Nations Unies ECE/CES/2017/8

# Conseil économique et social

Distr. générale
27 mars 2017
Français
Original : anglais

## Commission économique pour l'Europe

Conférence des statisticiens européens

**Soixante-cinquième session plénière**
Genève, 19-21 juin 2017
Point 8 a) de l'ordre du jour provisoire
**Résultats des examens approfondis réalisés par le Bureau de la Conférence des statisticiens européens**

# Examen approfondi de l'intégration des données

## Note du Groupe de haut niveau sur la modernisation de la statistique officielle

*Résumé*

Le présent document offre une base pour l'examen approfondi de l'intégration des données que le Bureau de la Conférence des statisticiens européens a effectué en février 2017.

Il fait le bilan du Projet d'intégration des données 2016 mené dans le cadre du Groupe de haut niveau sur la modernisation de la statistique officielle de la Commission économique pour l'Europe. Il donne un aperçu des expériences effectuées en matière d'intégration des données et recense les perspectives, problèmes et enjeux dans ce domaine.

Les conclusions et recommandations sont formulées dans la section VI. Les résultats de l'examen figurent dans le document ECE/CES/2017/8/Add.1.

GE.17-04862 (F) 110417 120417
*1704862*

# I. Résumé

1. Le présent examen porte sur le bilan des expériences en matière d'intégration des données effectuées dans le cadre du Projet d'intégration des données 2016 auquel ont participé les services nationaux de statistique des pays suivants : Brésil, Canada, Colombie, Hongrie, Italie, Nouvelle-Zélande, Pays-Bas, Pologne, Serbie et Slovénie. Plusieurs types d'intégration ont été examinés : données d'enquête et sources administratives, données d'enquête et nouvelles sources de données (mégadonnées incluses), sources traditionnelles et données géospatiales, et intégration de données pour valider les statistiques officielles. Selon les types susmentionnés, plusieurs expériences ont été menées et ont permis de déterminer les perspectives, problèmes et enjeux, d'élaborer de nombreuses recommandations et de tirer des enseignements.

2. L'intégration des données offre de nombreuses perspectives, notamment la production à moindre coût de statistiques nouvelles ou plus d'actualité qui réduisent la charge de travail pour les répondants et peuvent être de meilleure qualité. Les principaux problèmes recensés étaient les suivants : stabilité des sources de données, compétences et techniques nécessaires, différences conceptuelles, besoins en métadonnées, gestion de la qualité, partenariats efficaces, transition de l'expérimentation à la production, coûts de mise en place et de gestion, mesures à prendre pour éviter le chevauchement des activités et perception du public concernant le respect de la vie privée. Il devient de plus en plus nécessaire d'intégrer différentes sources de données à mesure qu'elles se multiplient et que les capacités des technologies de l'information (TI) et de l'infrastructure des données se renforcent.

3. De nombreux types de données et de sources peuvent être intégrés et aucune méthode commune ne peut donc être recommandée pour l'intégration de tous les types de données. Cela étant, il existe plusieurs procédures normalisées parmi les différents types d'intégration. Une série similaire de mesures pourrait être suivie dans les projets d'intégration, par exemple : détermination des besoins et clarification des exigences opérationnelles, identification des partenaires, choix des sources de données potentielles, prise en compte des méthodes et de la qualité, analyse des coûts, des avantages et des risques, et acquisition des données ainsi que des outils, des compétences et des ressources requis. Lors de l'expérimentation, il faudrait évaluer les résultats et perfectionner les méthodes et les approches afin de mettre au point une solution reproductible en matière de production.

4. Il est important d'élaborer des cadres de qualité pour les données intégrées et de répondre aux besoins en métadonnées pour maintenir la valeur des statistiques officielles obtenues de l'intégration des sources de données. De par sa nature, l'intégration des données requiert divers types de partenariats, d'où des difficultés supplémentaires pour les services nationaux de statistique. Il est important d'élaborer des cadres de gouvernance solides pour gérer les projets d'intégration. Outre les compétences statistiques de fond essentielles, il conviendrait d'acquérir de nouvelles compétences en matière de technologie de l'information, des applications et du matériel informatiques, ainsi que les compétences indispensables pour obtenir, négocier et diffuser l'utilisation des sources de données extérieures.

5. Les différentes expériences d'intégration ont débouché sur l'élaboration de nombreuses recommandations, par exemple :

- Obtenir l'appui des hauts dirigeants et informer les responsables des services de statistique du projet ainsi que de ses objectifs et de ses résultats ;
- Maintenir les objectifs bien clairs et expliquer l'importance du projet, préciser les caractéristiques du problème à résoudre ;
- Établir une bonne collaboration avec les partenaires/fournisseurs de données : être clairs sur les besoins en données, prendre en compte les objectifs communs des institutions (ne collecter les données qu'une seule fois, réduire les dépenses inutiles), mettre en place des accords ;

- Collaborer avec les utilisateurs des données ;
- Consulter d'autres experts, en apprendre davantage sur les méthodes et solutions concrètes possibles ; procéder à des échanges de personnel pour acquérir de l'expérience et tirer des enseignements des bonnes pratiques ;
- Prendre en compte les recommandations et les normes internationales en matière de statistique ;
- Partager les résultats ;
- Commencer les travaux sur un échantillon, examiner les données avant de débuter, faire preuve de patience et de persévérance, garder à l'esprit une large gamme de solutions ; et
- Mesurer la qualité des données.

6. Le projet d'intégration de données du Groupe de haut niveau sur la modernisation de la statistique officielle (Groupe de haut niveau) se poursuit en 2017.

# II. Introduction

7. Le Bureau de la Conférence procède périodiquement à un examen approfondi de certains domaines statistiques dans le but d'améliorer la coordination des activités statistiques dans la région de la CEE, de déceler les lacunes ou le chevauchement des activités et d'aborder des questions d'actualité. Cet examen porte essentiellement sur des questions stratégiques et expose les préoccupations d'ordre théorique et en matière de coordination dont les services de statistique ont fait état.

8. En octobre 2015, le Bureau de la Conférence a choisi de soumettre la question de l'intégration des données à un examen approfondi à sa réunion de février 2017. Le présent document offre une base pour l'examen en résumant les activités statistiques internationales menées dans le domaine choisi, en identifiant les enjeux et les problèmes et en formulant des recommandations sur les mesures de suivi susceptibles d'être prises.

9. Par rapport aux approches classiques, l'intégration des données offre la possibilité de produire des statistiques plus d'actualité et plus désagrégées et ce, à des fréquences plus élevées. En 2015, le Groupe de haut niveau a reconnu que les organismes de statistique officiels avaient du mal à créer les capacités nécessaires pour intégrer de nouvelles sources de données dans leurs processus de production statistique, d'où le lancement de son projet d'intégration de données 2016. L'examen repose sur l'expérience acquise et les enseignements tirés par les pays participants.

10. La section III du document décrit les types d'intégration de données visés et la section IV un cadre général pour l'intégration des données ainsi que l'approche large adoptée pour les projets d'intégration. La section V résume les questions, les problèmes, perspectives et enjeux, les mesures d'atténuation des risques, les recommandations formulées, ainsi que les compétences et les ressources requises. La dernière section contient des conclusions et des recommandations succinctes ainsi qu'une proposition à l'intention du Bureau de la Conférence. Un résumé de chaque expérience proposée peut être consulté sur la page wiki du projet d'intégration de données 2016 du Groupe de haut niveau : www1.unece.org/stat/platform/x/HQazBw.

# III. Portée/définition du domaine statistique visé

11. Le projet d'intégration de données 2016 du Groupe de haut niveau 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. Des ressources ont été mises en commun dans le cadre d'activités concrètes conjointes.

12. Grâce à de nombreuses combinaisons de sources de données et de méthodes de modélisation, il est possible d'obtenir de nombreux types d'intégration de données. Tous les types n'auraient pas pu être englobés dans un projet s'étendant sur une seule année. Il a

donc été décidé de privilégier les expériences portant sur quatre principaux types d'intégration, comme indiqué ci-après (le nom du pays qui a proposé une expérience et participé au projet est indiqué entre crochets, mais un plus grand nombre de pays ont peut-être participé à l'expérience lorsqu'elle a été menée).

