Nations Unies ECE/CES/2013/20

Conseil économique et social

Distr. générale
18 mars 2013
Français
Original: anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Soixante et unième séance plénière
Genève, 10-12 juin 2013
Point 5 de l'ordre du jour provisoire
Travaux du Groupe de haut niveau sur la modernisation de la production et des services statistiques

Utilisation des «données massives» dans les statistiques officielles

Note du secrétariat

Résumé

Lors d'un séminaire de haut niveau sur la simplification de la production et des services statistiques, qui s'est tenu à Saint-Pétersbourg du 3 au 5 octobre 2012, les participants ont demandé que soit établi un document expliquant les questions relatives à l'utilisation des données massives par les milieux de la statistique officielle. Ils souhaitaient un document à visée essentiellement stratégique, destiné aux chefs et aux principaux responsables des organismes de statistique.

Afin de répondre à cette demande, le Groupe de haut niveau sur la modernisation de la production et des services statistiques a constitué une équipe informelle[1] composée d'experts nationaux et internationaux, coordonnée par le secrétariat de la Commission économique pour l'Europe des Nations Unies. Le Groupe de haut niveau s'est félicité des observations formulées à propos de la teneur et des recommandations de ce document.

Ce document, ainsi que d'autres informations sur les travaux du Groupe de haut niveau, peuvent être consultés à l'adresse suivante: www1.unece.org/stat/platform/display/hlgbas.

[1] Les membres de cette équipe étaient les suivants: Michael Glasson (Australie), Julie Trepanier (Canada), Vincenzo Patruno (Italie), Piet Daas (Pays-Bas), Michail Skaliotis (Eurostat) et Anjum Khan (Commission économique pour l'Europe des Nations Unies).

I. Introduction

1. Dans notre monde moderne, de plus en plus de données sont générées sur le Web et produites par les détecteurs des dispositifs électroniques présents en nombre toujours croissant dans note environnement. La masse des données et la fréquence à laquelle elles sont produites ont conduit à la notion de «*données massives*». Il s'agit d'ensembles de données dont le volume, la vitesse et la diversité ne cessent de croître; souvent ces données «3 V» sont en grande partie non structurées, c'est-à-dire qu'elles ne correspondent pas à un modèle prédéterminé et/ou ne s'intègrent pas facilement dans les bases de données relationnelles classiques. Non seulement elles créent de nouvelles opportunités commerciales dans le secteur privé, mais elles peuvent aussi constituer un apport très intéressant pour les statistiques officielles, soit utilisées seules, soit associées à des sources de données traditionnelles comme les enquêtes par sondage et les registres administratifs. Toutefois, il n'est pas facile de recueillir l'information à partir des données massives pour l'intégrer dans un processus de production de statistiques. Le présent document s'efforce de répondre aux deux questions fondamentales suivantes qui concernent le *contenu* et les *modalités*:

a) À quel sous-ensemble de données massives les organismes nationaux de statistique (ONS) devraient-ils s'intéresser, compte tenu du rôle des statistiques officielles;

b) Comment les ONS peuvent-ils utiliser les données massives et résoudre les problèmes qu'elles posent?

2. La question du contenu est abordée dans un document sur les données massives au service du développement, publié par l'initiative Global Pulse des Nations Unies[2]:

> *Au niveau le plus général, les données massives correctement analysées, peuvent fournir des instantanés du bien-être des populations à une fréquence élevée, avec un degré élevé de granularité et depuis des angles très divers, en réduisant les délais et en comblant les lacunes dans les connaissances. En pratique, l'analyse de ces données peut aider à comprendre ce que Global Pulse a appelé les «signaux de fumée numériques» − c'est-à-dire des changements anormaux dans la manière dont les collectivités ont accès aux services, qui peuvent être considérés comme des indicateurs approximatifs des modifications du bien-être profond.*
>
> *Les statistiques officielles continueront de produire des informations pertinentes mais la révolution des données numériques («les actions qui peuvent être suivies ou stockées numériquement, les choix et les préférences manifestés par les populations dans leur vie quotidienne») offre une possibilité immense de parvenir à une connaissance plus riche et plus profonde de l'expérience humaine qui peut compléter l'élaboration des indicateurs déjà existants*.

