Chroniques

Conseil national
de l'information statistique

Les enjeux des nouvelles sources de données

L'idée d'une rencontre pour débattre des enjeux des nouvelles sources de données est née au cours de l'élaboration du nouveau programme du moyen terme 2019-2023 du Conseil national de l'information statistique. La rencontre, qui s'est déroulée le 2 juillet 2018, a rassemblé un grand nombre d'acteurs représentants de la statistique publique mais également des chercheurs, des universitaires, des représentants du monde associatif, des partenaires sociaux. L'intervention de Mireille Elbaum, Présidente du Haut Conseil du financement de la protection sociale, en ouverture de la journée a posé les questions fondamentales de ces nouvelles sources : de quoi s'agit-il ? quels nouveaux défis posent-elles à la statistique publique ? quels changements vont-elles induire dans les relations qu'entretient la statistique publique avec la recherche mais aussi avec la société civile ? Cette intervention, reprise ici intégralement, constitue une base de tout premier ordre aux réflexions à venir.

Tout d'abord merci au Cnis et à son secrétariat général de m'avoir conviée à introduire cette rencontre. Les mots que je vais dire ici n'ont aucunement la prétention d'être une conférence ou même un cadrage introductif d'ensemble, dans la mesure où, je souhaite le dire d'emblée, je ne suis en rien spécialiste du sujet, et presque en position de « Candide » en la matière.

Mon expérience personnelle récente m'a conduite à le rencontrer sur trois terrains particuliers :
– les réflexions de l'*European Statistical Governance Advisory Board* (Esgab), où je viens d'être nommée, et qui dans son rapport de 2017 a voulu éclairer les opportunités, mais aussi les risques et les problèmes suscités pour ce nouvel environnement ;
– les travaux du Haut conseil du financement de la protection sociale (HCFiPS), qui, dans le rapport qu'il a consacré aux relations entre organismes sociaux et entreprises, s'est intéressé à l'utilisation potentielle du *datamining* pour repérer les situations potentielles soit de difficultés précoces, soit au contraire de fraude ;
– mes missions à l'Inspection générale des affaires sociales (Igas), et notamment le travail que j'ai conduit sur le contrat d'objectifs de Santé publique France (SPF), qui a mis en évidence les potentialités, mais aussi les difficultés et les arbitrages auxquels étaient confrontés les épidémiologistes eu égard à la multiplication des sources et des acteurs qui les produisent.

Mon point de vue est donc essentiellement celui d'une utilisatrice spécialiste des politiques sanitaires et sociales, mais aussi celui d'une personne profondément attachée au système statistique, avec la conscience que, non seulement il est partie constitutive de la démocratie, mais que,

Chroniques n° 16
Septembre 2018
Directrice de la publication : Françoise Maurel
Rédactrice en chef : Isabelle Anxionnaz
Maquette : Rose Pinelli Vanbauce
Publication diffusée gratuitement,
ne peut être vendue
ISBN 978-2-11-139683-8

comme l'ont montré les travaux d'Alain Desrosières[1], il contribue à forger, par les concepts et les pratiques qu'il promeut, les représentations sur lesquelles s'appuieront les acteurs politiques et sociaux, parfois des années plus tard, pour asseoir leurs positions et décisions.

C'est à partir de cette conviction qu'il nous faut à mon sens interroger les enjeux de ces nouvelles sources de données, ce terme choisi par le Cnis étant plus judicieux et plus large que celui de *Big Data*.

Les questions introductives que je vais soulever sont ici de trois types :
– De quoi parle-t-on lorsque l'on traite des « nouvelles sources de données » et qu'est-ce que cela a réellement de nouveau ?
– Quels sont les principaux enjeux mais aussi problèmes qui en découlent directement pour la production statistique publique ?
– Quels sont aussi ces enjeux et problèmes dans les relations que la statistique publique entretient d'une part avec la recherche économique et sociale, d'autre part avec la société civile et les citoyens, relations dont le Cnis est le lieu privilégié ?

De quoi parle-t-on à propos des « nouvelles sources de données » et qu'est-ce que cela a de réellement nouveau ?

