Conseil national
de l'information statistique

Montrouge, le 1er octobre 2018 – n°113/H030

Rencontre du 2 juillet 2018

LES ENJEUX DES NOUVELLES SOURCES DE DONNÉES

Centre de conférences Pierre Mendès-France
Bercy - Paris

ACTES DE LA RENCONTRE
« LES ENJEUX DES NOUVELLES SOURCES DE DONNÉES »
- 2 juillet 2018 -

Table des matières

.I OUVERTURE DE LA RENCONTRE

Patrice DURAN, Président du Cnis

Je suis heureux de vous accueillir pour cette rencontre organisée par le Conseil national de l'information statistique sur les enjeux des nouvelles sources de données. Cette rencontre s'inscrit dans la suite logique du précédent colloque sur l'économie numérique et ses enjeux pour la statistique publique. Ces travaux portent sur le positionnement même de la statistique publique. Dans cette période de préparation et de fixation du moyen terme, nous devons identifier les grands enjeux du présent et de l'avenir. Ces travaux permettent aussi d'affirmer le rôle du Cnis en tant qu'interface entre les producteurs et les usagers de la statistique publique et, à travers son positionnement dans le moyen terme, dans le regard que la statistique publique peut porter sur le contexte qui est le nôtre et les problèmes auxquels il faudra faire face.

Le compte rendu complet de la journée précédente sur l'économie numérique est en ligne sur le site du Cnis et un condensé sera disponible dans les prochains jours. L'organisation de deux rencontres relativement rapprochées dans le temps peut surprendre. Nous préparons actuellement le moyen terme 2019-2023. Or pour élaborer celui-ci, il nous faut réaliser une analyse prospective des besoins à cinq ans et identifier les changements qu'il convient d'apporter au système d'information. Ces rencontres visent donc à nourrir cette dimension prospective du Cnis.

Après la numérisation de l'économie et ses enjeux pour la statistique publique, il est apparu que les réflexions portant sur le système statistique des cinq prochaines années devaient nécessairement intégrer la production en continu du volume considérable des données produites, notamment par les acteurs privés et l'apport de ces sources massives nouvelles pour la statistique publique. Dans son dernier rapport, Henri Verdier souligne que la donnée n'est plus pensée comme un simple outil, mais comme le principal carburant de l'action publique. Cette évolution soulève la question essentielle de la fiabilité et de la disponibilité de la donnée. Big data ne va pas forcément de pair avec rich data, car les bénéfices escomptés ne sont pas toujours très clairs et les défis techniques et méthodologiques se révèlent très importants. C'est bien pour cela qu'il faut saisir tout à la fois la portée pratique, politique, technique et scientifique du phénomène que nous vivons aujourd'hui avec l'explosion de ces données numériques.

Dans ce contexte de profusion des données, le positionnement de la statistique publique se trouve au cœur de nos réflexions. L'expansion considérable des données et de leurs usages appelle d'autant plus à clarifier le rôle et la responsabilité de la statistique publique. Quel usage peut-il être fait de ces données massives par la statistique ? En quoi les pratiques du système statistique public seront-elles modifiées ? Quels sont les obstacles à l'utilisation de ces données et quelles précautions d'usage doivent être respectées ? Telles sont quelques-unes des questions qui seront abordées aujourd'hui. La réponse est à la fois technique, éthique et juridique. Ces dimensions vont structurer notre réflexion, puisque la session de ce matin sera centrée sur la qualité des données. La session de l'après-midi abordera quant à elle les questions éthiques posées par l'utilisation des données massives et leur inscription dans le droit. Dans la mesure où il s'agissait de réfléchir sur la statistique publique, nous ne pouvions qu'inviter Jean-Luc Tavernier à intervenir dans une conclusion en forme d'introduction, puisque les sujets que nous traitons concernent le futur.

Nous allons accueillir immédiatement Mireille Elbaum que je remercie d'avoir accepté cette tâche toujours difficile d'ouvrir un colloque. Spécialiste des questions sociales, elle a été titulaire de la chaire « politiques et économie de la protection sociale » au Conservatoire national des arts et métiers. Elle a également été directrice de la Drees (Direction de la recherche, des études, de l'évaluation et des statistiques). Aujourd'hui, elle préside le Haut conseil du financement de la protection sociale, et a été nommée membre du Conseil consultatif européen pour la gouvernance statistique (Esgab).

Je vous remercie et vous souhaite une très bonne journée.