3. Les données massives sont capables de produire des statistiques plus pertinentes et plus actuelles que les sources traditionnelles de la statistique officielle. Celle-ci s'appuie presque exclusivement sur les données recueillies par sondage et sur les données administratives fournies par les programmes publics, souvent une prérogative des ONS prévue par la législation. Tel n'est pas le cas des données massives qui sont pour la plupart aisément disponibles ou qui appartiennent à des entreprises privées. C'est pourquoi le secteur privé peut en tirer parti et produire de plus en plus de statistiques qui s'efforcent d'être plus actuelles et plus pertinentes que les statistiques officielles. Il est peu probable

[2] http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

que les ONS perdent leur statut «officiel» mais, s'ils n'évoluent pas, ils risquent de perdre peu à peu leur réputation et leur pertinence. Un grand avantage des ONS est qu'ils disposent d'infrastructures assurant l'exactitude, la cohérence et l'interprétabilité des statistiques produites. En incorporant des sources de données massives pertinentes dans leur processus d'établissement de statistiques officielles, les ONS sont les mieux à même de mesurer leur exactitude, d'assurer la cohérence des systèmes de statistique officielle et de les interpréter tout en améliorant constamment leur pertinence et leur actualité. Le rôle et l'importance des statistiques officielles seront ainsi protégés.

II. Définitions

4. Les statistiques officielles peuvent être définies sur la base des Principes fondamentaux de la statistique officielle. Principe 1[3]:

> *La statistique officielle constitue un élément indispensable du système d'information de toute société démocratique fournissant aux administrations publiques, au secteur économique et au public des données concernant la situation économique, démographique et sociale et la situation de l'environnement*.

5. Les statistiques officielles sont supposées dépeindre une situation et fournir la *description* d'un pays, de son économie, de sa population, etc. Lorsqu'on utilise les données massives comme source supplémentaire d'information, cette *description* doit être examinée.

6. Les données massives peuvent être définies comme une variante de la définition donnée par Gartner[4]:

> *Les données massives sont des sources de données qui peuvent être − en général − décrites comme: «une masse importante, une grande vitesse et une forte diversité des données qui exigent des formes novatrices rentables de traitement pour améliorer les connaissances et la prise de décisions.».*

III. Sources

7. Le phénomène des données massives nous fait prendre conscience de ce que nous vivons désormais dans un monde où les données sont omniprésentes. Ce fait ne peut être ignoré, d'où l'intérêt pour les statistiques officielles. Jusqu'ici, les ONS et les organisations internationales produisaient des données en utilisant deux voies différentes: les enquêtes par sondage et les sources administratives y compris les registres. La question à laquelle il faut répondre est la suivante: comment les données massives peuvent-elles aider à mesurer avec plus d'exactitude et de manière plus actuelle les phénomènes économiques, sociaux et environnementaux?

8. En règle générale, les sources de données massives peuvent être classées comme suit:

a) Sources administratives (provenant de l'administration d'un programme, qu'il soit public ou non), par exemple les dossiers médicaux électroniques, les consultations et séjours à l'hôpital, les dossiers d'assurance, les dossiers bancaires, les banques alimentaires, etc.;

[3] http://unstats.un.org/unsd/methods/statorg/FP-English.htm.
[4] http://www.gartner.com/it-glossary/big-data/.

b) Sources commerciales ou transactionnelles (provenant d'une transaction entre deux entités), par exemple les transactions au moyen de cartes de crédit, les transactions en ligne (y compris à partir de dispositifs mobiles), etc.;

c) Données fournies par des détecteurs, par exemple l'imagerie satellite, les capteurs routiers, les capteurs climatologiques, etc.;

d) Données des dispositifs de surveillance, par exemple des téléphones portables, du Système mondial de localisation (GPS), etc.;

e) Données comportementales, c'est-à-dire recherches en ligne (sur un produit, un service ou tout autre type d'information), système page vue en ligne, etc.;

f) Données indiquant des opinions, par exemple commentaires affichés sur les réseaux sociaux, etc.