Le « monde » des données statistiques sur lesquelles nous avons l'habitude de nous appuyer, notamment en matière sociale, repose traditionnellement sur une **dichotomie entre deux catégories de données**, qui ont chacune des avantages et des limites bien connus.

Du côté des données administratives, leur caractère de « produit fatal » (ou de sous-produit) des systèmes de gestion publics permet, lorsque ceux-ci ont une portée générale, une production à un moindre coût, à un niveau très détaillé (local ou par sous-catégorie de contribuables ou de bénéficiaires), et, en principe, une disponibilité régulière. En revanche, ces données peuvent être attachées à des dispositifs particuliers, n'ont une qualité suffisante que si elles découlent directement de la gestion, et peuvent donc être pauvres en informations sociodémographiques et de contexte, ainsi que, qui plus est, en éléments d'appréciation sur les besoins exprimés par les bénéficiaires et les réponses qui leur sont apportées.

Du côté des enquêtes auprès des entreprises et surtout des ménages, les avantages et les inconvénients sont inverses : à la possibilité d'adapter les recueils d'informations à des questionnements jugés pertinents, et de les mettre en relation avec une diversité de caractéristiques socioéconomiques ou d'environnement, s'oppose le coût de ces enquêtes, qui peut en faire une « donnée rare », et le fait que les acteurs locaux ou spécialisés n'y trouvent pas toujours leur compte.

En matière d'évaluation et plus largement d'enquêtes sur les politiques sociales, la tendance a d'ailleurs été au couplage et à l'appariement de ces deux types de données, par exemple en ce qui concerne les bénéficiaires de minima sociaux ou les consommateurs de soins de santé.

Au regard de « cette dichotomie » traditionnelle, qu'entend-on alors par « nouvelles sources de données » ?

Le principal point à noter est qu'il s'agit d'un ensemble de sources divers et composite dont il faut clarifier la nature et les propriétés, dans la mesure où elles n'entraînent pas le même type de problèmes.

On a d'abord ce que j'appellerai des « données administratives +++ », qui naissent du perfectionnement, de l'ouverture et surtout de l'appariement des données issues des systèmes de gestion publique (cf. le système national des données de santé ou les informations provenant de la déclaration sociale nominative - DSN- ou du futur du prélèvement à la source) : leurs caractéristiques et leurs limites restent celles des données administratives, les nouvelles potentialités étant liées aux élargissements et aux appariements permis par ces dispositifs.

On a ensuite, sur des sujets qui relèvent de domaines déjà couverts par la statistique publique, la possibilité d'utiliser des données collectées, comme sous-produit de leur activité, par des acteurs privés. Elles ont par nature le même type de caractéristiques que les données administratives, sachant toutefois que, si elles ouvrent la possibilité de réduire notablement les coûts de collecte, elles peuvent avoir un champ incomplet et que leur stabilité peut dépendre des politiques suivies par les entreprises à des fins industrielles ou commerciales.

On a enfin un ensemble divers et diffus de données, provenant notamment de la téléphonie mobile ou des réseaux sociaux, dont on ne sait pas complètement aujourd'hui dans quelle mesure elles peuvent contribuer à améliorer les modèles de prévision (cf. les informations obtenues à partir de *Google Trends* ou des offres d'emploi publiées sur Internet)[2], et *a fortiori* si elles pourront servir de base d'abord à des études pertinentes, puis à une éventuelle collecte statistique, cette question se posant par exemple dans le domaine sanitaire pour les informations provenant des réseaux de patients.

Qu'y a-t-il alors véritablement de nouveau ?

La nouveauté ne réside donc pas dans une éventuelle remise en cause de la distinction entre données de gestion et enquêtes, notamment auprès des ménages, qui conserve sa pertinence. Les « nouvelles » sources de données sont par contre associées à :
– des caractéristiques habituellement mises en avant à propos des *Big Data*, à savoir des volumes massifs et une granularité très fine ; dans certains cas, une grande rapidité d'obtention et de traitement ; et surtout potentiellement une plus grande variété, démultipliée par les appariements possibles ;
– l'irruption plus marquée d'acteurs de la sphère privée comme producteurs et détenteurs de ces données, parfois en tant que sous-produits de leur activité principale (données de caisse…), mais parfois aussi comme éléments clés de cette dernière (gestionnaires d'accès, réseaux sociaux…) ;
– un champ élargi d'acteurs, de réseaux et de potentialités à explorer, qui comportent un intérêt manifeste, mais aussi des inconnues et des risques.