.II QUESTIONS INTRODUCTIVES

Mireille ELBAUM, Haut Conseil du financement de la protection sociale

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, comme l'ont montré les travaux d'Alain Desrosières, 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 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), 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 pour 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.

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 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.

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 pré-construits et pré-formatés. Ce processus est au cœur même de l'indépendance du système statistique, et, comme l'a noté Jacky Fayolle, de la résistance opposée par les statisticiens à ce qu'Alain Supiot 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.
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 ; 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 ?
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 ?
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 de leurs caractéristiques et de leurs comportements.
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.

SESSION 1 - DE L'ENREGISTREMENT A LA DONNÉE STATISTIQUE : LA QUANTITÉ FAIT-ELLE LA QUALITÉ ?

Gunther CAPELLE-BLANCARD, Université Paris 1 Panthéon Sorbonne

Je remercie le Cnis d'avoir organisé cette rencontre. Le big data entraîne un grand changement pour les producteurs de données, les entreprises, les chercheurs. Ce sujet clé recouvre un nombre de thèmes incroyable et soulève des problèmes à la fois techniques, économiques, technologiques, scientifiques, de gouvernance, de déontologie, etc. La matinée s'organisera en deux temps. Tout d'abord, nos intervenants nous livreront leur témoignage dans différents domaines, puis nous échangerons de manière très interactive.

Quel apport des données du web pour connaître les offres d'emploi ?

Alexis EIDELMAN, Direction de l'animation de la recherche, des études et des statistiques (Dares)

Je vais effectivement vous livrer un témoignage très concret de l'activité que nous menons au sein du département Métiers et qualifications de la Dares sur les offres d'emploi en ligne. Historiquement, nous utilisions les offres diffusées par Pôle emploi. Or le monde a changé et il existe de nombreux acteurs de diffusion d'offres d'emploi, notamment lebon coin.fr pour les offres d'emploi peu qualifié, RégionsJob, Viadeo, etc. Certains acteurs apparaissent, d'autres disparaissent ou fusionnent. Tous n'ont pas le même positionnement ni le même fonctionnement.

Ces données permettent-elles de remplacer des enquêtes, notamment l'enquête sur les emplois vacants (Acemo) ? Cette enquête européenne nous permet d'identifier le nombre d'emplois vacants avec une définition établie au niveau européen. Un groupe de travail européen a réuni différents pays pour travailler sur le scraping des offres d'emploi en ligne et examiner la possibilité de remplacer l'enquête par ces données. Or sa conclusion apparaît décevante : l'exercice se révèle difficile, car seule une partie des offres d'emploi vacant se traduit par une offre d'emploi en ligne. Surtout, la définition de l'emploi vacant ne correspond pas à la diffusion d'une offre en ligne, ce qui entraîne une dissonance entre les deux sources de données. Dans ce cas, ces nouvelles sources ne permettront pas de se passer d'une enquête, en particulier pour obtenir des informations de cadrage un peu macroscopiques. Pour autant, nous ne laissons pas ces données en ligne de côté, car elles permettent de compléter les données d'enquêtes, notamment grâce à leur plus grande réactivité. Ainsi, outre le point annuel réalisé par l'enquête, nous pourrons disposer de points trimestriels. Nous devrions également gagner beaucoup en représentativité. Nous pourrons contrôler les biais via l'enquête et diffuser des données plus fines au niveau local. Outre la quantité, la précision et la vitesse, ces offres en ligne présentent un intérêt en termes de contenu. Aujourd'hui, nous exploitons le texte des offres d'emploi pour repérer les compétences qu'une enquête ne permet pas forcément d'identifier. Nous pouvons ainsi identifier si les personnes doivent détenir un permis de conduire ou connaître une technique particulière par exemple pour accéder à l'emploi proposé.

Nous ne pouvons pas remplacer une enquête par une nouvelle source, mais pouvons-nous travailler sans les données collectées sur le web ? Nous pensons, à la Dares, qu'un tel exercice se révélerait difficile. En 2017, nous avons constaté une augmentation spectaculaire des offres d'emploi collectées par Pôle emploi au dernier trimestre, surtout dans le domaine du BTP. Or cette augmentation ne marque pas seulement une reprise de l'emploi. Le site leboncoin.fr venait de changer de modèle économique, rendant payante la diffusion d'une partie des offres, et en particulier des offres d'emploi. Dès lors, la plupart des entreprises qui publiaient leurs offres dans le bâtiment ont choisi de revenir vers le canal gratuit de Pôle emploi. Conclusion : si nous ne captons pas ces données, nous risquons de créer des biais dans l'observation.