9. Les données administratives sont l'une des principales sources utilisées par les ONS pour établir les statistiques. Elles sont recueillies à intervalles réguliers par les services de statistique et servent à produire les statistiques officielles. Habituellement, les données sont reçues, souvent des administrations publiques, traitées, stockées, gérées et utilisées par les ONS de manière très structurée. Peut-on considérer que les données administratives sont des «données massives» au sens de la définition mentionnée ci-dessus? Pour l'instant, la réponse serait sans doute négative. Les données administratives peuvent devenir «massives» si leur vitesse augmente, par exemple lorsqu'on utilise largement les données administratives qui sont recueillies tous les jours ou toutes les semaines au lieu du rythme annuel ou mensuel, habituel.

IV. Problèmes

10. L'utilisation de données massives dans la statistique officielle pose de nombreux problèmes qui sont examinés ci-dessous:

a) Problèmes législatifs, concernant l'accès aux données et leur utilisation;

b) Respect de la vie privée: obtenir la confiance du public et faire en sorte qu'il accepte la réutilisation des données ainsi que l'établissement de liens avec d'autres sources;

c) Problèmes financiers: coût potentiel par rapport aux avantages des sources de données;

d) Problèmes de gestion, concernant les politiques et les directives sur la gestion et la protection des données;

e) Problèmes méthodologiques, concernant la qualité des données et la pertinence des méthodes statistiques;

f) Problèmes technologiques, concernant la technologie de l'information.

A. Problèmes législatifs

11. Alors que dans certains pays comme le Canada, la législation peut autoriser l'accès aux données des administrations publiques et d'organismes non publics, dans d'autres pays comme l'Irlande, seul l'accès aux données provenant des autorités publiques est autorisé. Il peut en résulter des restrictions d'accès à certains types de données massives comme il est décrit au paragraphe 8.

12. Il est reconnu[5] que:

> *Souvent, le droit des ONS à l'accès aux données administratives, inscrit en principe dans la législation, n'est pas suffisamment appuyé par des obligations particulières imposées aux détenteurs des données*.

13. Même si la législation prévoit l'accès à tous les types de données, l'objectif statistique justifiant cet accès pourra devoir être plus ou moins explicité selon les pays.

B. Respect de la vie privée

14. Les définitions peuvent varier d'un pays à un autre mais le respect de la vie privée est en général défini comme le droit des individus à contrôler ou à déterminer les informations les concernant qui peuvent être divulguées. On peut établir un parallèle entre les entreprises qui souhaitent protéger leur compétitivité et les consommateurs. Le respect de la vie privée est un fondement de la démocratie. Le problème avec les données massives est que très probablement les utilisateurs des services et des dispositifs générant ces données n'ont pas connaissance de ce processus ni du but dans lequel les données peuvent être utilisées. Le volume des données peut devenir encore plus important si celles-ci sont rassemblées, ce qui aggrave les préoccupations concernant le respect de la vie privée.

C. Problèmes financiers

15. Il est probable que les ONS devront payer pour acquérir des données massives, en particulier si ces données sont détenues par le secteur privé et si la législation ne dit rien des modalités financières concernant l'acquisition de données extérieures. C'est pourquoi les ONS doivent faire les bons choix en pesant les avantages qualitatifs (pertinence, actualité, exactitude, cohérence, accessibilité et interprétabilité) par rapport aux coûts et à la diminution de la charge administrative. Les coûts peuvent être non négligeables pour les ONS mais les avantages potentiels sont de loin supérieurs, puisque les données massives recèlent des informations susceptibles d'augmenter l'efficacité des programmes publics (par exemple le système de santé). Il faut tenir compte aussi des règles concernant les marchés publics.