1 A. Desrosières, *Introduction aux deux livres : Pour une sociologie historique de la quantification et Gouverner par les nombres*, Presses de l'École des Mines, 2008.
2 D. Blanchet, P. Givord, « Données massives, statistique publique et mesure de l'économie », *L'économie française - Comptes et dossier*, Insee Références, édition 2017.

Quels sont les principaux enjeux et problèmes qu'induisent ces nouvelles sources de données pour la production statistique publique ?

Ces enjeux et problèmes sont justement différents selon les types de données et de détenteurs, et suscitent des dilemmes pas forcément aisés à arbitrer pour les décideurs de la sphère statistique publique.

Le premier enjeu, sans doute le plus clairement apparent, concerne l'adaptation des compétences et des méthodes de la statistique publique au traitement de ces nouvelles sources.

Cela passe notamment :
– par l'accès à des capacités de stockage et de traitement (*cloud* - ou *nebulae* pour certains *geek* pratiquant le latin-, réseaux d'ordinateurs et programmes) permettant de traiter des masses de données très importantes ;
– par la prise en compte dans les méthodes d'estimation du fait que, sur des données en nombre très important, les coefficients estimés sont systématiquement significatifs, mais que leur validation peut impliquer d'autres méthodes, telle la reproduction de l'exercice ;
– par l'enrichissement des prévisions d'activité et d'emploi (grâce à des outils comme *Google Trends* ou les offres d'emploi sur Internet), en ayant toutefois conscience que plus les modèles prédictifs retracent fidèlement le passé, plus ils peuvent être inadaptés à percevoir des changements d'environnement ou même de fonctionnement de certaines sources (cf. l'expérience du recours *Google Flu* pour anticiper la diffusion de l'épidémie de grippe) ;
– par une réflexion renouvelée sur certains concepts ou pratiques clés de la statistique, à savoir :

- les biais liés aux différents types de collecte (ce n'est pas parce que les données sont massives qu'elles ne sont pas biaisées…) ;
- la distinction entre des corrélations qui peuvent dans certains cas être mises en évidence presque « par hasard », et les interprétations ou explications qui peuvent en être données en termes de causalités ;
- les cadres et modalités d'agrégation des données, qui peuvent être conçus de manière plus fluide et plus adaptée à de nouveaux « groupes », mais qui nécessitent une réflexion quant à leur pertinence.

L'adaptation des compétences au sein de la sphère statistique publique nécessite enfin, et c'est à ne pas oublier, le développement de capacités, pas toujours innées chez les *data scientists*, à expliquer de façon claire et transparente les traitements effectués, leur portée et leurs limites, et ce à la fois en direction des spécialistes et des citoyens.

Un deuxième enjeu pour le système statistique public concerne le rôle et les relations à organiser avec les opérateurs privés dans ce nouveau contexte.

Ces opérateurs peuvent d'un côté être demandeurs de données, comme c'est le cas vis-à-vis du système national de données de santé (SNDS), avec à la clé des questions de conditions d'accès et de garantie de la confidentialité des informations individuelles. Ils peuvent aussi être producteurs ou offreurs potentiels de données en direction de la statistique publique, en permettant à cette dernière d'alléger et de rationaliser son système de collecte (données de caisse) ou d'investir dans de nouvelles approches (téléphonie mobile).

Les expériences présentées et discutées lors de cette rencontre mettent en évidence la nécessité d'établir des conditions précises d'utilisation de ces données en vue de « l'intérêt statistique général. Cela implique des règles de **confidentialité tant individuelle que commerciale**, mais aussi des garanties de transparence et de stabilité dans leur mode de production et dans leur accès, dans un contexte de volatilité des opérateurs et des projets économiques.