L'utilisation de ces données présente néanmoins de nouveaux enjeux techniques. Devons-nous récupérer ces données en concluant des conventions avec les acteurs ou en recueillant les données directement sur leur site par des systèmes automatiques ? Nous avons, comme d'autres pays européens, choisi la seconde voie du scraping, qui ne repose pas sur la bonne volonté de l'opérateur privé qui serait en situation de pouvoir nous manipuler. En travaillant au niveau européen, nous avons constaté que la France était moins exposée à des problématiques de langue ou d'alphabet comme la Belgique ou la Grèce. En revanche, nous sommes confrontés à un problème de déduplication. Rien n'empêche en effet qu'une offre soit publiée sur plusieurs sites. Cette opération implique des calculs lourds ; elle représente un véritable enjeu, d'autant que les employeurs n'écrivent pas forcément la même chose sur les différents sites. La recherche des compétences dans le texte de l'offre exige aussi une expertise dans le traitement du langage et donc, pour nous service statistique, de recruter aussi de nouvelles compétences. Enfin, nous avons connu la semaine dernière notre première contrainte de volume, mais l'augmentation de notre capacité à 1 To devrait nous permettre de passer l'année sereinement. Le code Rome (Répertoire opérationnel des métiers et des emplois), la nomenclature des offres d'emploi définie par Pôle emploi soulève aussi une question statistique intéressante. Les offres diffusées sur leboncoin.fr ne sont pas associées à un code Rome. Dans notre diffusion, nous affectons ce code aux offres par un travail de classification supervisée. Pour cela, il nous faut des données de Pôle emploi, ce qui ne va pas sans poser de questions techniques. Nous allons pouvoir bien classer les offres qui ressemblent à celles de Pôle emploi, mais moins bien les autres.

Les données ouvertes sur le marché du travail existent et de nombreux acteurs privés s'en saisissent aujourd'hui pour réaliser des outils de visualisation et de restitution. Elles peuvent se révéler parfois meilleures que les nôtres. Ces acteurs, parfois moins vigilants sur les problèmes de déduplication ou de classification, diffusent leurs produits et certaines collectivités locales les achètent ce qui n'est pas sans poser des questions sur le positionnement du service public et la nécessité d'utiliser ces données.

Données satellitaires et mesure de l'occupation des sols : usages actuels et perspectives

Béatrice SÉDILLOT, Service de la statistique et de la prospective (SSP)
Mon propos portera sur les enjeux pour la statistique agricole de l'utilisation des données satellitaires, qui ne sont ni des données administratives ni véritablement des données privées, puisque nous pouvons accéder à certaines d'entre elles de façon gratuite. Disponibles en très grande quantité, ces données doivent néanmoins être interprétées pour pouvoir être utilisées. J'évoquerai également l'un de nos projets sur l'occupation des sols.
Aujourd'hui, on peut identifier trois grands types de domaines de la statistique agricole dans lesquels les données satellitaires pourraient être potentiellement utilisées :

- l'occupation et l'usage des sols ;
- la pousse des prairies (indicateur utilisé pour les bilans fourragers et la mesure des calamités agricoles) ;
- les surfaces et rendements des terres labourables.