16. Le rapport établi par la Commission fédérale sur les données massives de la TechAmerica Foundation aux États-Unis[6] a conclu notamment que le succès de ces données tient à ce qu'elles permettent:

> *De connaître les impératifs et les exigences commerciales critiques d'un organisme donné, de décider qu'elles sont les questions à poser et de pratiquer l'art du possible, et de prendre les premières mesures pour un ensemble de cas d'utilisation bien définis.*

17. Cette approche peut certainement être transposée dans le contexte d'un ONS.

[5] http://essnet.admindata.eu/WorkPackage?objectId=4251 (p41).

[6] http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf.

D. Problèmes de gestion

18. Pour la statistique officielle, les données massives signifient des informations supplémentaires parvenant aux ONS qui doivent se conformer aux politiques et aux directives sur la gestion et la protection de l'information.

19. Un autre problème de gestion est celui qui concerne les ressources humaines. Pour l'instant, la science des données[7] associée aux données massives qui apparaissent dans le secteur privé ne semble pas intéresser les milieux de la statistique officielle. Les ONS pourraient devoir procéder à des recherches systématiques en interne et au niveau national (dans les milieux universitaires, et les secteurs public et privé) pour trouver des spécialistes de la science des données et entrer en contact avec eux.

E. Problèmes méthodologiques

20. La représentativité est un point fondamental dans le cas des données massives. Les difficultés rencontrées pour définir la population cible, la population sondée et le cadre de l'enquête remettent en question le mode de pensée traditionnel des statisticiens officiels et leur manière de procéder à une induction statistique sur la population cible (et finie). Dans le cas d'une enquête traditionnelle, les statisticiens identifient une population cible/soumise à l'enquête, construisent un cadre d'enquête pour atteindre cette population, choisissent un échantillon, recueillent les données, etc. Ils construisent une boîte et la remplissent de données de façon très structurée. Dans le cas des données massives, les données arrivent en premier et le réflexe des statisticiens officiels serait de construire une boîte! On se pose alors la question: *Est-ce le seul moyen d'obtenir un système national cohérent et intégré de statistique officielle?* Le temps est-il tenu de se dispenser de la boîte?

21. Une autre question tient essentiellement à la nature même de la technologie de l'information (TI) et de la méthodologie employée. Lorsqu'on analyse de plus en plus de données, les méthodes statistiques traditionnelles conçues pour une analyse très approfondie de petits échantillons ne sont plus satisfaisantes; dans le plus simple des cas, elles sont simplement trop lentes. Il devient donc nécessaire de disposer de méthodes et d'outils nouveaux:

a) Méthodes permettant d'extraire rapidement les informations des masses de données disponibles; par exemple les méthodes et données de visualisation, et les techniques d'extraction en continu qui sont capables de «réduire les données massives». Dans un premier temps, augmenter la puissance de l'ordinateur est un moyen de faciliter cette démarche;

b) Méthodes capables d'intégrer l'information mise au jour grâce au processus statistique, par exemple l'établissement massif de liens, l'intégration des macro/mésodonnées, et les méthodes statistiques spécialement adaptées aux grands ensembles de données. Il faut élaborer des méthodes qui produisent rapidement des résultats fiables lorsqu'elles sont appliquées à de très grandes masses de données.

[7] Wikipedia définit la science des données comme une science qui intègre divers éléments et s'appuie sur des techniques et des théories empruntées à de nombreuses disciplines, y compris les mathématiques, les statistiques, l'ingénierie des données, la reconnaissance des formes et l'apprentissage, l'informatique avancée, la visualisation, la modélisation de l'incertitude, le stockage des données et le calcul à haute performance, afin d'extraire le sens des données et de créer des produits de données.