13. Les différents types d'intégration et expériences retenus étaient les suivants :

**1. Intégration des données d'enquête et des sources administratives**

a) Intégration des informations concernant l'éducation sur la base de l'emplacement géographique des écoles (Colombie) ;

b) Intégration des sources d'information potentielles pour la production de statistiques sur les vacances de poste (Hongrie) ;

c) Corrélation entre le registre des statistiques de l'emploi et l'enquête sur la population active (Slovénie).

**2. Intégration des nouvelles sources de données, telles que les mégadonnées, et des sources traditionnelles**

a) Stratégie d'extraction de données sur le Web pour les statistiques officielles − étude de cas de la Nouvelle-Zélande et étude de l'expérience d'autres pays (Nouvelle-Zélande) ;

b) Intégration des données extraites sur le Web pour la compilation de statistiques relatives aux prix (Hongrie) ;

c) Intégration des données scannées et des données sur les prix extraites du Web afin de permettre la mesure de prix comparables au niveau international (Nouvelle-Zélande) ;

d) Estimation et comparaison des indices de prix provenant de différentes sources de mégadonnées entre les pays (Nouvelle-Zélande) ;

e) Intégration des sources d'information potentielles pour la production de statistiques sur les vacances de poste (Hongrie).

**3. Intégration de données statistiques et géospatiales**

a) Intégration d'objets spatiaux utilisés en statistique et en géodésie − « modèle de niveau de 10 » pour l'harmonisation des cadres de référence en statistique et en géodésie (Pologne) ;

b) Analyse des modèles existants d'intégration des données statistiques et géospatiales (Colombie).

**4. Validation des statistiques officielles à l'aide des données provenant d'autres sources**

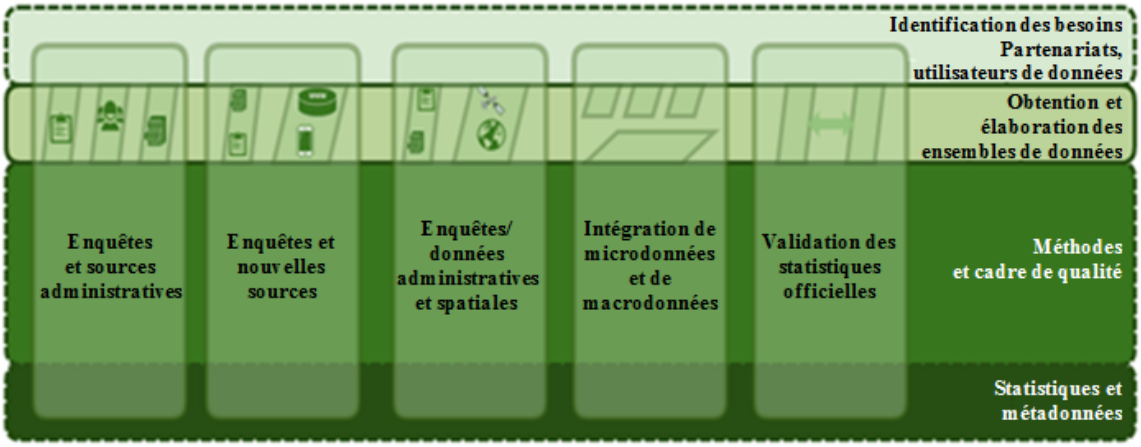
a) Analyse comparative des données sur les revenus issues de l'enquête sur les revenus et des données administratives de la Nouvelle-Zélande (Nouvelle-Zélande) ;

b) Corrélation entre le registre des statistiques de l'emploi et l'enquête sur la population active (Slovénie).

14. Le projet 2016 du Groupe de haut niveau avait mis au jour un cinquième type d'intégration, l'« intégration des microdonnées et des macrodonnées », mais aucune expérience n'avait été entreprise dans ce domaine.

15. Certaines activités concernaient plusieurs types d'intégration. Il existait également des activités transversales dont l'objet était d'obtenir et d'élaborer des ensembles de données, et de faire une synthèse des expériences s'agissant des partenariats, des méthodes, des cadres de qualité et des besoins en métadonnées. Le projet comprenait également une activité visant à identifier et à tester les outils utilisés pour l'intégration des données dans les diverses expériences énumérées ci-dessus. La figure 1 donne un aperçu de la structure du projet.

Figure 1
**Structure du projet 2016 du Groupe de haut niveau**




16. Le présent examen fait également fond sur les informations obtenues grâce au projet 2016 du Groupe de haut niveau et sur l'expérience tirée d'autres travaux d'intégration déjà achevés ou en cours. Il porte sur les problèmes rencontrés et les enseignements tirés, le but étant d'élaborer des lignes directrices pour appuyer et promouvoir les activités d'intégration des données dans les statistiques officielles. Sont mises en relief certaines des questions qui doivent être examinées pour mener des projets fructueux.

17. Les services de statistique ci-après ont contribué au projet : Institut brésilien de géographie et de statistique (IBGE), Bureau central de statistique de la Pologne (CSO), Office central de statistique hongrois (HCSO), Institut national italien de statistique (Istat), Département administratif national de statistique de la Colombie (DANE), Office de la statistique de la République de Serbie (BSRS), Bureau de statistique de la Slovénie (SURS), Statistique Canada, Bureau de statistique des Pays-Bas (CBE), Statistics New Zealand, Eurostat et Commission économique pour l'Europe (CEE).

# IV. Aperçu des expériences concernant l'intégration des données

18. La présente section décrit brièvement un cadre pour les projets d'intégration de données, ainsi que les projets mis en œuvre selon différents types d'intégration. Un résumé de chaque expérience peut être consulté sur la page wiki du projet d'intégration des données 2016 du Groupe de haut niveau : www1.unece.org/stat/platform/x/HQazBw.

19. Lors des expériences, les tâches concrètes commencent par les données. Si aucun ensemble de données adéquat ne peut être partagé entre les organisations, un échantillon ou des ensembles de données synthétisées peuvent être obtenus ou établis afin de partager les méthodes, les outils et les expériences d'intégration. Le « bac à sable » (sandbox) du Centre de calcul de haute performance de l'Irlande a été utilisé pour le traitement et l'expérimentation des grandes sources de (méga)données. Il s'agit d'un environnement relativement ouvert qui n'est pas conçu pour traiter des données sensibles du point de vue du respect de la vie privée ou d'autres données confidentielles. Il est indispensable de prendre en considération diverses questions statistiques, méthodologiques, juridiques et éthiques, mais l'utilisation d'ensembles de données uniquement à titre expérimental rend cette tâche plus facile. Étant issus d'une collaboration, des méthodes, procédures et outils appropriés peuvent être transférés aux environnements sûrs des diverses organisations pour que d'autres essais soient menés sur des données réelles.

## A. Cadre pour l'intégration des données

20. Il existe de nombreux types d'intégration et pour chacun d'entre eux de nombreuses combinaisons de sources de données et de méthodes de modélisation. Les données peuvent être intégrées au niveau des microdonnées, au niveau d'un dénominateur commun, aux niveaux agrégés (macrodonnées) à l'aide de méthodes de modélisation ou d'une combinaison de ces dernières.

21. Il existe cinq types d'intégration courants, à savoir :

a) Association des sources administratives aux données d'enquête et à d'autres données traditionnelles ;

b) Association des nouvelles sources de données (telles que les mégadonnées) aux sources traditionnelles ;

c) Association des données géospatiales aux informations statistiques ;

d) Association des microdonnées aux macrodonnées ; et

e) Validation des données officielles à l'aide de données provenant d'autres sources.

22. Les nouvelles perspectives offertes par le recours à des sources de données multiples et les problèmes posés par l'utilisation des enquêtes comme moyen le plus courant pour produire des statistiques officielles ont conduit la directrice du Committee on National Statistics (Comité de la statistique nationale) des États-Unis, M^me^ Connie Citro, à dire ce qui suit : « Nous devons passer du paradigme de production des meilleures estimations possibles à partir d'une enquête à la production des meilleures estimations possibles pour répondre aux besoins des utilisateurs, à partir de sources de données multiples. » (Citro, 2014)[1]. Le défi à relever est donc d'intégrer divers ensembles de données incohérentes et de créer des produits stables à partir d'apports souvent instables et en constante évolution. Au lieu d'essayer de produire des statistiques à partir d'une seule enquête, le service de statistique officiel doit essayer de trouver la meilleure combinaison de sources afin de fournir les indicateurs/statistiques répondant le mieux aux besoins des utilisateurs.

