
Neuchâtel, le 31.05.2022

Utilisation de sources de données diverses par la statistique publique

1. Contexte

Les considérations ci-après sont fondées sur la conviction du Conseil d'éthique que la statistique publique, qui jouit d'une large acceptation et d'une grande crédibilité aujourd'hui, ne peut maintenir son niveau que si elle tire profit du défi de la transformation numérique. Elle doit p. ex. exploiter de nouvelles sources de données, qui sortent pour l'heure des sentiers battus, et fournir des adaptations méthodologiques et des prestations de service dans le cadre de la gestion des données ou de la science des données. Notamment en raison de la concurrence des fournisseurs de statistiques privés, dont les standards de qualité ne correspondent pas toujours à ceux de la statistique publique ou du moins ne satisfont pas à l'impératif de la transparence, il y a lieu de donner des réponses stratégiques innovantes allant au-delà des méthodes et technologies traditionnelles. Cela implique bien entendu d'aborder des problématiques touchant l'éthique des données.

À l'Office fédéral de la statistique (OFS), la mise en œuvre de l'élargissement des tâches est clairement perceptible dans l'adaptation de la structure d'organisation opérée en 2021 (nouvelles divisions Interopérabilité et registres IOR et Science des données et méthodes statistiques DSSM). Les services statistiques régionaux prennent en main l'élargissement des tâches avec un certain retard. Mais ils s'établissent de plus en plus comme centre de compétences pour la gestion des données, et certains services, surtout ceux de plus grande taille, assument en outre des tâches dans le cadre de l'Open Government Data (OGD). Ainsi, tous les services de la statistique publique devraient tôt ou tard être confrontés à de nouvelles problématiques de déontologie et d'éthique des données et procéder le cas échéant à des adaptations de directives éthiques comme celles de la Charte. Mais à l'heure actuelle, c'est surtout l'OFS qui est ici à l'œuvre en raison des nouvelles tâches que le Conseil fédéral lui a confiées.

Il n'est dès lors pas étonnant que les suggestions et questions touchant ce domaine émanent actuellement plutôt de l'OFS. Leur traitement par le Conseil d'éthique relève cependant de la responsabilité de toute la statistique publique, ce qui inclut donc aussi les services statistiques régionaux. Ces derniers sont en outre associés aux réflexions via divers canaux (groupes de travail OFS-CORSTAT, Regiostat, etc.).

L'OFS et le Conseil d'éthique procèdent de manière itérative pour traiter les problématiques liées à l'éthique des données. L'OFS s'adresse ponctuellement au Conseil d'éthique, en fonction de l'évolution du processus d'élargissement et des projets. Ainsi, les réflexions présentées ci-après portant sur l'utilisation par la statistique publique de données diverses ont été élaborées l'année dernière.

2. Problématique: codes de conduite et nouvelles grandes sources de données (Big Data)

La question abordée en premier est celle de savoir si les codes de conduite tels que les Principes fondamentaux, le Code de bonnes pratiques ou la Charte prévoient certaines restrictions dans l'utilisation par la statistique publique de nouvelles grandes sources de données. Diverses publications sont disponibles à ce sujet. Nous nous référons ici à une contribution de D. Rozkrut et al. intitulée «Mapping the UN Fundamental Principles of Official Statistics against new and big data sources»¹. Cette contribution examine les Principes fondamentaux de la statistique publique de l'ONU dans le contexte de huit nouvelles grandes sources de données: téléphonie mobile, compteurs d'électricité, images satellites, réseaux sociaux, web scraping, capteurs de trafic routier et suivi des passagers, données scannées et données du système de caisse, données de vidéosurveillance.

Les auteurs de la contribution arrivent à la conclusion que toutes ces sources de données peuvent être utilisées dans la production de la statistique publique en conformité avec les principes fondamentaux. Ces sources devraient même être exploitées si les systèmes statistiques nationaux veulent respecter le premier des principes fondamentaux, à savoir celui de satisfaire au droit d'accès des citoyens à l'information publique par l'établissement et la mise à disposition de statistiques publiques.