Je parlerai ici des réflexions que l'on conduit actuellement sur la mobilisation des données satellitaires, en complément d'autres sources, pour la mesure de l'occupation et l'usage des sols car c'est le domaine où les réflexions sont les plus avancées.
S'agissant de l'occupation et de l'usage des sols, nous menons depuis la fin des années 1960 l'enquête Teruti qui mesure annuellement l'occupation physique des sols et l'utilisation des terres. Cette enquête permet de bien décrire les différentes utilisations du sol et d'apprécier avec un pas annuel les changements d'occupation. Nous pouvons ainsi observer les phénomènes d'artificialisation, de déprise agricole et de disparition éventuelle des forêts, jusqu'à un niveau départemental.
Depuis 2017, nous avons revu la conception de ces enquêtes en enrichissant le nombre de sources mobilisées pour qualifier les points afin de tirer au mieux parti des données disponibles et limiter le nombre de déplacements sur le terrain. Une grille couvrant l'ensemble du territoire avec un pas de 250 mètres a ainsi été établie et on a cherché à qualifier les neuf millions de points qu'elle comporte en mobilisant les couches géographiques disponibles, notamment celles produites par l'IGN (Institut national de l'information géographique et forestière) dans le cadre des observations aériennes de la BD Topo (description vectorielle 3D) et les données administratives de la PAC (Politique agricole commune) avec le registre parcellaire graphique qui isole, pour les bénéficiaires des aides, l'endroit du territoire où leurs cultures peuvent être éligibles à ces aides.
Lorsque les points ne peuvent être qualifiés à partir de ces sources, nous recourons à une collecte terrain ou à la photo-interprétation pour les points difficiles d'accès en procédant alors par échantillon. Aujourd'hui, 80 % des points peuvent être imputés à partir de sources géographiques (surfaces en eaux, sols revêtus non bâtis, sols bâtis, sols cultivés déclarés à la PAC et massifs forestiers). 4 % des points peuvent être photo-interprétés et nous en tirons un échantillon de 15 000 par an. Enfin, 16 % des points ne peuvent être qualifiés : il s'agit généralement de points sortis des bases administratives de la PAC, ou se trouvant dans une tâche urbaine un peu dense ou à proximité des forêts. Nous tirons un échantillon de 70 000 de ces points par an pour les enquêter sur le terrain. À l'issue d'une première année d'enquête, nous disposons d'une estimation au niveau national. Il nous faut cumuler trois années d'enquête pour obtenir une bonne précision au niveau départemental. Trois échantillons sont donc successivement tirés et chaque point est revisité au bout de trois ans.
Les données satellitaires peuvent, dans ce cadre, avoir deux types d'usages. Tout d'abord, avant d'aller sur le terrain, les enquêteurs disposent de plusieurs informations sur le point à qualifier : une photographie aérienne de l'IGN avec une précision de 50 centimètres mise à jour tous les trois ans ; une photo satellite issue du dispositif Spot, avec une précision un peu moins bonne (1,5 mètre) mais mise à jour annuellement. Dans certains cas, l'utilisation de ces photos suffit pour qualifier le point et évite un déplacement sur le terrain. Le deuxième usage envisagé est plus nouveau. Il consiste à expertiser les mesures de l'occupation des sols réalisées par modélisations à partir de la télédétection automatique pour les mobiliser éventuellement ultérieurement dans le cadre du dispositif Teruti. Nous travaillons avec le Centre d'études spatiales sur la biosphère de Toulouse (CesBio), unité mixte de recherche de l'Université Paul Sabatier (Toulouse III), du CNRS (Centre national de la recherche scientifique), du Cnes (Centre national d'études spatiales) et de l'IRD (Institut de recherche pour le développement). Cette unité a développé une couche OSO (couche d'occupation des sols) obtenue par télédétection automatique à partir d'images satellites Sentinel 2, des images européennes disponibles gratuitement dans le cadre du dispositif Copernicus (programme européen de surveillance de la terre).

Aujourd'hui, cette couche couvre la France entière avec 17 classes d'occupation des sols et le Cesbio travaille actuellement sur la qualité du classement opéré à partir des données satellites. Afin de vérifier les écarts de classement entre notre enquête Teruti et cette source de données OSO, nous avons construit ensemble une grille pour comparer les nomenclatures issues des deux sources, détecter les zones de cohérence et examiner plus précisément les anomalies. Ce travail partenarial a pour objectif de valider la qualité du modèle du Cesbio et de vérifier s'il est possible d'imputer certaines occupations du sol lors du renouvellement des échantillons de l'enquête ou lorsque nous revisiterons les points au bout de trois ans. La modélisation serait alors utilisée comme détecteur du changement et permettrait de cibler les déplacements d'enquêteurs sur les points pour lesquels il y a une forte présomption de changement d'occupation.
Ces travaux devraient se concrétiser au cours de l'année 2019 et pourraient s'inscrire dans le cadre de l'ESSnet Big Data II que va lancer Eurostat.
En définitive, nous retrouvons dans le domaine des données satellitaires, certains des enjeux évoqués en introduction de ce séminaire, en particulier les enjeux de compétences qui se posent ici de façon particulièrement nette. En interne, nous ne sommes pas formés et outillés pour interpréter seuls les données auxquelles nous pouvons accéder. Il importe donc de nouer des partenariats avec les acteurs publics ou privés, d'expertiser ce que l'on peut attendre de ces sources (avec souvent des enjeux de cohérence des nomenclatures) et d'évaluer le coût de ces opérations pour le système statistique. Nous nous trouvons aujourd'hui à cette croisée des chemins et les travaux que nous lançons visent à améliorer progressivement notre expertise dans ce champ.