23. Il faudrait envisager d'utiliser des données d'enquête, des données administratives, des mégadonnées et d'autres sources non traditionnelles. Certes, un certain nombre de tentatives ont été faites pour intégrer diverses sources de données et produire des statistiques, mais il n'existe pas de méthode ou de cadre de qualité généralisé. Pour relever ce défi urgent et complexe courant, le projet du Groupe de haut niveau avait pour objet de mettre en commun les ressources pour commencer par une approche plus systématique en vue de l'élaboration d'un nouveau cadre commun de production statistique.

24. De nombreuses questions devraient être examinées lorsque l'on entreprend des projets d'intégration de données, notamment identifier les besoins, créer des partenariats, obtenir des données, trouver des méthodes de modélisation adéquates, et gérer la qualité, les risques, la comparabilité et les besoins en métadonnées. Une série de thèmes généraux a été établie pour structurer les questions à examiner, à savoir :

- Exigences opérationnelles ;
- Perspectives ;
- Enjeux ;
- Atténuation des risques ;
- Procédures normalisées ;
- Méthodes recommandées ;
- Considérations liées aux technologies de l'information et de la communication (TIC) ;
- Qualité ;
- Normes ;
- Besoins en métadonnées ;
- Travaux connexes menés dans le cadre d'autres projets et organisations ;
- Compétences ;

[1] Citro, C. F. (2014). Des modes multiples pour les enquêtes à des sources de données multiples pour les estimations. *Techniques d'enquête, 40*(2), 151 à 181.

- Ressources ;
- Partenariats ;
- Gouvernance ;
- Promotion et sensibilisation ; et
- Recommandations.

25. Les participants aux expériences ont examiné ces questions, tiré les enseignements pertinents et, le cas échéant, élaboré des recommandations. Le projet a pour but de rassembler les expériences acquises pour formuler, à partir de données synthétisées, des recommandations plus générales ayant trait à plusieurs types d'intégration de données.

## B. Intégration des données d'enquête et des sources administratives

26. Certains pays ont déjà une longue expérience de l'intégration des données d'enquête et des sources administratives. Des projets concertés ont été menés dans ce domaine, par exemple à Eurostat.

27. Il est possible que les données administratives existent depuis un certain temps, mais qu'elles n'aient pas été utilisées. Elles peuvent être intégrées par corrélation de fichiers, par appariement statistique ou en appliquant des méthodes de modélisation. Cette opération peut consister à mettre en commun ou à regrouper des informations émanant d'enquêtes multiples, y compris celles qui ne sont pas menées par les services nationaux de statistique eux-mêmes.

28. Des difficultés sont courantes avec ce type d'intégration. La qualité des ensembles de données administratives peut être suffisamment bonne pour des raisons administratives mais ne pas suffire à des fins statistiques. Pour transformer des données administratives en données statistiques, il peut être indispensable d'améliorer la qualité et de remédier à des différences conceptuelles, en particulier lorsque les données administratives devraient être utilisées directement. Dans le cas des enquêtes fondées sur l'utilisation de données provenant de sources administratives, il est essentiel de rassembler toutes les données (enquêtes et sources administratives) dans une base de données.

29. Comme exemples de sources susceptibles d'être intégrées, citons les enquêtes sur la population active et les registres de l'assurance sociale et/ou de l'éducation, les données des ministères de la culture et des associations culturelles (pour produire des statistiques sur la fréquentation des musées). Il existe plusieurs exemples où des données administratives sont combinées avec des données d'enquête afin de produire des indicateurs traditionnellement recueillis par voie de recensement.

### 1. Description succincte des programmes

30. Les expériences suivantes ont eu lieu :

- Intégration des sources d'information potentielles pour la production de données statistiques sur les vacances de poste ;
- Corrélation entre le registre des statistiques de l'emploi et l'enquête sur la population active ; et
- Système de consultation et de localisation des écoles.

31. Les expériences ayant trait à ce type d'intégration comprenaient généralement les mesures suivantes :

- Définition du cadre statistique et des limites des travaux ;
- Choix des sources administratives et d'un ensemble de données d'enquête ;
- Définition d'un cas à partir des enregistrements issus des ensembles de données administratives au lieu des informations provenant d'une enquête (base de sondage plus petite) ou adjonction d'autres variables émanant de l'enquête administrative (questionnaire plus court), ou les deux ;

- Conception des résultats attendus de l'intégration ;
- Conception et expérimentation de la transformation de l'ensemble de données administratives en données statistiques, une méthode de nettoyage devant être conçue ;
- Conception et expérimentation de l'intégration des données administratives et des données d'enquête ; et
- Évaluation des produits statistiques obtenus à partir de l'intégration des ensembles de données administratives et des données d'enquête.

## 2. Travaux connexes menés dans le cadre d'autres projets et organisations

32. L'intégration des données d'enquête et des sources administratives n'est pas un thème nouveau pour la statistique officielle. Au fil des ans, différentes organisations ont mené des activités dans ce domaine, notamment dans le cadre du programme de travail de la Conférence. Parmi les travaux en cours, on peut citer les suivants :

- Projet ASSIST de la CEE : base de connaissances sur l'utilisation des sources administratives et secondaires en statistique. www1.unece.org/stat/platform/display/adso/ASSIST ;
- Projet ESS.VIP ADMIN : son objet est d'aider les États membres de l'UE à récolter les bénéfices (diminution des coûts et de la charge de travail, augmentation de la disponibilité des données) obtenus en utilisant des sources de données administratives pour la production de statistiques officielles. Il vise aussi à garantir la qualité du produit issu des sources administratives, en particulier la comparabilité des statistiques nécessaires au plan européen. https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources_en ;
- Projet ESSnet sur l'intégration des données : le projet portait en particulier sur les méthodes d'intégration des données (corrélation des enregistrements, appariement statistique, intégration des microdonnées) et sur les aspects statistiques à prendre en compte pour rendre ces méthodes concrètement applicables par les instituts nationaux de statistique. http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en ;
- Projet ESSnet : intégration des données d'enquête et des données administratives. Le but était de promouvoir les connaissances et l'application dans la pratique de méthodes rationnelles pour l'utilisation conjointe des sources de données existantes dans la production de statistiques officielles. http://ec.europa.eu/eurostat/cros/content/isad-finished_en ;
- Eurostat (2013) : utilisation des registres dans le cadre des statistiques communautaires sur le revenu et les conditions de vie (EU-SILC) − défis et perspectives. http://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF ;
- CEE (2007) : statistiques fondées sur des registres dans les pays nordiques. Examen des meilleures pratiques, l'accent étant mis sur les statistiques démographiques et sociales. http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2764 ;
- Projet de la CEE relatif aux indicateurs de qualité pour le modèle générique du processus de production statistique (GSBPM). Il s'agit d'un projet en cours dont l'objet est de mettre au point des indicateurs de qualité pour contrôler la qualité du processus de production statistique pour chacune des phases du GSBPM, y compris la sous-phase « Intégration des données ». Il comporte l'examen et la mise à jour des indicateurs de qualité en vue d'inclure l'utilisation des données administratives dans la production de statistiques officielles. On peut consulter les travaux en cours à l'adresse www1.unece.org/stat/platform/display/QI/Quality+Indicators+Home.

## C. Intégration des nouvelles sources de données (telles que les mégadonnées) et des sources traditionnelles

33. La présente section porte sur l'intégration de nouvelles sources de données (comme les mégadonnées des entreprises de téléphonie mobile, les réseaux sociaux, les capteurs intelligents, les images satellitaires, les pages Web, les transactions par carte de crédit, etc.) aux sources traditionnelles de la statistique officielle.

### 1. Description succincte des expériences

34. Il s'agit des expériences ci-après :

- Stratégie d'extraction de données sur le Web : méthodes d'extraction de données sur le Web pour les statistiques officielles − étude de cas de la Nouvelle-Zélande et étude des expériences d'autres pays ;
- Mesure intégrée de mégadonnées concernant les prix : estimation et comparaison des indices de prix provenant de différentes sources de mégadonnées parmi les pays (Nouvelle-Zélande) ;
- Intégration des données extraites sur le Web pour la compilation de statistiques relatives aux prix (Hongrie) ;
- Intégration des sources d'information potentielles pour la production de données statistiques sur les vacances de poste (Hongrie).