Sans entrer dans les détails de ce document, le Conseil d'éthique approuve ce constat et le complète par les considérations générales et les concrétisations suivantes:

2.1 Considérations générales

Les codes de conduite existants

- «Fundamental Principles of Official Statistics» de l'ONU (établis en 1992 / révisés en 2013),
- «Code of Practice» (Code de bonnes pratiques) du Système statistique européen (CoP, 2005 / 2011, 2017) et
- «Charte de la statistique publique de la Suisse» établie sur la base du CoP (Charte, 2002 / 2008, 2012)

... sont en premier lieu des directives déontologiques pour les services de la statistique publique

- sur le mandat d'information
- sur l'indépendance professionnelle
- sur la protection de la personnalité et des données
- sur l'efficacité
- sur la qualité

Sont ici prioritaires la crédibilité que la statistique publique doit viser par la **transparence de ses tâches et de ses méthodes de travail** et par la production de statistiques publiques répondant à des exigences qualitatives élevées. Il s'agit en premier lieu de savoir **comment** le traitement des données (soit l'acquisition, la préparation et la publication des résultats) doit se faire dans le cadre de la statistique «classique». Aucune restriction n'est explicitement formulée quant aux sources de données utilisées et notamment quant aux possibilités offertes par les Big Data. Si les services statistiques respectent les directives déontologiques des codes de conduite dans le traitement des données, ils peuvent alors traiter des données tirées des sources de leur choix.

¹ Statistical Journal of the IAOS 37 (2021) 161–169

2.2 Concrétisations

2.2.1 Mandat d'information

Pour donner suite à ce qui a été décrit ci-dessus, la statistique publique est tenue de recourir à de nouvelles sources de données permettant de remplir son mandat fondamental de couvrir les besoins en informations statistiques pertinentes d'intérêt général. Ce mandat est formulé dans le principe 1 des Principes fondamentaux et de la Charte. Dans le CoP, la pertinence des informations statistiques figure dans le principe 11, le principe 1 étant consacré à l'indépendance des services statistiques.

2.2.2 Les codes de conduite ne contiennent pas de directives sur les sources de données

Des contraintes telles que par exemple des restrictions relatives aux sources de données sont-elles formulées dans les codes de conduite considérés?

- Dans les Principes fondamentaux, le principe 5 indique que les données utilisées à des fins statistiques peuvent être tirées de toutes sortes de sources.
- Le préambule de la version révisée en 2017 du Code de bonnes pratiques mentionne entre autres que «L'objectif est d'intégrer les changements et les innovations les plus récents concernant l'élaboration, la production et la diffusion des statistiques officielles dans le système statistique européen et au-delà, par exemple les nouvelles sources de données émergentes, l'utilisation des nouvelles technologies, la modernisation du cadre juridique et les résultats des revues par les pairs relatives à la mise en œuvre du code de bonnes pratiques.». La référence à des «informations provenant de multiples sources de données» dans la version révisée est reprise dans le principe 2 ainsi que dans plusieurs indicateurs du Code de bonnes pratiques.
- La Charte ne prévoit aucune restriction relative aux sources de données. À la différence des Principes fondamentaux et du Code de bonnes pratiques, elle n'évoque cependant pas la possibilité d'exploiter des sources de données multiples. **La Charte devrait être complétée en conséquence lors de la prochaine révision.**

2.3 Conclusion

Du point de vue des trois codes de conduite considérés, le recours à des sources de données diverses ne pose aucun problème fondamental. Mais le traitement de nouvelles grandes sources de données soulève des questions à différents niveaux, que nous évoquons brièvement ci-après (sans garantir leur exhaustivité):

3. Nouvelles questions

3.1 Acquisition de données auprès de fournisseurs privés

Comme les fournisseurs privés ne sont le plus souvent pas tenus légalement de remettre des données à la statistique publique, ils le font sur une base volontaire.

3.1.1 Garantie de l'indépendance professionnelle

Pour la statistique publique, il est primordial d'assurer l'indépendance professionnelle. Elle peut le faire en faisant preuve de transparence dans la désignation et le choix des fournisseurs de données. En contrepartie, les services statistiques doivent assurer aux fournisseurs privés qu'ils utiliseront leurs données exclusivement à des fins statistiques.