Je suis en outre frappée par le fait que le rapport qu'a consacré le CNIS à cette question[3] et les expériences présentées ici se rapportent principalement à la réutilisation de données produites par des entreprises à l'occasion d'une autre activité principale (commerciale, bancaire, téléphonique), même si certaines de ces informations peuvent donner lieu à une valorisation économique.

Or, va aussi se poser à l'avenir la question des entreprises, souvent multinationales, dont la valeur même de l'activité économique réside dans la production de données et dans la « segmentation » des publics cibles que ces données permettent à des fins commerciales.
Il n'est d'ailleurs pas impossible que, dans ce cadre, le système statistique public se trouve en situation d'être concurrencé, voire contourné par la production d'informations, certes biaisées et reconstruites ad hoc, mais fournies très rapidement à partir d'échantillons massifs.

Est-ce que les citoyens et les décideurs publics continueront alors à considérer que les productions statistiques publiques « valent la peine », s'il s'agit avant tout de compléter des champs et de redresser des biais dans les délais forcément plus importants ?

Quels arbitrages opérer entre, selon la terminologie communautaire, *timeliness* et *reliability*, et comment empêcher que le système statistique public ne voie ses ressources contestées au vu de la disponibilité de « statistiques » privées d'accès apparemment immédiat et direct ?

Cela conduit à évoquer un troisième enjeu, sans doute le plus important et qui s'inscrit dans un contexte de contrainte budgétaire forte.

S'il est évident, comme cela est affirmé au niveau européen, que l'accès à ces nouvelles sources de données peut réduire à la fois la charge de réponse et les coûts de la collecte statistique, il faut garder à l'esprit les limites de ces processus, ne pas « lâcher la proie pour l'ombre », et conserver le « cœur » et l'identité du système statistique public.

Les enquêtes, notamment auprès des ménages, font partie de cette identité, et l'accès potentiel à de nouvelles données « d'essence administrative » ne saurait justifier de trop restreindre la voilure en ce domaine, sachant que les pistes les plus intéressantes, par exemple en matière d'évaluation des politiques publiques, consistent souvent à coupler et apparier les deux types de sources.

3 F. Dupont, S. Grégoir, *La réutilisation par le système statistique public des informations des entreprises*, Rapport du groupe de travail Insee-Cnis, mai 2016.

Les enquêtes sont en effet l'occasion de partir de questionnements de fond débattus avec les chercheurs et la société civile, et de dépasser la contrainte « d'indicateurs » ou de cadres administratifs préconstruits et pré-formatés. Ce processus est au cœur même de l'indépendance du système statistique, et, comme l'a noté Jacky Fayolle[4], de la résistance opposée par les statisticiens à ce qu'Alain Supiot[5] a appelé « la Gouvernance par les nombres ».
C'est particulièrement important dans le champ des politiques sanitaires et sociales, où les enquêtes ont permis de montrer le poids des inégalités sociales dans des domaines où certains les attendaient peu (par exemple le handicap ou la dépendance), ou d'appréhender la question complexe des discriminations « ressenties ».

La question pour le système statistique public est donc au bout du compte d'affecter au mieux ses ressources pour conserver la maîtrise de ce qu'il mesure, non seulement dans sa qualité, mais aussi dans sa définition et dans son contenu.

De façon plus imagée, il lui faut donc déterminer dans quels paniers mettre ses œufs : dans celui un peu percé des enquêtes, dans celui un peu plus solide des données administratives, mais où on trouve parfois un « petit canard » au lieu du poussin espéré, ou dans l'espoir d'une corne d'abondance (les *Big Data*) qui miroite au loin, mais avec à coup sûr une part d'illusion.
Et c'est aussi cette question du « contenu de ce que l'on mesure » qui est selon moi l'enjeu principal des relations à venir entre système statistique public, recherche et citoyens.

