Nations Unies ECE/CES/2019/26

Conseil économique et social

Distr. générale
12 avril 2019
Français
Original : anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Soixante-septième réunion plénière
Paris, 26-28 juin 2019
Point 6 a) de l'ordre du jour provisoire
Le nouveau rôle des organismes nationaux de statistique dans le domaine de la gestion de données
Première séance : nouvelles possibilités et problématiques relatives au système de données

Même rôle mais changement de fonction ? Conséquences de la donnéification pour les statistiques officielles

Note d'Eurostat

Résumé

Le présent document traite de l'évolution des conditions d'accès aux données, de leur traitement et de leur diffusion, ainsi que de leur incidence sur les fonctions du système statistique. Les auteurs partent du principe que la mission des statistiques officielles reste inchangée, mais que les moyens et les conditions d'exécution évolueront. Pour cela, il faudra suivre une approche systémique selon laquelle les organismes de statistique devront repenser leur modèle de production et de diffusion.

Les auteurs présument que les organismes de statistique ne pourront pas centraliser l'ensemble des données et de la production statistique. Les tâches de ces organismes pourraient donc être renforcées et étendues à la fourniture d'orientations sur l'utilisation, le traitement et la communication des statistiques. Les organismes de statistique pourraient jouer un rôle de premier plan au sein des administrations publiques en tant que dépositaires des données utilisées à des fins statistiques. La présente note porte aussi sur l'idée de certifier les données, les processus et les produits, ainsi que sur les conséquences pour les organismes de statistique. La demande croissante de statistiques de qualité qui répondent à des besoins précis et qui sont diffusées juste à temps exige anticipation et souplesse pour éclairer les différents groupes de la société dans les débats publics.

Le présent document est présenté, pour examen, au séminaire de la Conférence des statisticiens européens de 2019 intitulé « Le nouveau rôle des organismes nationaux de statistique dans le domaine de la gestion de données », la première séance étant consacrée au thème « Nouvelles possibilités et problématiques relatives au système de données ».

GE.19-06167 (F) 160419 070519
1906167

I. Introduction

1. L'évolution technique des technologies de l'information et de la communication a vu naître un processus mondial de numérisation global de la société tout entière depuis l'avènement d'Internet intervenu au cours de la dernière décennie du siècle dernier. Les principaux facteurs qui ont facilité le processus ont été la propagation d'Internet et du World Wide Web, qui sont devenus une plateforme générale pour les informations et les applications. Avec le développement des téléphones mobiles, l'accès à Internet ne s'est plus limité aux ordinateurs personnels disposant d'une connexion câblée, mais a permis de recourir à des services en ligne en tout lieu, à tout moment et pour tout individu.

2. Grâce à la miniaturisation et à l'intégration de plus d'informations, ces dispositifs sont devenus des appareils polyvalents qui accompagnent les gens 24 heures sur 24 et 7 jours sur 7. En parallèle ont été élaborés des services et des plateformes, qui ont remplacé les activités ou objets physiques par leurs équivalents numériques. Avec la constitution de réseaux est apparue une tendance à établir des monopoles pour des applications particulières telles que la recherche d'informations et l'utilisation des réseaux. Depuis peu, de plus en plus de machines et d'appareils sont équipés de technologies de l'information (TI), le but étant de les rendre plus intelligents. Ces « appareils intelligents » sont capables de communiquer entre eux, créant ainsi ce que l'on appelle l'Internet des objets. L'intelligence artificielle est de plus en plus intégrée dans ces dispositifs ou dans des systèmes créés par le réseau de dispositifs pour les rendre plus intelligents. Cette évolution a pour objectif final de mettre en place un service pour les utilisateurs des appareils ou systèmes intelligents au lieu d'offrir un produit. Par exemple, les entreprises offrent la mobilité au lieu de vendre une voiture.