22. Si l'on utilise les données massives pour la statistique officielle, il est indispensable de disposer de nouvelles techniques pour résoudre les problèmes méthodologiques suivants:

a) Mesures de la qualité des produits obtenus à partir de données extérieures difficiles à traiter. Le fait de dépendre de sources extérieures limite l'éventail des mesures applicables par rapport aux produits obtenus grâce aux techniques de collecte d'informations ciblées;

b) Application et valeur limitées des données provenant de sources extérieures;

c) Difficulté d'intégrer des informations provenant de diverses sources pour obtenir des produits à forte valeur;

d) Difficulté d'identifier une proposition de valeur en l'absence de rétroaction en boucle fermée dans les organisations commerciales.

F. Problèmes technologiques

23. Comme on l'a vu au paragraphe 9, accroître la vitesse de l'accès aux données administratives suppose aussi d'utiliser intensivement les interfaces de programmation standards d'applications (API) ou (parfois) des API en «streaming». Il est ainsi possible de connecter directement des données administratives à des applications pour la saisie et le traitement des données. La collecte de données en temps réel ou quasi réel augmente au maximum le potentiel des données, offrant de nouvelles possibilités pour combiner des données administratives et des données à grande vitesse provenant d'autres sources, telles que:

a) Les données commerciales (transactions avec les cartes de crédit, transactions en ligne, ventes, etc.);

b) Les dispositifs de suivi (téléphones portables, GPS, caméras de surveillance, systèmes automatisés d'achat et de paiement) et capteurs physiques (circulation, météorologie, pollution, énergie, etc.);

c) Les réseaux sociaux (Twitter, Facebook, etc.) et moteurs de recherche (recherche en ligne, page vue en ligne);

d) Les données des collectivités (appels de citoyens ou données résultant d'une externalisation ouverte).

24. En un temps où les données massives sont omniprésentes, ce changement de paradigme pour la collecte de données permet de recueillir et d'intégrer de nombreux types de données provenant de nombreuses sources différentes. Associer les sources de données pour produire de nouvelles informations est un défi supplémentaire intéressant à relever dans un avenir proche. Associer des sources de données «traditionnelles» telles que les enquêtes et les données administratives, à de nouvelles sources de données ainsi qu'aux nouvelles sources de données combinées, offre des possibilités pour décrire les comportements des communautés «ingénieuses». C'est encore un domaine inexploré qui peut ouvrir de nouvelles perspectives.

V. Comment les milieux de la statistique officielle peuvent-ils utiliser les données massives

25. Plusieurs exemples d'études sur les données massives en cours ou prévues sont examinées dans cette section. Les deux premières concernent les Pays-Bas.

26. **Statistiques de la circulation et des transports**: Aux Pays-Bas, on compte chaque jour environ 80 millions d'enregistrements des détecteurs de véhicules à boucle électromagnétique. Ces données peuvent être utilisées comme source d'information pour les statistiques de la circulation et des transports et peut-être aussi pour des statistiques concernant d'autres phénomènes économiques. Les données sont fournies à un niveau très détaillé. Plus précisément, pour plus de 10 000 détecteurs à boucle magnétique installés sur les routes des Pays-Bas, on dispose minute par minute du nombre de véhicules en mouvement de diverses longueurs. La longueur permet de distinguer entre voitures et camions. L'inconvénient de cette source tient à un grave défaut de couverture et de sélectivité. Le nombre de véhicules détectés n'est pas disponible pour chaque minute et certaines routes importantes ne sont pas encore équipées de boucles de détection. Au niveau le plus détaillé, correspondant à chaque boucle, le nombre de véhicules détectés témoigne d'une (forte) instabilité indiquant la nécessité d'une approche plus statistique. Recueillir le très grand nombre d'informations à partir des données constitue une difficulté majeure. Utiliser pleinement ces informations permettrait d'obtenir plus rapidement des statistiques plus fiables sur la circulation et des informations plus détaillées sur la circulation des grands véhicules qui indiquent une évolution du développement économique.