35. Dans le cadre de ces expériences, des mesures telles que les suivantes ont été prises :

- Définition d'un cadre statistique et des limites des travaux ;
- Conception et expérimentation de l'intégration entre les données extraites d'Internet et les ensembles de données statistiques traditionnelles, y compris l'extraction et la reconnaissance d'entités et l'appariement d'objets. Étude de nouvelles techniques de corrélation d'enregistrements et d'appariement d'objets (c'est-à-dire lorsqu'un objet peut avoir une structure plus lâche qu'un enregistrement) ; et
- Conception et évaluation des produits statistiques issus de l'intégration, de la combinaison ou de la fusion des données extraites sur Internet et des ensembles de données statistiques traditionnels.

36. En 2016, les participants à cette activité ont enquêté sur les travaux en cours dans ce domaine (au sein de leur propre organisation et d'autres groupes (par exemple ESSNET et projet de mégadonnées du Groupe de haut niveau)). Les activités se poursuivront en 2017, les objectifs proposés étant les suivants :

- Élaborer un guide pratique pour l'intégration des données d'enquête, des données administratives et des mégadonnées (y compris des études de cas) sur la base des travaux accomplis par les organisations participantes ;
- Encourager la participation d'autres participants et projets ; et
- Décrire les mesures nécessaires au moyen du modèle générique du processus de production statistique (GSBPM).

### 2. Travaux connexes menés dans le cadre d'autres projets et organisations

37. Les travaux menés au sein d'organisations internationales ou dans le cadre d'autres expériences du Projet d'intégration de données 2016 du Groupe de haut niveau sont indiqués ci-après :

- Mégadonnées de la CEE, mégadonnées d'Eurostat ;
- Travaux relatifs aux prix (voir la section B) ;
- Indice des loyers de la Nouvelle-Zélande extrait du site Trade Me (non ajusté en fonction de la qualité) ; et
- Vacances de poste (Serbie).

## D. Intégration des données géospatiales et statistiques

38. La présente section traite de l'intégration des données géographiques et des données statistiques.

39. Les données statistiques sont presque toujours liées à un certain espace physique, comme une municipalité, un État, un pays ou une région. Chaque niveau est utile pour les différents acteurs et les différents types de décision. Nombre de ces décisions sont subordonnées aux éléments physiques de l'environnement, et au-delà, influeront sur ce dernier. La localisation des ressources naturelles et leurs quantités, les types de sols, les conditions météorologiques, l'infrastructure de communication ainsi que les installations sont autant d'exemples d'informations géographiques qui sont indispensables pour bien comprendre les chiffres produits par la statistique officielle. Les travaux en matière de classification géospatiale pourraient servir de point de départ pour établir un lien entre les domaines de la statistique et la géographie.

40. L'intégration des données géographiques dans la statistique officielle vise à améliorer la valeur des données statistiques en cours de production.

### 1. Description succincte des expériences

41. L'ensemble de ces travaux, en raison de sa complexité de fait, n'était pas censé couvrir toutes les activités en une seule année. Les travaux ne faisaient que débuter puisque les activités futures dans ce domaine exigeraient probablement des efforts allant au-delà des activités d'une seule équipe spéciale. Les expériences portant sur ce type d'intégration des données mettaient l'accent sur les points suivants :

- Détermination de l'intégration des objets spatiaux utilisés en statistique et en géodésie − « modèle de niveau de 10 » pour l'harmonisation des cadres de référence en statistique et en géodésie (Pologne) ;
- Analyse d'un système d'intégration des données géospatiales et statistiques (Colombie) ;
- Intégration des informations concernant l'éducation sur la base de l'emplacement géographique des écoles (Colombie).

42. Les mesures suivantes ont été prises :

- Examen des initiatives internationales menées dans ce domaine (systèmes GEOSTAT et GISCO d'Eurostat et système de gestion de l'information géospatiale à l'échelle mondiale (GGIM) de l'ONU) pour créer des synergies et améliorer les capacités de travail dans le cadre du projet ;
- Inventaire des sources et des projets liés aux informations géographiques qui pourraient être utiles pour la production de statistiques ;
- Formulation de propositions concernant la possibilité de produire des informations de grande valeur en intégrant données statistiques et données géographiques ;
- Recherche sur les méthodes d'intégration et de production d'informations pertinentes obtenues en combinant ces deux types de données ; et
- Expériences et projets pilotes visant à évaluer la valeur des données statistiques et géographiques intégrées pour ajuster la portée du projet.

43. Les activités ci-après ont été proposées aux fins d'un suivi en 2017 :

- Élaborer un arbre de décision − série de questions pratiques à résoudre avant de se lancer dans l'intégration des données géospatiales et des données statistiques. Cela permettra d'aider les organisations à évaluer leur maturité et leurs capacités en matière de statistiques spatiales ;
- Examiner les travaux accomplis par les participants au projet GEOSTAT 2 et l'Australie pour référencer les capacités géospatiales et les normes relatives aux données (y compris les normes de la communauté géospatiale) et envisager de les

incorporer dans les composantes du modèle générique du processus de production statistique (GSBPM) et du modèle générique d'informations statistiques (GSIM) ;

- Passer en revue et étudier les activités conjointes avec l'Initiative des Nations Unies sur la gestion de l'information géospatiale à l'échelle mondiale (UNGGIM), l'Europe et les Amériques pour obtenir de meilleurs résultats dans l'intégration des données statistiques et géospatiales ;
- Comparer et analyser les résultats des enquêtes menées dans différentes régions. Mener, si besoin est, des enquêtes supplémentaires fondées sur GEOSTAT2 si les méthodes appliquées dans ces enquêtes sont différentes et si les résultats ne sont pas comparables. Analyser l'existence de points communs et de divergences entre les approches appliquées aux niveaux national et international aux objets spatiaux utilisés en statistique et en géodésie pour harmoniser ces deux cadres de référence (application du « modèle 10 » dans les couches inférieures du cadre mondial de la statistique géospatiale (GSGF)) ;
- Identifier les risques communs pour l'intégration des données géospatiales et des données statistiques.

### 2. Travaux connexes menés dans le cadre d'autres projets et organisations

44. En février 2016, le Bureau de la Conférence a soumis à un examen approfondi la question du développement des services d'information spatiale. Un séminaire sur ce thème a eu lieu en avril 2016 et le Bureau a examiné la suite donnée à ces activités en octobre 2016. À la réunion de février 2017, le Bureau a examiné, au titre du point III k) de l'ordre du jour, une proposition visant à ce que les statisticiens et la communauté géospatiale mènent des activités conjointes dans le cadre de la Conférence.

45. Le cadre d'intégration des données géospatiales et statistiques est très complexe car de nombreux acteurs y interviennent au niveau mondial. Le cadre mondial de la statistique géospatiale (GSGF-UNGGIM), le cadre statistico-spatial (SSF), le projet de fusion des statistiques et des informations géospatiales dans les États membres de l'UE (ESSnet), l'European Forum for Geodesy and Statistics (EFGS) ainsi que des initiatives telles que GEOSTAT2 (Eurostat) sont essentiels pour l'élaboration d'une approche cohérente et systématique concernant la corrélation des données géospatiales et statistiques. Ces mesures risquent de prendre un certain temps.

46. En raison de la complexité de ce domaine, il serait utile d'organiser une séance marathon virtuelle ou en face à face entre les acteurs clefs (notamment l'Australie, la Pologne, la Colombie, le Mexique, la Finlande, la Suède et d'autres entités à déterminer). Ce point devrait être incorporé dans un atelier sur les normes géospatiales et statistiques.

## E. Validation des statistiques officielles

47. Un autre domaine où l'intégration des données entre en jeu est celui où l'on utilise des données provenant d'autres sources pour valider les statistiques officielles. Comme elle permet de recenser les enregistrements émanant de sources de données multiples qui appartiennent à un seul individu ou à une seule entité, l'intégration des données peut servir à valider les statistiques officielles :

- Soit par le recours à des sources de données extérieures pour déterminer l'exactitude des résultats d'une enquête ;
- Soit par le recours aux résultats d'une enquête pour remettre en cause les résultats issus d'autres sources de données de fournisseurs de statistiques.

48. Dans certains cas, d'autres sources ont été jugées comparables aux statistiques officielles, qui ont été remises en cause en cas de divergence. Au Royaume-Uni, un exemple montre comment la répartition des entreprises figurant dans les « Pages jaunes » des répertoires téléphoniques a été évaluée par comparaison aux registres statistiques des entreprises (www.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf) Un autre exemple concerne la comparaison entre les chiffres de l'inflation donnés par le Billion

Prices Project du MIT et les indices des prix officiels. Il ressort de ces exemples que les « autres sources » atteignent un niveau de crédibilité qui remet en question le rôle des statistiques officielles. Les spécialistes des études de marché sont aux prises avec des problèmes analogues. Leurs activités subissent déjà une pression croissante de groupes opérant sur Internet et dont les services sont peu onéreux ou gratuits.

