



---

**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****Vers la certification de nouvelles statistiques d'intérêt  
général : facteurs et modalités****Note de l'Institut national de la statistique  
et des études économiques (France)***Résumé*

Le présent document étudie la question de la qualification de nouvelles statistiques d'intérêt général en France et soumet une proposition en la matière. Il décrit la situation actuelle en France concernant la qualification des statistiques produites en dehors du système statistique national, qu'elles soient publiques ou privées, et les problématiques et difficultés liées à leur essor, ces statistiques pouvant compléter mais parfois aussi remettre en cause les statistiques officielles. Les auteurs proposent un mécanisme à trois niveaux, le premier se limitant à l'auto-documentation d'une statistique selon un modèle prédéfini, le deuxième correspondant à l'approbation de cette documentation et le troisième à l'approbation de la documentation, ainsi que de la méthode et des résultats.

Le présent document est présenté pour examen au séminaire organisé en 2019 par la Conférence des statisticiens européens sur le thème « Le nouveau rôle des organismes nationaux de statistique dans le domaine de la gestion de données » dont la première séance porte sur les « nouvelles possibilités et problématiques relatives au système de données ».



## I. Introduction

1. En France, les statistiques publiques englobent les produits statistiques d'enquêtes ou de sources administratives qui émanent non seulement du système statistique national (SSN) mais aussi d'institutions publiques ou d'organismes privés chargés de missions de service public. C'est donc en puissance un très vaste ensemble de statistiques constituant un bien public mis à la disposition de tous les utilisateurs potentiels.
2. À l'ère du libre accès aux données publiques et de la numérisation de l'économie, les nouveaux ensembles de données et statistiques publics ou privés voient leur champ s'élargir. La production de données ou de statistiques s'est donc généralisée à de nouveaux acteurs qui ont leurs propres sources de données de par leur activité ou qui réutilisent les sources publiques grâce à l'offre croissante de données ouvertes (brutes). Ces nouvelles statistiques peuvent coïncider avec le domaine des statistiques publiques, en offrant un éclairage complémentaire ou en induisant un certain volume de bruit, si les messages diffèrent, et couvrent aussi de nouveaux champs d'observation.
3. Ce phénomène n'est pas nouveau : par exemple, il y a toujours eu des enquêtes, effectuées par des organismes professionnels, qui ont produit et publié des statistiques largement diffusées dans les médias. Ce qui a changé avec la révolution numérique, c'est la capacité qui se répand parmi les acteurs du secteur privé et de la société civile à produire des données en masse et à extraire de nouvelles statistiques d'intérêt général.
4. L'INSEE, dans sa stratégie pour 2025, a inscrit une réflexion prospective visant à « attribuer au service statistique public la fonction de certification des données ». Dans un contexte où le phénomène des « fausses nouvelles » prend de l'ampleur, cette démarche semble d'autant plus légitime. Comment établir cette fonction de qualification des nouvelles statistiques pour donner davantage de visibilité à celles qui amènent une plus-value ? Quel serait le champ d'application pertinent ? Quelles sont les méthodes possibles pour la mise en œuvre de cette nouvelle fonction ?
5. Pour répondre à ces questions, l'Inspection générale de l'INSEE a été chargée d'effectuer une mission dont les principales conclusions figurent dans le présent document.