3.1.2 Question des coûts

Il convient bien sûr de discuter de la question des coûts liés à l'acquisition par la statistique publique de données de fournisseurs privés. Cet aspect financier touche des problématiques politiques et éthiques, notamment des questions liées à l'étendue du service public dans le domaine de l'infrastructure des données et de son utilisation entre autres par la statistique publique.

3.2 Un point particulier à considérer: information des milieux interrogés en cas de recours à de nouvelles sources de données?

Le recours à de nouvelles sources de données telles que les réseaux sociaux ou les données issues de la télécommunication n'est pas prévu dans les lois sur la statistique, du moins pas directement. Les milieux interrogés, soit ici les exploitants / propriétaires des sources de données concernées, doivent être informés des buts et des bases légales, etc. des enquêtes en vertu de l'art. 4, al. 4 LSF et du principe 2 de la Charte (base légale). Une telle information n'est pas vraiment problématique vis-à-vis de ces instances. Est-ce suffisant d'informer leurs fournisseurs de données que leurs données personnelles sont communiquées à la statistique publique à des fins statistiques via les Conditions générales (CG)? Doivent-ils absolument en être informés? Ou faut-il limiter la communication à des données non personnelles?

3.3 Nouveaux défis méthodologiques et qualitatifs

Les nouvelles grandes sources de données ne sont généralement pas conçues à des fins statistiques. Elles ne correspondent pas à la statistique publique ni en termes de volume ni à ses modèles traditionnels au niveau de la structure ou du contenu. Les nouvelles sources de données ne répondent pas nécessairement aux exigences statistiques en ce qui concerne les définitions, les standards, la représentativité, la couverture ou la continuité. Cette situation amène donc de nouvelles questions et possibilités quant aux méthodes statistiques et aux exigences qualitatives; elles sont abordées actuellement dans le cadre de la stratégie d'innovation sur les données de l'OFS. Quels critères sont valables p. ex. pour le contrôle de qualité de données dont le service statistique ne sait pas comment elles ont été créées et pour lesquelles les métadonnées s'y rapportant sont parfois lacunaires? Il faut donc renforcer les divisions de recherche dans la statistique publique ainsi que la collaboration avec les hautes écoles et instituts de recherche; la division DSSM nouvellement créée tient compte de ces enjeux.

3.4 Labellisation des résultats «en dehors» de la production statistique

Dans le cadre de la stratégie d'innovation sur les données, il est possible d'obtenir des résultats qui sont en adéquation avec les directives de la Charte et peuvent dès lors être intégrés dans la production statistique. Lorsque ce n'est pas le cas, si p. ex. la qualité ne correspond pas (encore) aux exigences en cours ou qu'on se trouve seulement à la phase de développement de projets, il faut chercher des possibilités répondant aux besoins des parties prenantes sans compromettre la crédibilité de la statistique publique. La voie généralement suivie actuellement à l'OFS (ainsi que dans le système statistique européen SSE) est de publier de tels résultats issus de la statistique expérimentale sous un label et avec les métainformations s'y rapportant.

3.5 Limites

La production statistique établie en recourant à de nouvelles méthodes doit respecter les directives des codes de conduite valant pour la statistique publique. Quelles solutions devons-nous adopter pour des cas comme ...

3.5.1 ... la statistique expérimentale

Dans quelle mesure peut-on outrepasser ces directives, par ex. dans la statistique expérimentale, et comment prévoir l'information à ce sujet et la labellisation?

- Est-il possible d'apporter des adaptations / des compléments à la Charte ou faut-il établir d'autres directives éthiques?

3.5.2 ... les données administratives et statistiques

Les délimitations entre «statistique» et «non statistique» de même qu'entre p. ex. «données administratives» et «données statistiques» peuvent être floues.

- Peut-on fixer des limites et en fin de compte des définitions sur la base de directives d'éthique des données?
- Comment procéder en sachant que les informations conçues au départ pour la statistique sont aussi de plus en plus souvent sollicitées par l'administration et que cette situation affecte la législation en vigueur et les codes de conduite.

3.6 Autres (*apparaîtront au fil de l'avancement du processus*)

Au nom du Conseil d'éthique

Peter Laube