49. Une chose est claire. Les « autres sources » susmentionnées sont là pour durer et augmenteront en nombre et en influence. Plusieurs ODD utiliseront des indicateurs établis à partir de sources officielles qui pourraient ne pas recourir à des méthodes et à des sources de données normalisées. Les différentes approches pourraient entraîner des divergences dans les résultats, lesquelles devront être analysées et expliquées.

50. Les présentes activités portent sur les questions en jeu dans l'intégration des sources de données de remplacement dans les processus de validation utilisés pour produire des statistiques officielles. Il s'agit : d'évaluer l'origine et la qualité de la source, la fiabilité et les intérêts commerciaux ou autres des parties qui exploitent les données ; de concevoir des processus et des techniques de modélisation qui soient durables et officialisés (car il serait difficile de justifier les ajustements ponctuels apportés aux statistiques) ; et d'éduquer les utilisateurs (tant le grand public que des groupes d'utilisateurs plus spécifiques) à bien utiliser et interpréter les informations.

51. En 2017, il est prévu d'étoffer les principes élaborés en 2016, y compris les indicateurs de qualité de la procédure de validation. Les résultats devraient être communiqués au projet du Groupe de haut niveau, qui établit des indicateurs de qualité pour le modèle générique du processus de production statistique (GSBPM).

### 1. Description succincte des expériences

52. Concernant l'utilisation de l'intégration des données dans le but de valider les statistiques officielles, deux expériences ont été réalisées :

- Analyse comparative des données sur les revenus provenant de l'enquête sur les revenus et des données administratives de la Nouvelle-Zélande (Nouvelle-Zélande) ; et
- Corrélation entre le registre des statistiques de l'emploi et l'enquête sur la population active (Slovénie).

53. Les activités suivantes étaient incluses :

- Recensement des questions liées à l'utilisation systématique de données provenant d'autres sources dans la validation des statistiques officielles ; et
- Recommandation des approches et des techniques de modélisation possibles.

54. En 2016, les travaux ont porté sur l'identification des différentes applications et méthodes de validation des statistiques officielles. Les problèmes identifiés concernant l'utilisation des données administratives pour valider les statistiques officielles et les enseignements tirés des expériences de validation ainsi que d'autres projets de validation réalisés au sein de l'organisation des membres contributeurs sont bien étayés et fournissent des lignes directrices initiales sur le recours à des données administratives pour valider les statistiques officielles et recommander des approches et des techniques visant à résoudre les problèmes recensés.

55. Pour 2017, il a été proposé de poursuivre les travaux en vue de mener les actions suivantes :

- Étudier si le référentiel de règles du projet ESSNET et le manuel des définitions en matière de validation du projet ESS.VIP sont pertinents en tant qu'outils de validation ;
- Tester les approches et techniques de modélisation recommandées ;
- Décrire les différentes applications de l'utilisation de l'intégration des données pour valider les statistiques (par exemple, validation des statistiques existantes, remplacement des sources, conception de nouvelles statistiques, amélioration de la

conception des statistiques existantes, remise en cause des résultats provenant d'autres sources/fournisseurs de statistiques) ;

- Élaborer des lignes directrices initiales concernant l'utilisation des données administratives pour valider les statistiques officielles. Inclure une description d'un premier ensemble de prescriptions minimales (il peut s'agir de métadonnées, d'étapes de processus, de méthodes minimales) pour lancer le processus de validation ;
- Déterminer quelles méthodes de validation existent ou pourraient être créées. Inclure les logiciels disponibles pour appliquer des méthodes de validation ;
- Rassembler des informations sur les questions recensées concernant l'utilisation des données administratives pour valider les statistiques officielles ;
- Recommander les approches et les techniques de modélisation nécessaires pour résoudre les problèmes recensés dans l'utilisation des données administratives pour valider les statistiques officielles ;
- Tester les approches et techniques de modélisation recommandées ;
- Rassembler des informations sur l'expérience et les enseignements tirés ;
- Si nécessaire, étoffer les lignes directrices concernant l'utilisation des données administratives pour valider les statistiques officielles. Inclure des mesures ou des indicateurs de qualité qui sont indispensables pour rendre compte de la qualité du processus de validation. Communiquer les conclusions concernant l'élaboration, dans le cadre du projet du Groupe de haut niveau, d'indicateurs de qualité pour le modèle générique du processus de production statistique (GSBPM) ; et
- Mettre au point des supports de formation pour mener des processus de validation fondés sur l'intégration des données.

### 2. Travaux connexes menés dans le cadre d'autres projets et organisations

56. Les projets connexes menés dans le cadre des projets ESSnet et ESS.VIP sont les suivants :

- Projet ESS.VIP ADMIN : son but est de trouver des moyens d'optimiser l'utilisation et l'accessibilité des sources de données administratives dans la production de statistiques officielles, tout en garantissant la qualité et la comparabilité de ces statistiques. Des précisions sont disponibles sur https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources_en ;
- Projet ESSnet sur l'intégration des données : ce projet, qui est achevé, portait sur les méthodes et les aspects méthodologiques de l'intégration des microdonnées. Des précisions sont disponibles sur http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en ;
- Projet ESSnet concernant l'intégration des données d'enquête et des données administratives : il avait pour objet de développer les connaissances et les compétences des services nationaux de statistique participants en matière d'utilisation de données administratives et de données d'enquête intégrées dans la production de statistiques officielles. Des précisions sont disponibles sur http://ec.europa.eu/eurostat/cros/content/isad-finished_en ;
- Projet ESSnet concernant l'intégration des macrodonnées : ce projet examine les différentes méthodes d'intégration des sources de données au niveau agrégé ou au niveau des macrodonnées. Les résultats sont disponibles sur https://ec.europa.eu/eurostat/cros/content/macro-integration_en.

# V. Perspectives, problèmes et enjeux

## A. Perspectives et avantages potentiels

57. L'intégration des données provenant de sources de données multiples permet aux services nationaux de statistique d'étendre leur utilisation des sources de données extérieures à la production de statistiques officielles et offre un grand nombre de perspectives et d'avantages potentiels, ce qui incite fortement les services de statistique à améliorer leurs capacités en la matière. L'intégration des données peut :

- Fournir des statistiques plus d'actualité et plus détaillées ;
- Permettre de créer de nouvelles statistiques officielles ou d'améliorer les statistiques existantes en leur ajoutant des éléments provenant de la source de données extérieure ;
- Répondre à des besoins en données, nouveaux ou non satisfaits ;
- Réduire la charge de travail imposée aux répondants ;
- Vaincre les effets de la réduction des taux de réponse ou remplacer les sources de données existantes ; et
- Améliorer la qualité et régler les problèmes de distorsion dans les enquêtes.

58. L'intégration des données permet également d'améliorer la qualité des sources statistiques traditionnelles en trouvant les sources d'erreur et en déterminant les questions méthodologiques et opérationnelles qui ont une incidence sur la qualité. De nouvelles sources de données peuvent offrir un meilleur échantillon ou même une couverture complète. Elles peuvent être moins chères que les données d'enquête et éventuellement de meilleure qualité. Les données en ligne présentent l'avantage d'être obtenues rapidement et à des fréquences élevées.

59. En Nouvelle-Zélande, par exemple, la promotion des compétences en matière d'intégration des données a conduit à la création de l'infrastructure de données intégrée (IDI) de Statistics New Zealand, qui regroupe des ensembles de données reliés provenant de divers organismes publics (y compris les propres ensembles de données de Statistics New Zealand). L'IDI, qui ne cesse de s'agrandir, est une vaste base de données de recherche renfermant des microdonnées ayant trait aux personnes et aux ménages. Elle a permis de commencer à répondre à des questions de recherche complexes, le but étant d'améliorer les résultats pour les Néo-Zélandais.

60. Un cas particulier de l'intégration des données est l'intégration des données administratives et des données d'enquête. Cela peut servir des fins différentes : compléter les enquêtes par sondage pour une partie de la population ou un ensemble de variables, à des fins d'estimation ou aux fins du processus de validation et d'édition des données. Dans certains cas, une enquête par sondage peut être remplacée par des données reposant entièrement sur des sources administratives. Les données administratives peuvent aussi servir à établir et à tenir à jour les registres statistiques, qui sont ensuite utilisés dans la réalisation d'enquêtes.