II. Un nouveau paradigme : la donnéification

3. La donnéification, qui transforme chaque activité et état en données, passe par la numérisation. Ces données sont recueillies, stockées, traitées, analysées et échangées. Leurs échelles, en matière de volume, de dimensionnalité, de fréquence, de densité et de diversité, sont de plusieurs ordres de grandeur supérieures à celles des collectes de données antérieures à l'ère numérique. De fait, les collectes de données coûteuses du passé se sont transformées en données en tant que sous-produits de la donnéification. La collecte de données et la fourniture d'informations ne sont plus un quasi-monopole de structures particulières telles que les organismes de statistique, mais peuvent être effectuées par tout organisme doté des compétences nécessaires pour transformer les données en informations.

III. Évolution des habitudes et des attentes des utilisateurs

4. En plus de la production de données, les plateformes Internet collectent des données dans le cadre de leur modèle économique. En échange de la prestation de services ou d'autres incitations, les utilisateurs communiquent des informations personnelles ou non personnelles, qui sont de nouveau utilisées pour mettre au point de nouveaux services, mais servent également à engendrer des revenus, par exemple au moyen de la publicité, pour fournir gratuitement des services de plateforme aux citoyens. Outre les services directs aux utilisateurs, les plateformes et les applications Internet ont recours à des moyens supplémentaires pour inciter les usagers à utiliser un service. Cela pourrait être la ludification, qui offre des éléments ludiques types tels que le comptage de points ou la compétition avec d'autres utilisateurs. Comme autre exemple, on peut citer les systèmes de récompense communautaires, qui encouragent la création de communautés en vue de la réalisation d'un objectif commun et récompensent les individus et les communautés dans leur ensemble. On pourrait aussi par exemple encourager les activités d'économie d'énergie dans un quartier, qui pourraient en outre être associées à des approches ludiques. La reconnaissance publique pourrait être une autre raison de s'engager dans un service numérique et de fournir des données.

5. Des incitations financières ou pseudo-financières sont également employées pour recevoir des données des citoyens. Un exemple d'incitation pseudo-financière est la carte de fidélité dont les points peuvent être utilisés pour acheter des biens. Il suffit parfois de promettre la possibilité de remporter des prix pour convaincre les citoyens de fournir des données personnelles. Les citoyens sont à la fois des consommateurs de services et des producteurs de données. Ils prennent de plus en plus conscience de ces mécanismes et échangent des données contre des services personnels ou d'autres incitations. Le comportement des citoyens placés dans des situations similaires et dans des contextes différents, comme la fourniture de données aux organismes de statistique ou pour le bien public, changera en raison des mécanismes décrits, qu'ils expérimentent au quotidien.

6. La personnalisation est un facteur important lorsqu'il s'agit de fournir des services et d'encourager les citoyens à communiquer leurs données personnelles. Si l'on prend comme exemple la communication de données par les appareils portables de suivi de l'état de santé à la plateforme de vente, la collecte et la fourniture de données se font essentiellement en arrière-plan grâce au transfert automatique des données. En contrepartie, les utilisateurs reçoivent des informations sur leur état de santé et des conseils personnalisés en matière de comportement. Ils peuvent aussi se comparer à d'autres utilisateurs ou ont la possibilité de créer des groupes, par exemple, pour faire des exercices ensemble.

7. La grande disponibilité des informations et des services a changé les attentes des utilisateurs. Les informations et les services devraient être accessibles juste à temps, 24 heures sur 24 et 7 jours sur 7. Grâce aux moteurs de recherche et aux encyclopédies en ligne, les informations sont accessibles au moment même de la demande. Il est possible d'acheter tout produit à tout moment de la journée, la livraison se faisant peu après le paiement. Le volume des informations est adapté à la demande. Leur fourniture est plus ou moins immédiate.