Les enjeux concernant les relations du système statistique avec la recherche, la société civile et les citoyens

Du côté de la recherche, je ne puis ici qu'évoquer brièvement quelques-unes des interactions liées aux nouvelles sources de données. Elles concernent notamment :

– l'enjeu de l'accès aux données administratives et à leurs appariements constitués à des fins statistiques : il implique, comme l'ont montré les travaux d'un groupe du CNIS, des procédures claires, facilitées et suffisamment rapides[6] ;
– l'enjeu de l'exploration commune de questions tant de fond que de méthode, parmi lesquelles on peut par exemple citer :

- Quelles sont les possibilités d'identifier et de repérer de façon « signifiante » de nouveaux comportements économiques ou de nouveaux « groupes » sociaux et/ou culturels ?
- Comment apprécier la valeur économique intrinsèque de la production de données qui est à la base de l'activité de certaines firmes, et que penser de la valorisation financière qui en est faite ?
- Quelles implications ont les modalités de constitution, de recueil et d'agrégation des informations collectées, sachant « qu'une donnée n'est jamais donnée » et rétroagit sur les hypothèses de recherche, que les éléments recueillis sur les comportements des individus dans certaines bases sont à la fois multiples, massifs et incomplets, et que l'agrégation de ces données devient une question clé, mais qu'elle peut donner lieu à des algorithmes « boîte noire » ne permettant pas toujours d'en maîtriser les incidences ?

L'enjeu le plus important à évoquer concerne enfin la logique même de la recherche en économie et en sciences sociales. Va-t-on ainsi passer, comme le craignent certains chercheurs en éthique, à une science dite *data driven*, qui en viendrait à succéder à une recherche empirique décrivant les phénomènes, à une recherche tentant de les expliquer par des hypothèses théoriques et des modélisations et à une recherche fondée sur la simulation calculatoire de phénomènes complexes[7] ?

La question peut paraître un peu obscure, mais je prends le risque de la reformuler brutalement en : « Est-il possible d'avoir des réponses sans avoir pesé et posé les questions ? ».

Cela peut parfois sembler tentant dans le cas des données de santé, qui peuvent faire apparaître des corrélations inattendues et susceptibles d'alerter sur des phénomènes ignorés, mais cela pose aussi des problèmes redoutables de maîtrise, d'explication et d'interprétation des phénomènes et de leur causalité.

Ce sont finalement des problématiques du même ordre que l'on retrouve du côté de la société civile et des citoyens, dont les interactions avec le système statistique public sont la mission première du Cnis (au-delà des sujets de libertés individuelles que je n'aborde pas ici).

Tout d'abord, même si c'est un peu provocateur, je me suis interrogée sur le fait que cette rencontre consacrée au moyen terme ait pour entrée exclusive les sources de données, et non plus les domaines d'observation (démographie, emploi, conditions de vie) et les questions qu'ils peuvent susciter dans un monde en mutation.

Si cette appropriation globale des potentialités et des enjeux des nouvelles sources de données est à l'évidence nécessaire, elle ne saurait remplacer la co-construction des problématiques et des outils d'observation propres à chacun des différents domaines, pour éviter justement que des réponses, établies à partir de données non conçues à cet effet et donnant lieu à des traitements de type « boîte noire », ne se substituent aux questions que porte légitimement le débat social.

En matière de politiques sociales où se posent des problèmes essentiels d'évaluation mais aussi de non recours, faut-il par ailleurs se contenter d'analyser les comportements « administratifs » des individus observés *ex-post*, ou aussi les interroger sur leurs appréciations et leurs attentes ?

4 J. Fayolle, « À propos de la gouvernance par les nombres, pour une articulation de la raison juridique et de la raison statistique », *Droit et Société*, 98/2018.
5 A. Supiot, *La Gouvernance par les nombres : cours au Collège de France 2012-2014*, Fayard, coll. « Poids et mesures du monde », 2015.
6 A. Bozio, P.-Y. Geoffard, *L'accès des chercheurs aux données administratives, État des lieux et propositions d'actions*, Rapport du groupe de travail du Cnis, mars 2017.
7 L. Coutellec, P.-L. Weill-Dubuc, « *Big data* ou l'illusion d'une synthèse par agrégation : une critique épistémologique, éthique et politique », *Journal international de bioéthique et d'éthique des sciences*, vol. 28, 2017/3.