61. Les enquêtes par sondage constituent en général des sources plus souples que les sources administratives car elles sont conçues pour répondre à un objectif précis. En revanche, les sources administratives offrent habituellement une meilleure couverture des populations cibles et le taux de réponse est généralement élevé. Il est rentable et moins coûteux d'acquérir des données auprès des sources administratives existantes que de procéder à une enquête par sondage, et il n'y a pas de charge de travail supplémentaire pour les répondants. Du fait que les données administratives englobent des populations entières, il est possible de produire des données locales à un niveau de détail impossible à obtenir avec les enquêtes par sondage. Cela représente aussi un avantage dans la mise en œuvre des politiques locales. Toutefois, les sources administratives posent également des problèmes, qui sont décrits dans la section suivante.

62. L'intégration des données statistiques et des données d'autres sources à des informations géospatiales ajoute de la valeur tant aux données statistiques qu'aux données spatiales. Les autres avantages de l'intégration aux données géospatiales sont les suivants :

- Amélioration de la qualité des données géospatiales et statistiques ;
- Meilleure collaboration entre les organismes de cartographie et les services de statistique, amélioration de la gestion et de l'actualité des ensembles de données ;
- Meilleure interopérabilité des ensembles de données, facilitation des méthodes de corrélation des ensembles de données ;
- Possibilités supplémentaires concernant les analyses (spatiales) et la présentation des données ;
- Création de nouveaux types de services et de données pour répondre aux besoins des utilisateurs ;
- Flexibilité pour les utilisateurs externes ; et
- Utilisations plus larges : pour les responsables politiques et les décideurs, en particulier pour les décideurs régionaux ; à des fins scientifiques ; pour la protection de l'environnement, etc.

## B. Problèmes et enjeux

63. L'intégration des données s'accompagne de nombreux défis. L'expression « intégration des données » peut être largement interprétée : il existe de nombreux types de données et de nombreuses sources qui peuvent être intégrés. Il est donc nécessaire de prévoir différentes approches.

### 1. Questions juridiques et institutionnelles

64. La première et la plus importante question juridique vise le fondement juridique de l'utilisation des sources administratives et d'autres sources extérieures à des fins statistiques. Les projets d'intégration des données menés par les services nationaux de statistique sont soumis à des lois, à des codes de pratique, à des protocoles et à des politiques, dont certains sont énoncés dans la loi sur les statistiques des pays concernés. L'utilisation des sources administratives préexistantes pour les statistiques officielles peut être incluse dans la législation nationale. Dans le cas contraire, il est indispensable d'établir une telle base.

65. Souvent, il n'existe aucun fondement juridique lorsqu'il s'agit d'utiliser d'autres sources de données administratives. Selon le cinquième Principe fondamental de la statistique officielle, « [l]es données utilisées à des fins statistiques peuvent être tirées de toutes sortes de sources, qu'il s'agisse d'enquêtes statistiques ou de fichiers administratifs. Les organismes responsables de la statistique doivent choisir leur source en tenant compte de la qualité des données qu'elle peut fournir, de leur actualité, des coûts et de la charge qui pèse sur les personnes sondées ». Les services nationaux de statistique sont ainsi habilités à utiliser des données de différentes sources, mais rien n'oblige les propriétaires d'autres sources à partager leurs données avec les services de statistique.

66. La législation relative à la confidentialité et à la perception du public est un sujet particulièrement préoccupant lorsqu'il s'agit d'intégrer des données administratives, des mégadonnées ou des données spatiales de faible granularité.

67. Il est important de gérer la perception du public et la communication liées à l'intégration des données pour maintenir la confiance dans les statistiques officielles. Il est important également de faire en sorte que les utilisateurs de données et le grand public comprennent et soutiennent les travaux d'intégration de données menés dans le domaine des statistiques officielles. Pour s'assurer de l'acceptation par le public, il est nécessaire de suivre la législation relative à la protection des données, par exemple la loi sur la protection des données personnelles. Cette législation fixe les règles ayant trait au traitement des données personnelles, de façon qu'il ne soit pas porté atteinte aux droits juridiques des particuliers concernant la vie privée et l'intégrité des données relatives aux personnes.

68. Il est indispensable de collaborer avec les fournisseurs de données administratives pour veiller à ce que la couverture, la qualité des données, les classifications, etc., utilisées dans les sources administratives répondent aux besoins statistiques. Une bonne solution pour venir à bout de ce problème consiste à favoriser la collaboration entre les services nationaux de statistique et les autorités administratives lors de l'élaboration de documents juridiques établissant et gérant une source administrative. L'approbation des services nationaux de statistique lors de l'adoption de lois sur les fichiers administratifs peut être énoncée dans une loi sur les statistiques. Il sera peut-être nécessaire de signer des accords de coopération pour répartir les tâches entre les parties aux accords et de définir les règles et les conditions de transfert des données telles que l'actualité, la mise en œuvre technique et les métadonnées.

69. L'intégration des données offre la possibilité d'établir des partenariats en vue d'une collaboration et d'un échange de données d'expérience avec les instituts de statistique, les producteurs de données, les universités, les établissements de recherche, le secteur privé, les institutions publiques, les organisations commerciales, y compris les fournisseurs d'infrastructures TIC. Ces partenariats vont de pair avec tous les problèmes liés à la mise en place de partenariats stratégiques (qui ont été examinés lors du séminaire de la Conférence en avril 2016). Dans certains cas, les mêmes données sont utilisées par plusieurs institutions, de sorte qu'il est recommandé de mettre en place une collaboration permanente entre les groupes méthodologiques institutionnels afin d'élaborer un système qui soit satisfaisant à des fins administratives et statistiques.

70. La transition de projets d'intégration de données du stade de la recherche à celui de l'établissement reproductible et fiable de statistiques n'est pas exempte de défis, notamment en ce qui concerne la gouvernance des projets d'intégration. Plusieurs pays ont élaboré des cadres de gouvernance solides pour la gestion de ces projets. Par exemple, l'Australie dispose d'un cadre d'intégration des données statistiques (« National Statistical Service § Data Integration Need to know », 2014) qui a été approuvé et est utilisé par le Gouvernement australien. D'autres pays ont des cadres analogues.

### 2. Questions de gestion, notamment pour les ressources

71. Une difficulté générale pour l'intégration des données est de disposer des bonnes ressources (budget, ressources humaines et infrastructure TI) pour fournir le savoir-faire, les connaissances et les technologies nécessaires. Les restrictions budgétaires pourraient restreindre la capacité d'obtenir les ressources requises.

72. Les ressources humaines comprennent des statisticiens experts en la matière, des spécialistes de la méthodologie et des experts en technologies de l'information. Il est indispensable de disposer d'une grande variété de compétences, tant pour l'exploitation des données que pour l'obtention, la négociation et la communication de l'utilisation des sources de données extérieures. Il peut s'agir de compétences en matière de leadership, de négociation, de tissage de relations, de droit, de protection des données et de communication, etc. Le personnel travaillant dans le domaine de l'intégration des données doit connaître les diverses politiques et dispositions législatives en la matière, y compris la législation qui protège les données individuelles.

73. Les compétences de fond requises sont des connaissances spécialisées en contenu des données et en processus statistiques, la compréhension des mesures de la qualité, etc. Il est primordial que les ressources humaines soient dotées des compétences appropriées en matière de technologies de l'information, de logiciels et de bases de données. Pour des projets d'intégration spécifiques, tels que l'utilisation des mégadonnées, il pourrait être nécessaire de connaître les outils d'extraction de données sur le Web et de posséder des compétences en technologies de l'information pour la gestion et le traitement des mégadonnées. Si des sources de données géospatiales sont utilisées, il faudra disposer de compétences dans le domaine géospatial et connaître les procédures pertinentes existantes.

74. L'évolution rapide des technologies de l'information (matériel et large éventail d'outils logiciels) facilite l'intégration des données. L'infrastructure TI nécessaire comporte des serveurs, des logiciels et des outils de développement de bases de données hébergeant les microdonnées et les métadonnées, ainsi que des outils de traitement et de diffusion des

données. Les mécanismes de transfert de fichiers doivent être sûrs et efficaces. L'utilisation des mégadonnées requiert des ressources TI spécifiques pour le stockage et le traitement de volumes importants de données. Pour les données géospatiales, davantage de ressources infrastructurelles sont nécessaires puisque les données restent désagrégées au-delà du traitement initial et ce, jusqu'à la diffusion. Ces services peuvent être externalisés, auquel cas les fonds demandés à cet effet devraient être disponibles.

