, mais aussi des détenteurs de données administratives et d'autres sources de données. La liaison de données n'est pas une nouveauté, mais le type et le nombre d'ensembles de données qui peuvent être liés rendent possible une transformation profonde. Cette question conservera une importance stratégique dans les années à venir ;

c) Une approche systématique est nécessaire pour changer d'attitude et ne pas se contenter de réagir aux demandes. Les organismes nationaux de statistique doivent jouer un rôle formel dans la liaison de données et pourraient devenir des passerelles pour les

utilisateurs en coordonnant l'établissement de telles liaisons à l'intérieur du système statistique national ;

d) Le niveau d'acceptabilité sociale de la liaison de données varie suivant les pays : dans certains, elle est bien admise, tandis que dans d'autres, elle suscite des inquiétudes. Dans les pays où il n'existe pas de système d'identifiant unique ou de registre de population, l'établissement de liens entre les données soulève des difficultés non négligeables, d'où l'importance de la communication, de la transparence et des partenariats ;

e) Les organismes nationaux de statistique peuvent être chargés de lier des données provenant de différentes sources, mais ne pas être autorisés à transmettre les données liées à d'autres institutions pour des raisons juridiques ou d'acceptation par le public. Cette asymétrie peut être difficile à accepter pour d'autres services de l'administration publique. Il faudrait équilibrer les dispositions institutionnelles entre les organismes nationaux de statistique et le reste de l'administration ;

f) Les métadonnées, la terminologie et les classifications doivent être mieux harmonisées pour faciliter les liaisons de données. La notion de « données administratives » revêt par exemple des significations différentes selon les pays ;

g) La question de la liaison de données est liée à celle du partage des données. Une importance identique devrait leur être accordée, en fonction des demandes des utilisateurs. Les questions horizontales et celles qui correspondent à des demandes particulières doivent être examinées simultanément ;

h) La liaison de données permet de produire des données intégrées issues de plusieurs domaines. Il est donc d'autant plus nécessaire de venir en aide aux utilisateurs, qui peuvent être des experts dans leur propre domaine, mais pas dans d'autres ;

i) La liaison de données est un vaste sujet qui présente de multiples aspects, qu'il s'agisse de questions horizontales telles que la terminologie, la sensibilisation, l'acceptabilité sociale, les techniques et la communication, ou de questions propres à un domaine particulier qui pourraient être inscrites à l'ordre du jour de différents groupes thématiques ;

j) Plutôt que de créer un nouveau groupe, il faudrait confier l'examen des questions horizontales et l'établissement d'une feuille de route au Groupe de haut niveau sur la modernisation de la statistique officielle et à ses groupes de l'application de la science des données et des méthodes modernes, et de l'appui à la mise en œuvre des normes ;

k) Un bref questionnaire pourrait être élaboré pour déterminer comment différents pays lient des données. Il permettrait ainsi de recueillir des études de cas et des exemples ;

l) Au Royaume-Uni, la liaison de données joue un rôle important dans la production statistique de l'ensemble de l'administration publique et dans la coopération avec les universités et d'autres organismes de recherche. Au sein de l'Office for National Statistics, l'équipe chargée des liaisons de données souhaiterait être associée aux travaux futurs sur ce sujet ;

m) Le Comité inter-États de statistique de la Communauté d'États indépendants (CIS-Stat) est disposé à participer à l'élaboration d'une feuille de route.

61. Le Bureau est parvenu aux conclusions suivantes :

a) Le Groupe de haut niveau sur la modernisation de la statistique officielle envisagera d'inclure les questions horizontales ayant trait à la liaison de données dans son programme de travail et dans l'ordre du jour de ses groupes − à savoir le groupe de l'application de la science des données et des méthodes modernes, et le groupe de l'appui à la mise en œuvre des normes − dans la mesure du possible, lorsque ces questions sont liées aux mandats des groupes respectifs ;

b) Le Bureau invite le Groupe de haut niveau sur la modernisation de la statistique officielle à élaborer une feuille de route sur la liaison de données, en se fondant sur les résultats de l'examen approfondi ;

c) Les questions ayant trait à la liaison de données issues de différentes sources et de différents domaines devraient être intégrées dans le programme de travail des groupes thématiques travaillant sous l'égide de la Conférence et inscrites s'il y a lieu à l'ordre du jour des réunions d'experts. Dans cette optique, des exemples par pays à partager et à diffuser pourraient être rassemblés ;

d) Le Bureau entend suivre les progrès réalisés dans ce domaine au cours des prochaines années.

Bibliographie

Ci, W., et Hou, F. (2017). Immigrants' initial firm allocation and earnings growth. *Canadian Studies in Population*, *44*(1–2), p. 42 à 58.

Edmunds, R. (2005). Models of statistical systems. Paris: OCDE.

Équipe spéciale de la Conférence des statisticiens européens. (2023). Data stewardship and the role of national statistical offices in the new data ecosystem.

Fellegi, I., et Wolfson, M. (1999). Towards systems of social statistics – Some principles and their application in Statistics Canada. *Journal of Official Statistics*, *15*(3), p. 373 à 393.

Green, D., Morissette, R., Sand, B. M., et Snoddy, I. (2019). Economy-wide spillovers from booms: Long-distance commuting and the spread of wage effects. *Journal of Labor Economics*, *37*(S2), S643–S687.

Gueye, B., Lafrance-Cooke, A., et Oyarzun, J. (2022). Identification des propriétaires d'entreprises autochtones et des entreprises appartenant à des Autochtones. *Études analytiques : Méthodes et références*, No. 045 Statistique Canada, Catalogue nº 11-633-X.

Jeon, S-H., Liu, H., et Ostrovsky, Y. (2021). Mesurer l'économie à la demande au Canada au moyen des données administratives. *Revue canadienne d'économique*, *54*(4), p. 1638 à 1666.

Rancourt, E. (2019). The scientific approach as a transparency enabler throughout the data life-cycle. *Statistical Journal of the IAOS*, *35*(4), p. 549 à 558.