Ce n'est pas toujours pertinent ou possible, mais je voudrais en bout de course évoquer, à titre de boutade, le cauchemar que seraient pour moi des pratiques futuristes où, à l'instar de ce qu'anticipait Isaac Asimov de façon quasi visionnaire en 1955, il ne serait même plus jugé utile de faire exprimer aux individus, dans leur diversité, leurs préférences démocratiques, mais préférable de les inférer à partir leurs caractéristiques et de leurs comportements[8].

Et ce petit livre, que m'a donné ma fille ingénieure dans la « *tech* », peut aussi servir de clin d'œil et de *caveat* pour nos réflexions sur les évolutions à venir.

Programme de la Rencontre du Cnis du 2 juillet 2018.

Ouverture de la rencontre par l'allocution introductive de Mireille Elbaum, Présidente du Haut Conseil du financement de la protection sociale

Session 1- De l'enregistrement à la donnée statistique : la quantité fait-elle la qualité ? animée par Gunther Capelle-Blancard, Professeur à l'Université Paris 1 Panthéon Sorbonne

Avec des interventions de Alexis Eidelman, Chef du département des métiers et qualification à la Dares ; Béatrice Sédillot, Cheffe du Service de la statistique et de la prospective (SSP) ; Marie Leclair, Cheffe de la division des prix à la consommation à l'Insee ; et Benjamin Sakarovitch, Membre de l'Unité SSP Lab à l'Insee

Les interventions ont été suivies d'une **table ronde** avec la participation de Paul-Antoine Chevalier (Etalab), Pierre-Philippe Combes (CNRS-GATE Lyon Saint- Etienne), David Cousquer (Trendeo), Sylvie Lagarde (Insee), Michail Skaliotis (Eurostat)

Session 2- Le dilemme entre intérêt général et protection des données privées animée par Antoine Bozio, directeur de l'Institut des politiques publiques

Avec des interventions de Philippe Lemoine, Membre du Collège de la Cnil ; Javier Nicolau, Membre de la Mission accès aux données de santé à la Drees ; Jacques Fournier, Directeur général des statistiques à la Banque de France.