75. L'intégration des données s'accompagne de risques qu'il faut gérer. Les statistiques officielles doivent donner des produits stables alors que les données d'entrée à intégrer sont souvent instables. En raison des risques d'interruption des sources de données, il faudrait mettre en place des plans d'urgence en cas d'indisponiblité de ces sources.

76. Pour atténuer les risques, il faudrait utiliser des cadres organisationnels de gestion des risques ainsi que les directives relatives aux pratiques de gestion des risques dans les services de statistique (en cours d'élaboration par l'Istat). Des méthodes d'évaluation des risques peuvent être requises pour certaines sources de données.

77. L'intégration des données peut être coûteuse à mettre sur pied. La gestion d'un projet d'intégration de données continu implique aussi des coûts. Il peut donc être recommandé, dans un premier temps, de se préoccuper avant tout d'utiliser des ressources limitées pour produire des résultats tangibles et utiles.

78. Les sources de données extérieures ne sont pas forcément disponibles gratuitement. Évaluer leur qualité implique aussi des coûts. Les données administratives sont généralement moins onéreuses que les données issues des enquêtes par sondage puisqu'elles sont déjà recueillies à des fins administratives, mais un certain budget est quand même nécessaire. De même, l'utilisation des mégadonnées nécessite du temps et un budget correspondant afin d'établir et d'expérimenter les données.

79. Il peut être nécessaire de recourir à des méthodes différentes pour des sources potentielles différentes. Avant de mettre en œuvre un projet d'intégration, il faut donc tester/étudier chaque source de données et il faut déterminer et évaluer tous les coûts connexes.

80. Une autre difficulté réside dans la réticence à changer toute partie d'un processus de production en cours qui comprendra l'intégration d'une source de données extérieure, en particulier lorsque les méthodes existantes sont largement acceptées et qu'un savoir-faire bien établi a été mis en place.

81. L'utilisation de procédures normalisées qui sont communes à différents types d'intégration des données faciliterait grandement une telle intégration. Nombre de projets d'intégration pourraient suivre la même série de mesures, telles que :

- Détermination des besoins et clarification des exigences opérationnelles ;
- Recherche des travaux connexes menés dans d'autres organisations ;
- Identification des partenaires et des collaborateurs ;
- Choix des sources de données potentielles ;
- Identification des considérations méthodologiques et relatives à la qualité ;
- Analyse de la question de savoir si et comment les sources de données potentielles peuvent être utilisées (comme sources directes ou pour valider des données) ;
- Réalisation d'une étude de faisabilité (notamment concernant les coûts, les avantages, les risques, etc.) ;
- Acquisition des données ;
- Acquisition des outils, des compétences, des ressources nécessaires ;
- Expérimentation ;
- Évaluation des résultats ;
- Perfectionnement des méthodes et de l'approche (selon les besoins) ; et
- Mise au point d'un mode de production reproductible.

82. Il existe également un certain nombre de tâches pour lesquelles il serait utile de disposer de procédures normalisées, par exemple :

- Suppression des incohérences entre les enregistrements corrélés ;
- Imputation des valeurs d'une variable dans les enregistrements corrélés d'un ensemble de données intégré ;
- Détermination des coefficients de pondération pour tenir compte des liens interrompus dans un ensemble de données intégré ; et
- Application des méthodes de validation.

### 3. Méthodes, concepts et définitions

83. La méthode d'intégration des données dépend largement des sources de données utilisées et aucune méthode commune n'est recommandée pour l'intégration de tous les types de données. Le projet d'intégration du Groupe de haut niveau prévoit de constituer un catalogue des méthodes nouvelles et communément utilisées pour les différents types d'intégration. Dans l'idéal, les méthodes devraient être décrites sous une forme compatible avec l'architecture commune de la production statistique.

84. La plupart des sources de données à intégrer sont extérieures au service national de statistique. Ce dernier n'avait aucun contrôle sur la collecte des données et il faut s'attendre à des divergences en matière de concepts, de classifications, de populations et d'unités de collecte.

85. Lorsque des données sont intégrées, les concepts statistiques devraient être alignés sur les concepts appliqués aux nouvelles sources de données. Il est parfois indispensable d'élaborer de nouveaux concepts. Le service national de statistique doit comprendre les décisions conceptuelles pour pouvoir déterminer comment transformer des données extérieures en données statistiques. Ces différences influent sur les possibilités d'utilisation des sources de données extérieures dans la production de statistiques, en particulier en ce qui concerne : la couverture de la population, la validité des concepts visés, la disponibilité et l'exactitude des métadonnées descriptives, les erreurs d'échantillonnage, les distorsions, le fondement juridique des données, les méthodes de collecte de données et la conception des questionnaires, la charge de travail pour les répondants, la confidentialité des résultats, et les différentes conséquences pour les différents types de données fournis. Il faut clairement expliquer et étayer ces différences et les données correspondantes doivent être stockées pour garantir la réutilisation et l'amélioration des évaluations.

86. Les utilisateurs de données statistiques devraient être bien informés de la définition des concepts et des populations appliquée dans toutes les sources de données utilisées dans un produit statistique. Il faudrait également faire en sorte que le service national de statistique et les utilisateurs de données aient une position commune quant aux concepts statistiques.

87. Par exemple, lorsque des données administratives et des données d'enquête sont intégrées, les différences de concepts pourraient déboucher sur des problèmes de couverture et de distorsion. Dans certains cas, tels que les statistiques sur les entreprises, les unités utilisées dans les données administratives ne correspondent pas nécessairement à la définition des unités statistiques requises. Une certaine modélisation devrait être effectuée pour convertir les unités administratives en unités statistiques. Les définitions des variables divergeront probablement aussi. Il est important de connaître à fond les effets de ces divergences. Il est parfois possible d'influer sur la définition administrative en coopérant avec l'autorité responsable.

88. En présence de classifications différentes, on utilise habituellement des tableaux de correspondance et des outils de conversion fondés sur des variables supplémentaires peut-être disponibles pour la conversion à un code de classification plus correct. Cela étant, même des classifications identiques peuvent déboucher sur des données différentes, en particulier lorsqu'elles sont complexes ou que leurs règles sont difficiles à appliquer. S'agissant des sources administratives, le codage serait souvent effectué par le répondant, alors que dans une enquête par sondage il peut souvent exister des questions ouvertes et le codage est souvent effectué par des experts.

89. La coopération entre le service national de statistique et l'autorité administrative est un bon moyen de résoudre une partie des problèmes de classification. Le service national de statistique peut apporter son expérience et pourrait se charger de la gestion de la classification. Une autre question est celle de savoir s'il convient d'appliquer des classifications internationales directement converties ou des classifications nationales. Cela dépend du point de savoir quelles données nationales sont nécessaires. En cas de changement et de révision, la première solution est généralement plus difficile à mettre en œuvre que les classifications nationales. Il est plus contraignant de changer une classification en une source administrative car de nombreux fournisseurs de données peuvent devoir se familiariser avec les changements.

90. Lorsque l'on intègre des données géospatiales à d'autres sources de données, les niveaux de granularité ou d'agrégation des ensembles de données pourraient ne pas coïncider et la couverture pourrait être différente.

91. Les données manquantes et les erreurs peuvent aussi constituer des problèmes à surmonter. Dans les enquêtes statistiques, les données manquantes sont dues à l'absence de réponse au niveau d'une unité ou d'une variable, mais les causes peuvent être différentes s'agissant des sources administratives. Il est important de déterminer si les erreurs et les données manquantes sont systématiques et d'appliquer des règles de validation et d'édition adéquates.

92. Un problème courant avec les ensembles de données corrélés tient à la présence d'incohérences dans les enregistrements corrélés. Les ensembles de données corrélés devenant plus nombreux, les gains d'efficacité potentiels dans la détection et le traitement des incohérences augmentent avec la multiplication des variables. Cependant, cela peut aussi accroître le volume des contrôles requis.

93. Les sources éventuelles de distorsions devraient être recensées en ce qui concerne les ensembles de données intégrées. Il est possible que la question de la couverture et les questions conceptuelles ne s'appliquent qu'à certains groupes de population, de sorte que les résultats devraient être généralisés avec précaution. Certaines variables peuvent influer sur la qualité des corrélations et constituer une source de distorsions éventuelles lors de l'analyse des ensembles de données obtenus. Il peut être nécessaire d'étudier les taux de corrélation entre les différents groupes de population.