8. Le traitement automatique des données offre la possibilité d'obtenir des informations quasiment en temps réel. Les données recueillies peuvent être traitées, analysées et agrégées en très peu de temps. Les délais entre la collecte de données et la fourniture d'informations peuvent être considérablement réduits ; dans certains cas, des informations peuvent être offertes en temps réel. Par exemple, dans un réseau énergétique intelligent, la production et la consommation d'énergie peuvent être suivies en temps réel ; les données sur le trafic sont recueillies et traitées de manière à fournir en temps réel des informations sur les embouteillages et à formuler des recommandations pour optimiser la circulation.

9. Sur Internet, il n'existe aucune autorité qui certifie l'exactitude des informations ou la qualité d'un bien ou d'un service. Pour instaurer la confiance, des mécanismes ascendants ont été créés. Selon les caractéristiques d'Internet en tant que réseau, les prestataires de services s'efforcent de renforcer la fiabilité des informations en établissant des liens avec d'autres nœuds du réseau. Le nombre de liens vers une source d'information renforce la pertinence d'un résultat trouvé en ligne. Des plateformes collectent des informations sur les évaluations de produits faites par les personnes qui ont acheté un produit. Certaines plateformes ont encore perfectionné ce concept pour confronter les évaluations des vendeurs et des consommateurs de services en vue de créer un « Web de la confiance ». Dans ces cas, la réputation devient un indicateur de la fiabilité. Les clients sont encouragés à évaluer la qualité d'un bien ou d'un service qu'ils ont consommé. La qualité est souvent évaluée selon une approche spontanée et non systématique pour ce qui est des critères appliqués par une personne pour évaluer un produit.

IV. Mission des statistiques officielles

10. Les organismes de statistique fournissent à la société des données statistiques de qualité concernant la situation et les tendances de la société, de l'économie et de l'environnement. Pour s'acquitter de cette mission, le système statistique a élaboré un environnement ancré dans des cadres juridiques.

11. En outre, la fiabilité des données statistiques repose sur des codes de déontologie tels que le Code de bonnes pratiques de la statistique européenne et sur des cadres de

qualité qui permettent de recueillir systématiquement des informations (métadonnées) pour décrire le processus et le produit statistique. Des contrôles permettent de vérifier si les principes énoncés dans le code de déontologie sont respectés. Dans le cas du système statistique européen (SSE), un mécanisme de sécurité informatique du système est responsable de l'assurance et de la certification concernant la sécurité informatique et des examens par les pairs sont effectués pour établir la conformité avec le Code de bonnes pratiques de la statistique européenne. Comme les données constituaient un produit rare et cher, les principaux efforts ont porté sur la conception optimale de systèmes de collecte de données. Les cadres de qualité définissent et décrivent les critères, les procédures et les indicateurs destinés à la collecte de données et à l'évaluation de la qualité des produits statistiques. En conclusion, le système statistique a été conçu selon une approche descendante allant de la mission aux produits statistiques.

12. Cette mission générale des statistiques officielles est toujours d'actualité, mais les moyens d'exécution devront évoluer, en partie en raison des changements décrits ci-dessus, qui sont intervenus dans la collecte et le traitement de données et dans la recherche d'informations.

V. Conséquences pour les statistiques officielles

13. Les statistiques officielles ont été critiquées pour avoir perdu leur capacité à représenter le monde avec exactitude[1]. Ce phénomène pourrait être lié à un changement d'habitudes en matière de recherche d'informations et à une modification des attentes des citoyens en ce qui concerne la consommation d'informations, ceci étant le résultat de l'avènement d'un Internet omniprésent. Un changement d'habitudes s'accompagne d'une modification des attentes. Les utilisateurs exigent que les possibilités qui leur sont offertes dans un domaine soient transférées à d'autres domaines de la recherche d'informations et de la prestation de services. Les attentes changent également en raison de la demande de décisions axées sur des données, des preuves fondées sur des informations étant demandées avant la prise de décisions. Cette demande est présente dans le secteur économique, mais existe aussi dans le secteur public, en particulier pour les débats politiques au sein de la société. La prise de décisions fondée sur des données ou l'élaboration de politiques reposant sur la connaissance des faits aura des répercussions considérables sur les statistiques, étant donné que la demande de données fiables et de qualité augmentera lorsqu'elle sera prise au sérieux.