27. **Statistiques des réseaux sociaux**: Environ un million de messages sont affichés chaque jour sur les réseaux sociaux publics aux Pays-Bas. Ces messages sont accessibles à toute personne disposant d'un accès Internet. Les réseaux sociaux constituent une source de données potentielle puisque les membres échangent volontairement des informations, discutent des questions qui les intéressent et contactent leur famille et leurs amis. Afin de déterminer si les réseaux sociaux constituent une source de données intéressante pour les statistiques, Statistics Netherlands a étudié les messages du double point de vue de leur contenu et de l'appréciation subjective. Les études sur le contenu des messages de Twitter (le réseau social public le plus fréquenté aux Pays-Bas au moment de l'étude) a révélé que près de 50 % des messages étaient composés de «bavardages inutiles». Le reste concernait essentiellement les activités de loisirs (10 %), le travail (7 %), les médias (TV et radio 5 %) et la politique (3 %). L'utilisation de ces derniers messages, plus sérieux, était freinée par les messages de «bavardage», plus frivoles. Ces derniers affectaient aussi négativement l'exploration des textes. La détermination des appréciations subjectives dans les messages des réseaux sociaux a révélé des possibilités très intéressantes d'utilisation de cette source de données pour les statistiques. On a constaté que ces appréciations étaient étroitement corrélées à la confiance des consommateurs; en particulier s'agissant de la situation économique. Cette corrélation était stable sur une base mensuelle et hebdomadaire mais très instable pour ce qui était des chiffres quotidiens. Il est donc possible de produire des indicateurs hebdomadaires de la confiance des consommateurs et ceux-ci peuvent être établis le premier jour ouvrable suivant la semaine étudiée, ce qui démontre la possibilité de fournir des résultats rapides.

28. On trouvera ci-après une liste des études prévues au titre du programme de travail d'Eurostat qui comporte un certain nombre d'études de faisabilité visant à explorer le potentiel des données massives pour la statistique officielle.

29. **Statistiques des prix**: Utilisation et analyse des prix collectés sur l'Internet. C'est un projet de vingt-quatre mois débutant en janvier 2013 qui mettra au point un logiciel libre très élaboré de capture de données («scraping») des sites Web pour aider les spécialistes de l'indice des prix à la consommation (IPC) à procéder à la collecte automatisée des prix sur l'Internet. Ce logiciel présente des similitudes avec le Billion Prices Project du Massachusetts Institute of Technology (MIT) et sera testé dans cinq pays pour résoudre des problèmes techniques et méthodologiques. Il sera communiqué «en l'état» sous licence publique de l'Union européenne (EUPL) aux autres organismes statistiques qui en feront la demande.

30. **Statistiques du tourisme**: Étude de faisabilité sur l'utilisation des données de localisation mobile pour les statistiques du tourisme. Un projet de quinze mois devrait commencer en janvier 2013. Il évaluera l'intérêt de l'utilisation des données de localisation mobile pour les statistiques du tourisme (et de domaines apparentés) et en déterminera les avantages et les inconvénients. Les sujets étudiés comprennent: l'accès (et la continuité d'accès), la confiance (des producteurs et des utilisateurs des statistiques), les coûts, les concepts (transformation du concept actuel des statistiques du tourisme en une nouvelle source de données) et d'autres questions de méthodologie (par exemple la représentativité, le choix d'échantillons dans une grande masse d'observations). L'aptitude à traiter les dossiers importants des opérateurs mobiles est considérée comme une condition critique qui devra être remplie pour que le projet réussisse. Ce projet a été inclus dans le programme de travail en raison notamment des résultats prometteurs de la recherche dans un certain nombre de pays.