94. L'un des aspects les plus importants de l'intégration des données administratives est la normalisation des identificateurs (ou d'autres informations permettant de relier les données statistiques) entre différentes sources. En l'absence d'un tel système, il est beaucoup plus difficile d'établir un lien entre les différentes sources, et des méthodes de corrélation et d'appariement des données doivent être appliquées.

95. Il est indispensable de recourir à des méthodes permettant de mieux estimer les erreurs de corrélation afin de déterminer des modèles appropriés pour tenir compte de ces erreurs. Celles-ci contribuent à l'apparition d'éventuelles erreurs de couverture dans la population cible. Il faudrait aussi faire preuve de circonspection lorsque l'on crée des statistiques à partir d'ensembles de données intégrées dont l'un est extérieur puisque l'unité peut y être définie différemment.

96. Une extrême prudence est de mise pour la prévision et l'analyse rétrospective des données corrélées, en particulier pour les données longitudinales. Une personne peut faire l'objet d'une corrélation pour un trimestre mais pas pour un autre pour des raisons liées à la qualité des données (ou elle peut être corrélée à un enregistrement différent). L'utilisation d'ensembles de données appariés, même à des fins de validation, peut entraîner une rupture dans la série de données à gérer.

97. Les données provenant de sources extérieures peuvent souffrir d'erreurs de mesure, par exemple, en matière de validité, et ces erreurs se propagent lorsque les données sont intégrées à d'autres sources de données pour donner un produit statistique. De ce fait, les concepts cibles utilisés dans un ensemble de données provenant de sources extérieures doivent être bien compris avant d'être utilisés dans la production de statistiques officielles.

### 4. Qualité

98. La qualité des sources de données et des statistiques produites doit être mesurée, gérée et publiée. Les dimensions suivantes de la qualité doivent être évaluées : exactitude, pertinence, cohérence, accessibilité, comparabilité et actualité. Il est important de disposer de métadonnées descriptives détaillées afin de faciliter l'évaluation de la qualité des sources. L'intégration des données rend l'utilisation des métadonnées plus difficile. Le nombre minimal et idéal de métadonnées requises devrait être déterminé.

99. Un cadre de qualité serait nécessaire pour les données intégrées car cela permettrait de comprendre l'« incertitude » ou la fiabilité des estimations et de comprendre comment les indicateurs sont comparables dans le temps et avec des indicateurs analogues produits à partir de sources traditionnelles. Il est essentiel de disposer d'un cadre de qualité visant à déterminer comment concevoir de manière optimale une méthode permettant de combiner la ou les sources de données susceptibles de minimiser l'effet cumulatif des erreurs potentielles sur un produit statistique. On peut utiliser le projet de cadre de qualité concernant les mégadonnées établi au titre du projet relatif aux mégadonnées 2014 du Groupe de haut niveau, les cadres de qualité traditionnels pouvant servir à mettre au point un premier projet de cadre de qualité pour les statistiques produites à partir de sources de données intégrées.

100. D'autres problèmes tiennent à l'actualité des données et aux différences entre les périodes de référence. Les données administratives ou géospatiales peuvent ne pas être disponibles en temps voulu ou ne pas coïncider avec la période de référence statistique. On peut résoudre ce problème en analysant l'incidence et, si nécessaire, en l'ajustant grâce à des modèles.

### 5. Coopération internationale

101. Il est important d'exploiter l'expérience collective des milieux de la statistique officielle pour éviter le chevauchement des activités entre les pays et les organisations, et apprendre des expériences les uns des autres. L'échange d'expériences entre les organisations statistiques officielles contribuera à élaborer et à appliquer des approches communes en vue d'inclure plus efficacement et plus concrètement l'intégration des données dans les processus de production statistique.

102. Un cadre de collaboration internationale visant à résoudre les problèmes communs serait bienvenu. Le cas échéant, une coordination avec des groupes internationaux est nécessaire.

103. Il est envisageable de réunir certains fournisseurs de données et un groupe de services nationaux de statistique pour étudier les avantages mutuels et éventuellement conclure des accords pour la fourniture de données. Lorsque l'on utilise des ressources TI à l'extérieur du service de statistique, il est indispensable de mettre en place de bons partenariats avec les concepteurs et administrateurs de ces systèmes.

104. Étant donné qu'il faudrait dans la mesure du possible utiliser les outils d'orientation et les normes élaborés sous l'égide du Groupe de haut niveau (modèle générique du processus de production statistique (GSBPM), modèle générique d'informations statistiques (GSIM) et modèle générique d'activité des organismes statistiques (GAMSO) par exemple), des services conformes à l'architecture commune de la production statistique (CSPA) devraient être élaborés ou proposés. Il faut également identifier d'autres normes et cadres communs utilisés dans un environnement plus vaste et ayant trait aux tâches d'intégration des données.

105. L'absence de normes bien établies pour l'intégration de différentes sources de données offre l'occasion de proposer de nouvelles normes. La concordance entre les normes statistiques et autres est nécessaire. Par conséquent, il faut prendre en considération les normes utilisées par les secteurs sources (par exemple, éducation, voyages, services bancaires).

106. Il serait utile d'élaborer des approches communes pour certains domaines statistiques. Le bac à sable (sandbox) du Groupe de haut niveau et des ensembles de données d'échantillon peuvent être utilisés à cette fin. Les services nationaux de statistique de différents pays peuvent collaborer entre eux pour mettre au point une approche commune permettant d'obtenir des données auprès des sociétés multinationales.

# VI. Conclusions et recommandations

107. Par rapport aux seules approches classiques, l'intégration des données offre la possibilité de produire des statistiques plus d'actualité et plus désagrégées et ce, à des fréquences plus élevées. De telles activités ne feront donc que s'intensifier. Il deviendra de plus en plus nécessaire d'intégrer différentes sources de données à mesure qu'elles se multiplient et que les capacités des technologies de l'information (TI) et de l'infrastructure des données se renforcent.

108. Certains services nationaux ont mené des expériences pour certains types d'intégration. Très souvent, les expériences sont plus limitées et aucun service ne dispose de compétences pour tous les types d'intégration. Les directives sont inexistantes ou limitées dans ce domaine et un aperçu global des expériences fait défaut. La collaboration et l'échange de données d'expérience en matière d'intégration peuvent être des plus utiles pour la communauté statistique internationale.

109. Il est fortement recommandé de poursuivre le projet d'intégration de données du Groupe de haut niveau, de renforcer le partage de données d'expérience et d'intensifier la collaboration entre les services nationaux de statistique et d'autres organismes qui produisent des statistiques officielles ainsi que les fournisseurs de données. Le partage de données et la collaboration conduiront à des économies et à des synergies qui permettront d'améliorer la situation aux niveaux national et international.

110. Les différentes expériences d'intégration des données ont débouché sur l'élaboration de nombreuses recommandations, dont voici un résumé :

- Obtenir l'appui des hauts dirigeants et informer les responsables des services de statistique du projet ainsi que de ses objectifs et de ses résultats ;
- Maintenir les objectifs bien clairs, expliquer les objectifs et l'importance du projet et préciser les caractéristiques du problème à résoudre ;
- Établir une bonne collaboration avec les partenaires/fournisseurs de données : être clairs sur les besoins en données, prendre en compte les objectifs communs des institutions (ne collecter les données qu'une seule fois, réduire les dépenses inutiles), mettre en place des accords ;
- Collaborer avec les utilisateurs des données ;
- Consulter d'autres experts, en apprendre davantage sur les méthodes et solutions concrètes possibles ; procéder à des échanges de personnel pour acquérir de l'expérience et tirer des enseignements des bonnes pratiques ;
- Prendre en compte les recommandations et les normes internationales en matière de statistique ;
- Partager les résultats ;
- Commencer les travaux sur un échantillon, examiner les données avant de débuter, faire preuve de patience et de persévérance, garder à l'esprit une large gamme de solutions ; et
- Mesurer la qualité des données.

111. En ce qui concerne la dernière question concernant les cadres de qualité des données, des recommandations plus détaillées ont déjà été élaborées au titre du projet. Elles peuvent être consultées sur le site wiki du Groupe de haut niveau : http://www1.unece.org/stat/platform/x/axSzBw.