## II. Point sur la situation en France

6. Le système statistique national (SSN) est déjà confronté à la nécessité de qualifier les statistiques qui alimentent le débat public. Une partie des produits statistiques des organismes chargés d'une mission de service public est donc ainsi « labellisée » par l'Autorité de la statistique. Cette procédure a pour but de garantir le respect des principes du Code de bonnes pratiques de la statistique européenne : l'indépendance professionnelle dans l'élaboration, la production et la diffusion des statistiques publiques, ainsi que les principes d'objectivité, d'impartialité, de pertinence et de qualité des indicateurs établis.
7. D'autres statistiques, élaborées par des acteurs privés ou des organisations de la société civile, sont « étalonnées » pour employer la terminologie en vigueur au Conseil national de l'information statistique (CNIS). Toutefois, cette procédure est relativement lourde : elle prévoit non seulement l'évaluation de la documentation mais aussi la validation de la méthodologie. En conséquence, le nombre de statistiques « étalonnées » reste très faible par rapport au volume de données produites par le secteur privé ou les organisations de la société civile, alors que certaines d'entre elles trouvent un grand écho dans les médias ou dans le débat public.
8. Le système de statistique a déjà saisi certaines des possibilités offertes par les nouvelles sources publiques ou privées, plutôt que de les ignorer et de ne rien changer à ses habitudes, ce qui l'aurait clairement exposé au risque de marginalisation, car la réputation et la confiance dans le système de statistique ne peuvent pas se substituer à la recherche de l'innovation (Ricciatio *et al.*, 2018). De nombreux instituts nationaux de statistique (INS) ont élaboré des plans pour utiliser des mégadonnées comme ressources pour la production de leurs statistiques. L'INSEE a créé en mai 2018 un laboratoire d'innovation (le SSPLab)

qui est un centre de ressources et d'animation pour la recherche appliquée et le développement expérimental visant à promouvoir l'innovation et la nouveauté en matière de sources de données, de technologies et de science des données. Le SSPLab participe aux travaux de l'Équipe spéciale d'Eurostat sur les mégadonnées. Par exemple, il a expérimenté l'utilisation des données issues des téléphones portables pour produire des estimations de la population.

9. Ainsi, les systèmes statistiques nationaux ont réagi rapidement pour investir le domaine des mégadonnées et en développer l'utilisation pour produire des statistiques officielles. Toutefois, il ne peut être exclu que des statistiques entièrement privées provenant de nouvelles technologies mettent en cause la production de statistiques officielles. Sur un plan plus positif, les statistiques privées pourraient aussi compléter les statistiques officielles, le problème étant ensuite de les reconnaître comme des éléments d'information valables entre autres sources.

10. Par exemple, certaines entreprises privées produisent des rapports sur l'emploi qui dérivent de leur activité de gestion de la paie pour des sociétés privées ou utilisent des techniques d'extractions de contenus de sites Web pour collecter des extraits de presse sur les plans de licenciement ou de recrutement ou sur les investissements locaux. Le modèle sur lequel cette activité repose ne consiste pas à commercialiser les statistiques agrégées mais de chercher à obtenir de la visibilité et de la crédibilité auprès des médias afin de vendre des produits commerciaux tels que des renseignements détaillés destinés à des clients privés.

11. Toutefois, les principaux producteurs de statistiques non officielles appartiennent au secteur public : ce sont divers observatoires financés par des fonds publics ou des fonds de collectivités locales, par exemple de régions ou de grandes villes. Ils produisent des statistiques susceptibles de compléter les statistiques publiques dans un domaine spécifique (sectoriel ou géographique).

### III. Problématiques et difficultés

12. Le premier problème consiste à fournir des informations pertinentes au public en accompagnant les statistiques produites des métadonnées qui leur correspondent. Le but premier, fondamental des métadonnées statistiques est d'aider les utilisateurs à interpréter, comprendre et analyser les statistiques. Le niveau de détail des métadonnées dépend des besoins et des compétences des utilisateurs. Par exemple, un journaliste peut ne pas avoir le temps d'assimiler un grand volume de métadonnées détaillées et il lui faudra des informations minimales pour éviter les erreurs d'interprétation. À l'inverse, un spécialiste des sciences sociales pourra aller jusqu'à contester les hypothèses formulées par le producteur initial des statistiques et déduire de nouveaux résultats statistiques en se fondant sur d'autres hypothèses. Ce dernier type d'utilisateur devrait avoir accès à toutes les hypothèses pertinentes et à tous les autres paramètres relatifs à la collecte, l'élaboration et l'évaluation des données de processus, telles qu'elles ont été conçues et mises en œuvre par le producteur de statistiques (Dippo, Sundgren, 2000).

13. Au-delà des métadonnées, ce qui compte, c'est la qualité des données produites. Le Code de bonnes pratiques de la statistique européenne régit le système public de statistique en France et dans les États Membres. Il veille à ce que les statistiques produites se fondent sur les principes de qualité, d'indépendance professionnelle, de pertinence, de qualité des processus et des produits et soient accompagnées des métadonnées qui leur correspondent.