31. **Statistiques sur l'utilisation des technologies de l'information et des communications (TIC)**: Étude de faisabilité sur l'exploitation des flux de trafic Internet pour recueillir des statistiques sur la société de l'information. Avec ce projet, Eurostat souhaite tester et évaluer la faisabilité des méthodes de mesure «axées sur l'usager» et «axées sur le Web», dans une perspective pluridimensionnelle associant les questions techniques, méthodologiques, juridiques et sociopolitiques ainsi que les questions de coût. Parmi les autres produits attendus de ce projet, qui peut être intéressant pour les organismes de statistiques nationaux et internationaux, il convient de mentionner trois rapports sur les sujets suivants: i) comment élaborer une procédure d'accréditation; ii) manuel de méthodologie et de mise en œuvre du processus à l'intention des ONS; et iii) expérimentation du concept de «données libres fédérées». Ce concept est l'équivalent (ou le supplément) des données dites «libres publiques». Il fait référence à un sous-ensemble partagé de données massives communiquées par les entités du secteur privé, qui sera «ouvert» pour utilisation par les ONS.

32. Alors que les ONS ne sont qu'au début de l'exploration des potentialités des données massives aux fins de la statistique officielle, les premiers éléments suggèrent que l'expérimentation devrait porter sur trois grands domaines:

a) Association des données massives et des statistiques officielles;

b) Remplacement des statistiques officielles par les données massives;

c) Combler de nouvelles lacunes en matière de données, c'est-à-dire mettre au point de nouvelles mesures «fondées sur les données massives» pour étudier des phénomènes émergents (qui ne sont pas connus à l'avance ou auxquels les approches traditionnelles ne peuvent s'appliquer).

33. Les possibilités de combinaison des données massives avec les statistiques officielles présentent certaines analogies avec ce qui s'est fait au cours des dernières décennies lorsque les données administratives ont été associées aux statistiques officielles. Toutefois, ce qui sera probablement un peu différent, et potentiellement intéressant, est la possibilité de recourir plus largement à la modélisation statistique pour combiner les deux types de données. On pourrait ainsi obtenir des estimations qui, tout en conservant l'excellente qualité des statistiques officielles, les renforceraient grâce aux mesures en temps quasi réel que permettent les données massives.

VI. Conclusions/recommandations

34. Cette section examine les conclusions à tirer et propose des recommandations pour les travaux futurs. **Il est clair que, au cours des deux années qui viennent, il faudra identifier un petit nombre de projets pilotes qui serviront à valider le concept** (analogues à ceux qui ont été décrits à la section 5) avec la participation de quelques pays collaborateurs. Les résultats pourraient être présentés au Groupe de haut niveau.

35. Les données passives représentent pour les organismes nationaux et internationaux de statistique un certain nombre de défis et de responsabilités essentiels qui concernent principalement la méthodologie, la technologie, la gestion, les compétences ainsi que les questions juridiques. **Les organismes de statistique sont donc encouragés à inclure formellement les questions relatives aux données passives dans leurs programmes de travail annuels et pluriannuels en mettant en route des projets de recherche et des projets pilotes dans certains domaines et en affectant les ressources appropriées à cet effet.**

36. Utiliser d'énormes quantités de données n'est pas une tâche facile. En raison de leur seul volume, obtenir des informations à partir des données passives tout en assurant leur qualité peut s'avérer difficile. La phase d'exploration des données prendrait considérablement plus de temps pour les données passives que pour d'autres sources, souvent plus structurées, de données très nombreuses. Il faut donc disposer de «nouvelles» méthodes d'exploration et d'analyse. Le mot «nouvelles» est placé ici entre guillemets car de nombreuses méthodes sont déjà utilisées mais ce qui est nouveau c'est leur application à la statistique officielle. Trois d'entre elles se sont révélées particulièrement utiles, à savoir: les méthodes de visualisation, l'exploration des textes et le calcul à haute performance.