14. La disponibilité de nouvelles sources de données et l'évolution des habitudes en matière d'utilisation des informations ont une incidence sur la production de statistiques ainsi que sur la façon dont les statistiques sont consommées.

A. Apport de données

15. En général, les nouvelles données ne sont pas collectées aux fins des statistiques officielles. Il faut donc redéfinir leur objet pour produire des statistiques. Leur granularité est souvent différente de celle des microdonnées traitées dans le cadre des statistiques officielles. Ainsi, ces données ne comportent pas de renseignements sur l'état de santé d'un individu, mais sur sa pression artérielle ou son rythme cardiaque, grâce aux appareils portables de suivi de l'état de santé. De plus, la fréquence temporelle est généralement bien plus élevée que pour les microdonnées détenues par les organismes de statistique. Les données doivent être analysées et converties en informations qui ont un sens et sont adaptées aux statistiques officielles. À la différence des microdonnées utilisées dans les statistiques officielles, nous appelons ces données des données profondes (deep data) ou des nanodonnées.

16. Leurs volumes sont beaucoup plus élevés que ceux des données traitées dans les organismes de statistique, dont les capacités de traitement pourraient être dépassées si l'on

[1] William Davies (2017) : How statistics lost their power ; The Guardian ; https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy

utilisait la totalité ou une partie seulement de ces données. Il faudrait appliquer un traitement préalable aux nanodonnées pour les convertir en microdonnées afin de les rendre plus gérables et de les intégrer dans le processus de production statistique. Transférer le processus de calcul aux détenteurs de données comporte un autre aspect, celui de la sécurité des données. Centraliser toutes les informations possibles au même endroit augmente considérablement le risque d'atteinte à la sécurité des données étant donné que le lieu de stockage central sera très alléchant pour les attaques sécuritaires. Pour répartir les risques, il faudrait conserver les données sur leur lieu de production et les moyens techniques de traitement, d'agrégation et de corrélation devraient comporter des techniques de protection de la vie privée jusqu'au point où les produits ne seraient plus personnels mais seraient des agrégats.

17. En matière de production de données, ces principes mettent l'accent non plus sur la conception de la collecte de données, mais sur le traitement des données et privilégient, par rapport à l'accès aux données, la recherche d'informations et de statistiques finales ou intermédiaires, ultime objectif des organismes de statistique. L'enjeu ne serait plus de partager des données, mais de partager des algorithmes et des méthodes pour produire les statistiques (finales ou intermédiaires). Cette approche nécessiterait des partenariats étroits entre les détenteurs de données et les organismes de statistique, ainsi que le recours à des méthodes normalisées ou convenues pour traiter les données à la source.

18. En outre, il faudrait appliquer des mécanismes visant à assurer une qualité définie des sources et procéder à un examen minutieux des algorithmes de traitement, de préférence en tant que logiciels libres. Il faudrait aussi s'assurer que les programmes convenus soient appliqués aux données source convenues. Dans l'idéal, les résultats devraient être utilisables non seulement par les organismes de statistique mais aussi par les détenteurs de données en vue de créer une situation avantageuse pour tous. La création d'une telle situation faciliterait considérablement la collaboration entre les détenteurs de données et les organismes de statistique. Cela étant, des incitations supplémentaires pourraient exister, mis à part les mesures visant à favoriser les services de données ou d'information.