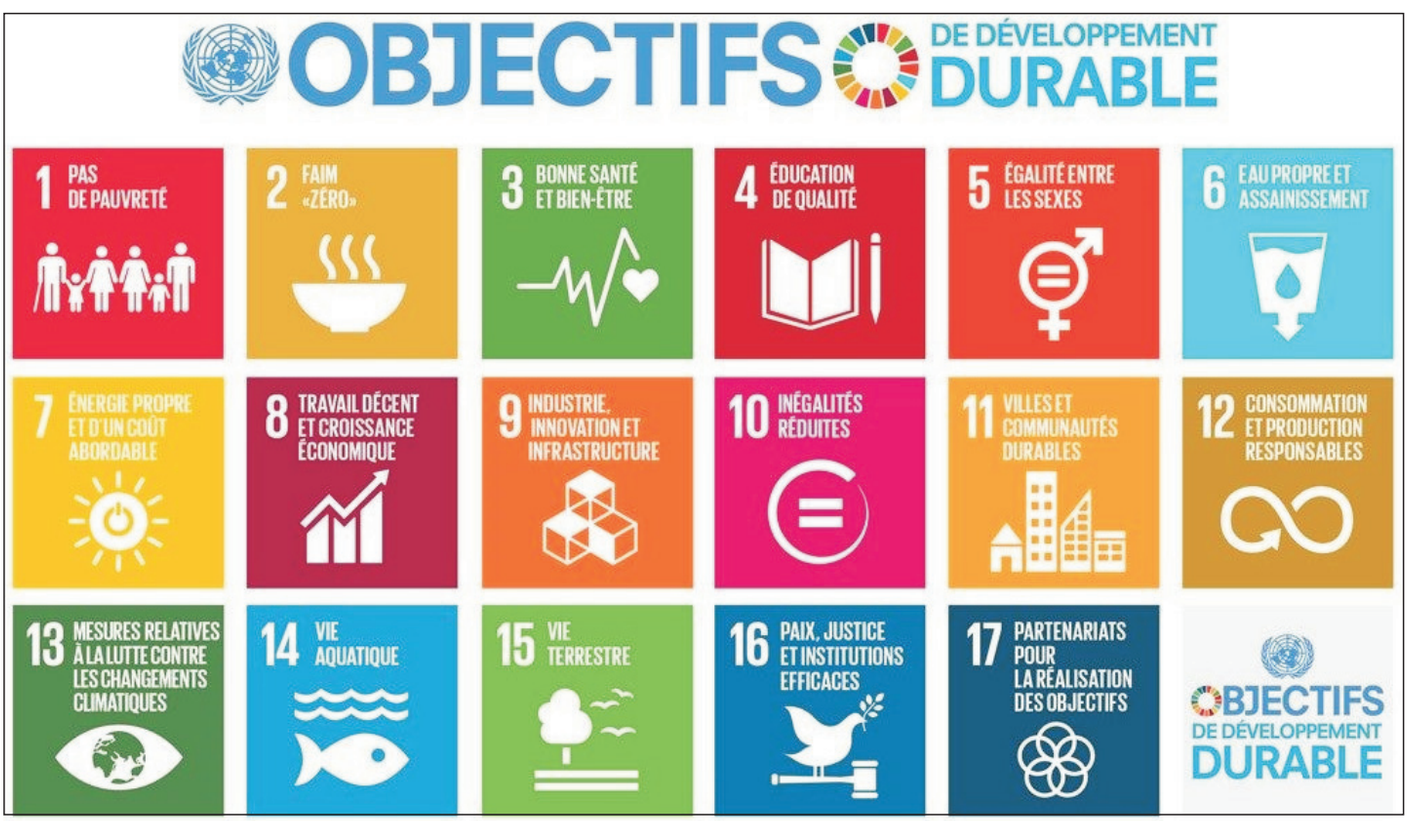
Les interventions ont été suivies d'une **table ronde** avec la participation de José Bardaji (Fédération française de l'assurance), Chantal Cases (Insee), Jacques Fournier (Banque de France), Philippe Lemoine (Cnil), Javier Nicolau (Drees), Bruno Ricard (Archives de France)

Clôture de la rencontre par Jean-Luc Tavernier, Directeur général de l'Insee

La synthèse, le compte-rendu détaillé ainsi que l'ensemble des documents préparatoires de la rencontre sont sur notre site Cnis.fr

8 I. Asimov, *A voté*, Le Passager clandestin, coll. Dyschroniques, 2016.

Conseil national
de l'information statistique



La déclinaison française des indicateurs des objectifs de développement durable

Groupe de travail sous la présidence de Jean-René Brunetière

Pour éclairer les stratégies publiques et privées, il importe de disposer d'éléments chiffrés permettant de suivre de manière objective les objectifs de développement durable dans toutes leurs dimensions. C'est ainsi que l'ensemble des pays de l'ONU s'est entendu sur 169 cibles et 232 indicateurs statistiques, jugés les plus pertinents pour suivre les 17 objectifs de développement durable au plan international.

Le groupe de travail du Cnis propose un tableau de bord de 98 indicateurs statistiques jugés pertinents pour la France et suffisamment pérennes et robustes pour être utilisés jusqu' en 2030 environ en complément des indicateurs internationaux.

Le groupe de travail a réuni une centaine de membres d'horizons très divers (société civile, organisations syndicales, associations, ONG, collectivités territoriales, chercheurs et experts, observatoires, producteurs de statistiques publiques, directions d'administration centrale des ministères ou établissements publics).

Le rapport ainsi que les différents documents produits pendant les travaux du groupe sont disponibles sur le site Cnis.fr

Secrétariat général du Cnis :
Timbre H030 - 88, avenue Verdier, CS 70058, 92541 MONTROUGE Cedex
Tél. : 01 87 69 57 02 - secretariat-general@cnis.fr - www.cnis.fr