37. Les ONS sont relativement peu nombreux à étudier les aspects technologiques des données passives et c'est principalement le secteur privé qui mène les travaux sur les outils et les solutions analytiques dans ce domaine. Adapter les outils et les systèmes analytiques des données passives aux statistiques officielles exigera inévitablement la participation des ONS. Les cas d'utilisation réussis devraient être portés à l'attention des milieux internationaux de la statistique.

38. Les synergies entre les ONS et le secteur privé ne se limitent pas aux questions technologiques. **La collaboration des ONS avec les propriétaires de sources de données du secteur privé revêt une importance capitale et concerne des questions sensibles telles que le respect de la vie privée, la confiance et la compétitivité des entreprises, ainsi que le cadre législatif des ONS. Des exemples nationaux dans ce domaine, traitant certaines des questions relatives à l'octroi aux ONS d'un accès privilégié à des sources privées de données passives devraient faire partie des initiatives prioritaires.**

39. Pour utiliser les données passives, les statisticiens doivent modifier leur état d'esprit et acquérir de nouvelles compétences. Le traitement d'un nombre toujours croissant de données pour les statistiques officielles doit être confié à des personnes conscientes des problèmes statistiques, dotées d'un sens de l'analyse, s'intéressant aux techniques de l'information (par exemple possédant des compétences en matière de programmation) et déterminées à extraire des «connaissances» utiles des données. Ces «spécialistes de la science des données» peuvent venir de diverses disciplines scientifiques. Aux Pays-Bas, des travaux intéressants ont été faits par des chercheurs titulaires d'un doctorat en mathématiques, en physique, en (bio) chimie et en économie/économétrie qui s'intéressaient aux technologies de l'information.

40. À long terme, les statisticiens des données massives et les «spécialistes de la science des données» pourront acquérir leurs compétences grâce à l'adaptation des programmes universitaires (certaines universités offrent déjà des cours à cet effet), mais à moyen et à

court terme, **les ONS devraient proposer une formation spécialisée pour développer en interne les capacités analytiques nécessaires. Une collaboration internationale à cet égard serait très bénéfique pour les milieux de la statistique officielle.**

41. **Le Groupe de haut niveau devrait aussi envisager, à moyen terme, l'élaboration de directives/principes pour l'utilisation efficace des données massives aux fins des statistiques officielles. Il devrait réfléchir à un nouveau rôle que pourraient jouer les ONS à l'avenir,** c'est-à-dire en termes d'étiquetage et de certification des statistiques dérivées de données massives et utilisées pour définir les politiques publiques. Avec la présence des données massives dans de nombreux domaines de la vie quotidienne, il devient de plus en plus important de créer un tiers de confiance; appartient-il aux ONS d'assumer cette responsabilité seuls ou doivent-ils être membres d'une autorité pluridisciplinaire indépendante?

42. La statistique officielle a pris un certain nombre d'initiatives à propos des données massives. On peut citer par exemple une proposition faite à la quarante-quatrième session de la Commission de statistique des Nations Unies visant à organiser les manifestations suivantes: une conférence mondiale à l'automne 2013, des sessions au Congrès mondial de la statistique de l'Institut international de statistique en août 2013, un débat prévu à la Réunion des directeurs généraux des instituts nationaux de statistique, à l'automne 2013 pour les chefs des organisations formant le Système européen de statistique, et un projet d'atelier Eurostat/CEE sur les applications pratiques des données passives, à la fin de 2013 ou au début de 2014. Plusieurs manifestations devraient aussi être organisées sur l'utilisation et l'intégration des données géospatiales (qui sont souvent considérées comme des données massives) dans les statistiques officielles. **Le Groupe de haut niveau devrait s'assurer que les produits de ces activités et d'autres activités analogues sont effectivement coordonnés et communiqués au niveau stratégique.**