19. Il sera nécessaire d'avoir largement recours aux mécanismes d'incitation pour promouvoir la recherche de données provenant de sources privées ou de particuliers. Ces incitations peuvent être la ludification, la participation des communautés, la reconnaissance publique, les services personnalisés ou l'attribution de récompenses monétaires ou pseudo-monétaires aux individus. Pour les entreprises, il pourrait s'agir d'une contribution au bien public, de l'adaptation des informations aux entreprises, de la promotion de services d'information, le bureau de statistique intervenant en tant que tierce partie, ce qui exige l'intégration de données provenant des concurrents, par exemple, des informations sur le nombre d'entreprises subissant des attaques de pirates informatiques sans révéler laquelle est effectivement atteinte. Les données d'itinérance pourraient aussi être regroupées pour produire des produits statistiques communs. Un moyen de stimuler les marchés de données pourrait consister à introduire des modèles de « freemium » (stratégie consistant à attirer un client avec une offre gratuite dans l'espoir de lui faire acheter un service payant).

20. Dans le contexte des données d'entreprises, il est important de rappeler que les nouvelles données ne sont en général pas des données sur les caractéristiques particulières d'une entreprise, mais qu'elles portent sur des tierces parties. En conséquence, la question de la charge de travail que représente la communication de données devrait être interprétée d'une nouvelle façon. Si des mesures d'incitation sont utilisées pour la collecte de données, la fourniture de données pourrait ne pas être perçue comme une charge de travail supplémentaire.

B. Traitement des données au sein des organismes de statistique

21. Nous partons du principe que les organismes de statistique reçoivent des données intermédiaires par des circuits de calcul qui assurent le respect de la vie privée, ces données pouvant faire l'objet d'un traitement plus poussé pour produire les statistiques finales.

Ces nouvelles données peuvent être souvent utilisées à différentes fins, c'est-à-dire la production de statistiques dans différents domaines. Ainsi, les données recueillies par les réseaux mobiles peuvent être utilisées pour améliorer les statistiques sur la population, les migrations, le tourisme ou la balance des paiements. Le processus doit donc être souple pour assurer des utilisations multiples. Dans le même temps, un domaine statistique peut tirer profit de différentes données d'entrée. Des statistiques sur le tourisme peuvent par exemple être établies à partir d'enquêtes, de données recueillies sur les réseaux mobiles, de données fournies par les compteurs du trafic routier, etc. L'architecture de l'infrastructure de production doit pouvoir prendre en charge la réception d'entrées multiples pour produire des résultats multiples.

22. De plus, il est fort probable que ces nouvelles sources de données ne pourront pas totalement remplacer la collecte traditionnelle de données, mais, ensemble, les nombreuses sources de données permettront d'améliorer divers aspects du produit final comme la granularité temporelle ou spatiale.

23. Avec cette nouvelle architecture de production, il sera davantage nécessaire d'assurer une coordination entre les différents domaines statistiques au sein des organismes de statistique. Les changements intervenus dans le processus de production pourraient avoir des répercussions sur de nombreux produits. Les possibilités offertes par les nouvelles sources de données pourraient être bénéfiques pour divers domaines statistiques.

24. En raison des caractéristiques techniques des nouvelles sources de données, il sera peut-être encore plus nécessaire pour les organismes de statistique de disposer de compétences supplémentaires offertes par le secteur de l'ingénierie. Afin de négocier et de concevoir efficacement le transfert du processus de calcul, il faudrait employer un personnel ayant des connaissances spécialisées dans différents domaines des nouvelles sources de données, par exemple, des spécialistes de la communication, des transports et de l'électronique ou des ingénieurs.

C. Production de données

25. Nous avons précédemment décrit l'évolution des habitudes des utilisateurs lorsqu'il s'agit de rechercher des informations sur Internet. Les organismes de statistique doivent accorder plus d'attention aux moyens de communiquer des informations statistiques à différents groupes d'utilisateurs, dans l'idéal en leur fournissant des informations aisément compréhensibles, dans les quantités adéquates et au moment même de la demande.

26. Les utilisateurs s'attendent à utiliser divers modes d'interaction pour recevoir les informations demandées. De plus en plus, les gens utilisent des assistants numériques qui interagissent avec eux en langage naturel. Cela pourrait les amener à s'attendre à ce que la reconnaissance de la parole leur permette de rechercher des informations.