14. Les statistiques produites à l'extérieur du système statistique national ne fournissent pas toujours les métadonnées nécessaires ni ne renvoient aux engagements pris en matière de qualité. Les entreprises privées fournissent quelques éléments décrivant la méthode suivie, mais souvent de façon lacunaire. Par exemple, il arrive souvent qu'elles ne mentionnent pas le domaine couvert, ou les variations possibles dans la collecte d'informations qui peuvent jouer sur la comparabilité temporelle des données. Même le concept, et la façon de le mesurer, peuvent être incomplets, ce qui est source d'erreurs d'interprétation.

15. Il est donc primordial que les utilisateurs aient de bonnes connaissances dans le domaine statistique pour pouvoir comprendre la façon dont les statistiques sont produites et évaluer leur qualité. Par exemple, Radermacher (2018) fait valoir que les statisticiens devraient se consacrer en amont au développement d'une culture des données dans la société pour former les citoyens à un meilleur usage des statistiques, officielles ou non.

#### IV. Mécanisme proposé

16. Dans un contexte où le phénomène des « fausses nouvelles » prend de l'ampleur, les journalistes se heurtent quotidiennement à la nécessité de vérifier les informations à leur disposition. Dans les médias français, comme ailleurs dans le monde, les rédactions sont de plus en plus nombreuses à employer des collaborateurs exclusivement à la vérification des faits. Le fondement même du travail journalistique est de recouper les sources et de contrôler leur fiabilité (Léchenet, 2015). Aussi les journalistes préfèrent-ils mener leurs recherches seuls à partir de sources existantes plutôt que d'être « guidés » par l'une des parties prenantes. Ainsi, les INS doivent « qualifier » les sources extérieures avec le plus grand soin, au risque d'être accusés de protéger leur monopole en pointant du doigt celles qui ne sont pas fiables.

17. Par conséquent, le mécanisme proposé est un processus de qualification des statistiques à trois niveaux assorti d'un gradient d'exigence croissante. Le premier niveau devrait prévoir l'auto-documentation des données statistiques selon un modèle normalisé à définir. Le deuxième niveau devrait correspondre à l'approbation (l'« enregistrement ») de la documentation, sur une base volontaire. Il ne s'agit pas d'approuver la méthode ou ses produits, mais de s'assurer seulement que la documentation est suffisamment complète pour fournir les informations nécessaires aux utilisateurs. Le troisième niveau devrait comprendre une évaluation de la qualité et de la pertinence des résultats.

18. Cette procédure concernerait principalement les producteurs de statistiques chargés de missions de service public, qui sont potentiellement beaucoup plus susceptibles de souscrire à ce type de mécanisme que les producteurs privés. On part de l'hypothèse que cela pourrait avoir des effets d'entraînement positifs, même dans la sphère privée.

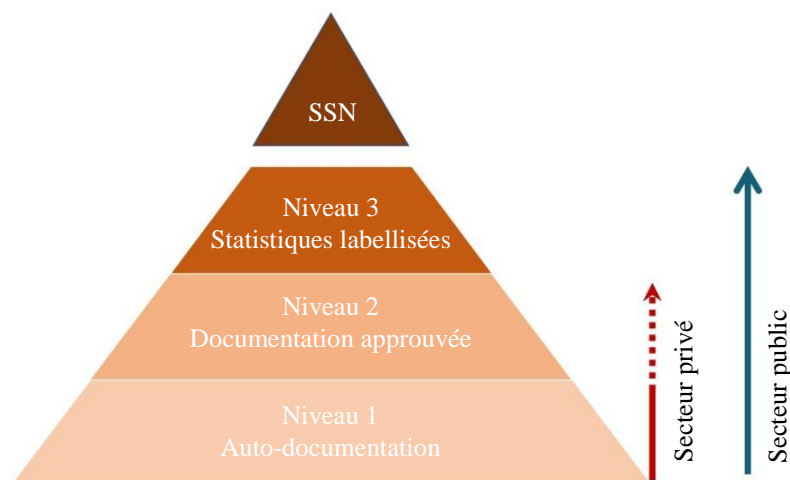
19. Le CNIS devrait d'abord se prononcer sur la recevabilité de la demande pour évaluer la documentation d'un produit statistique en fonction de sa pertinence au regard de l'information publique. Une fois l'avis de recevabilité obtenu, l'évaluation de la documentation serait confiée à un comité technique. Ce comité comprendrait des experts appartenant ou non au système statistique national. Cette instruction implique un coût variable, principalement à la charge du Comité : l'approbation de la documentation d'une statistique peut exiger des allers-retours avec le requérant. Pour réduire ce coût, il est nécessaire de préciser la norme ainsi que les conditions de son utilisation. L'avis du Comité ne porterait que sur la qualité de la documentation, sans avaliser la qualité du produit.