27. En général, les attentes augmentent quant à la granularité des statistiques. Les moyennes et les montants agrégés sur de grands territoires pourraient brosser un tableau plus large de l'ensemble de la société, mais pourraient ne pas représenter la réalité ou la perception de la réalité de certains groupes sociétaux dans de plus petites régions. Les gens demandent de plus en plus que les statistiques reflètent mieux la réalité de groupes sociétaux plus petits ou de parties plus petites du territoire.

28. Les données statistiques devraient être immédiatement disponibles dans des formats et des quantités adaptés aux besoins.

29. Du fait que les renseignements sont immédiatement disponibles et en raison de l'apparition de nouveaux systèmes en temps réel, la perception des utilisateurs évoluera quant à la question de savoir quand des informations sont fournies en temps voulu. Les gens pourraient ne plus admettre l'idée que des renseignements datant de plus d'un an sont des renseignements récents.

VI. Production de statistiques officielles par des tierces parties

30. La demande de statistiques est en augmentation, compte tenu de l'offre plus large de données produites par voie numérique et du paradigme des solutions fondées sur des données. Il est très probable que les nouvelles sources de données ne pourront pas remplacer complètement les collectes de données existantes. Partant, la démarche consistant à intégrer de nouvelles données dans le processus de production de statistiques officielles risque peu d'entraîner une réduction des ressources mais permettra, dans le meilleur des cas, de faire plus avec des ressources comparables.

31. Il faudra maintenir un mécanisme de définition des priorités pour décider des investissements à faire dans des domaines particuliers. En outre, l'objet des travaux ne sera plus la collecte de données, mais la conception du processus de diffusion d'informations à partir de données brutes. Nous partons du principe que ce processus consommera à l'avenir la plupart des ressources de la production statistique. Il s'agira par exemple de la conception de méthodes, d'algorithmes, d'infrastructures TIC, d'un cadre méthodologique, de critères de qualité et de métadonnées. Ce sont des éléments qui seraient en outre utilisés pour documenter le processus de production de données et la définition de la qualité des produits finals, à supposer que des tierces parties soient responsables du processus de production entier ou des étapes intermédiaires. Ces éléments pourraient aussi être appropriés pour la mise en place d'un processus de certification dans lequel les processus ou les produits sont certifiés comme étant conformes à des cadres prédéfinis.

32. Un processus type d'élaboration de la certification pourrait comprendre la mise au point de cadres pertinents, les spécialistes techniques et les statisticiens étant les mieux placés pour le faire. Les spécialistes techniques tels que les ingénieurs fourniraient, sur la source de données, les connaissances nécessaires pour convertir les données source dans des formats intermédiaires susceptibles de servir de base à des applications statistiques. Les cadres obtenus devraient servir à étayer le traitement des données par des documents et à décrire les caractéristiques des produits en résultant. Selon le niveau de qualité souhaité, certains éléments pourraient être choisis pour la documentation. Une tierce partie, qui pourrait être différente d'un bureau de statistique, pourrait certifier l'exactitude des documents. L'adéquation des données ainsi recueillies pour des applications statistiques pourrait être évaluée en fonction de la documentation.

33. Vu que la plupart des ressources seraient consommées dans le processus de conception méthodologique et que les organismes de statistique réaliseraient les activités connexes, la certification ne serait pas adéquate pour élargir considérablement l'offre globale d'informations statistiques sans accroître les ressources des organismes de statistique. Il serait également nécessaire de reproduire le processus de certification car les nouvelles sources de données ont tendance à être instables en raison de l'évolution technologique ou de changements intervenus dans le comportement des utilisateurs. En conclusion, la mise en place d'un processus de certification ne dispenserait pas les organismes de statistique de l'obligation d'investir dans les ressources humaines pour mener à bien le processus.