20. Ce mécanisme exige l'élaboration d'un modèle de documentation de statistique. Pour établir la norme, il est proposé de créer un groupe de travail composé de statisticiens officiels et d'experts extérieurs au service statistique national, ainsi que de producteurs de statistiques potentiels du secteur privé et du secteur public, l'objectif étant que le modèle fournisse l'information pertinente à communiquer aux utilisateurs.

21. L'approbation de la documentation serait accordée pour une durée limitée renouvelable et un répertoire en ligne donnerait la liste des statistiques approuvées.

22. Le troisième niveau de qualification correspond à la labellisation des statistiques publiques : il serait donc réservé aux statistiques produites par les organismes effectuant des missions de service public. Il comprend actuellement une évaluation de la pertinence de la production et de la méthode utilisée, ainsi que la vérification de la conformité aux principes fondamentaux du Code de bonnes pratiques de la statistique européenne. Comme l'INSEE et les départements de statistique des ministères, qui constituent le service national de statistique, sont en rapport avec de nombreux partenaires produisant des statistiques publiques, les membres du SSN pourraient avoir un rôle moteur à jouer pour accompagner leurs partenaires dans leurs travaux de labellisation des sources. Cela ferait intervenir une approche ascendante, et pourrait permettre d'élargir la portée des statistiques labellisées.

Figure 1  
**Mécanisme à trois niveaux pour la qualification des sources statistiques**



23. Toutefois, la procédure d’approbation restera inopérante si les candidats potentiels en ignorent l’existence et si les utilisateurs ne savent pas distinguer les différents niveaux de qualité. La pyramide constituée de niveaux croissants d’exigence de qualité, avec le SSN au sommet, et l’engagement pris de respecter le Code de bonnes pratiques de la statistique européenne doivent être illustrés de façon visuelle et pédagogique.

24. Enfin, l’un des moyens de contrôler ou de limiter l’audience des statistiques non officielles est de mener une action de communication plus convaincante concernant nos engagements et nos succès sur le plan de la qualité. Il est donc recommandé de mettre au point une communication sur la qualité des statistiques qui soit davantage orientée vers les utilisateurs, en particulier sur le site Web de l’INSEE. Par exemple, l’Autorité de statistique du Royaume-Uni a adopté ce type de communication axée sur l’utilisateur en organisant son code de pratique des statistiques suivant trois piliers (fiabilité, qualité et utilité) qui le rendent plus aisément compréhensible.

## V. Conclusions et recommandations

Le séminaire de la Conférence des statisticiens européens pourrait donner lieu à des échanges sur les diverses situations nationales et sur la possibilité d’adopter une approche commune de la qualification ou de la certification des statistiques non officielles.

## VI. Références

- Dippo, C, Sudgren, B. (2000). « The Role of Metadata in Statistics », document présenté à la Conférence internationale sur les enquêtes auprès des établissements (ICES-II), Buffalo, New York. <https://www.bls.gov/osmr/abstract/st/st000040.htm>.
- Léchenet, A. (2015). Le journalisme de données. Synthèse des débats. Cafés de la statistique, soirée du 14 avril 2015. Société Française de statistique.
- Radermacher, W. (2018). Official statistics in the era of big data opportunities and threats. *International Journal of Data Science and Analytics*. 6:225-231. Springer International Publishing AG.
- Ricciato, R. et al. (2018). Towards a reference architecture for trusted smart statistics. Équipe spéciale d’Eurostat sur les mégadonnées. Document présenté à la Conférence européenne sur la qualité dans les statistiques officielles de 2018.