Gunther CAPELLE-BLANCARD
Nous le voyons bien à travers ces deux premiers exemples, le big data a une actualité très concrète ; nous nous servons d'ores et déjà au quotidien de ces données. Un nouvel exemple va nous être présenté avec les données de caisse.

Données de caisse et webscraping : de nouvelles sources pour mesurer les prix à la consommation

Marie LECLAIR, Institut national de la statistique et des études économiques (Insee)
L'indice des prix à la consommation (IPC) mesure l'évolution des prix à la consommation à qualité constante. Pour ce faire, nous suivons tous les mois les prix d'un panier fixe de produits, que nous renouvelons annuellement pour préserver la représentativité de la consommation des ménages. Si un produit disparaît du panier au cours de l'année, nous le remplaçons par un autre et nous procédons à un ajustement de qualité. Aujourd'hui, le calcul de l'IPC repose essentiellement sur des prix relevés par des enquêteurs de l'Insee dans des points de vente physiques. Nous procédons à près de 200 000 relevés que nous complétons par 190 000 prix collectés centralement (relevés sur internet, tarifs, bases de données administratives). Certains prix proviennent aussi de données de transaction ou de collectes automatisées sur internet. Le fait d'utiliser ces nouvelles sources de données n'est pas le seul fait de la France ; tous les statisticiens européens et internationaux se posent la question de ces nouvelles sources et les utilisent déjà largement.
Les données de transaction représentent des données enregistrées lors de la transaction entre le consommateur et le vendeur, qui permettent d'enregistrer le prix, la quantité vendue et un identifiant assez précis des produits. Outre les données de caisse des hyper et supermarchés que nous projetons d'intégrer en 2020, nous utilisons d'ores et déjà les données des pharmacies pour les médicaments. Pour l'instant, la collecte automatisée de prix sur internet (webscraping) apparaît prometteuse pour les « services formulaires ». Lorsqu'il achète un service sur internet, le consommateur remplit un formulaire dans lequel il décrit ce qu'il achète. C'est le cas notamment pour l'achat de billets de transport ou la location de voitures. Le webscraping se révèle également prometteur, avec un peu plus de difficultés cependant, pour essayer de collecter des prix pour les biens, *via* l'aspiration de l'intégralité des prix des produits vendus sur un site. Il existe différentes techniques : soit nous mimons le comportement du consommateur, soit nous interrogeons directement la base de données qui sous-tend le site internet. Aujourd'hui, dans l'IPC, nous utilisons ces données pour le transport aérien *via* l'indice produit par la DGAC (Direction générale de l'aviation civile), mais aussi pour le transport maritime et nous le ferons à l'avenir pour le transport ferroviaire.
Les données de transaction offrent une connaissance très fine des quantités consommées. Aujourd'hui, nous connaissons la répartition de la consommation à un niveau assez agrégé grâce à la comptabilité nationale. Au niveau le plus fin, en revanche, nos enquêteurs fonctionnent par quota, car nous n'avons pas d'idée du poids du produit particulier dans la consommation des ménages. En outre, ces données de transaction sont exhaustives sur le champ qu'elles couvrent et permettent d'économiser en coûts de collecte terrain. Elles soulèvent néanmoins un certain nombre de questions. La loi numérique a permis de répondre à la question de l'accessibilité en permettant d'exiger la transmission de ces données privées à des fins statistiques. Tel est le cas pour les données de caisse depuis un arrêté de mai 2017. Cependant, une transmission obligatoire ne résout pas tous les problèmes. Nous devons quand même nouer un certain nombre d'échanges avec les enseignes. Pour que ces données soient facilement exploitables, il faut d'abord

que les entreprises les centralisent pour pouvoir les extraire à moindre coût et que ces données soient relativement concentrées sur un champ.
Le volume de données à traiter nécessite des architectures informatiques particulières et certains traitements manuels effectués jusqu'à présent par les enquêteurs doivent désormais être automatisés. Il s'agit notamment de s'assurer de la permanence d'un produit donné du panier ou de veiller à son remplacement s'il disparaît. En France, ce problème est rapidement résolu, car nous achetons un dictionnaire de code-barres, mais tel n'est pas le cas pour nos partenaires européens. Enfin, ces données privées ne sont pas construites à des fins statistiques, mais à des fins de gestion et elles peuvent mesurer des éléments différents de ceux que nous souhaitons mesurer. Les promotions peuvent ainsi être enregistrées d'une manière différente de celle qu'aurait choisie le statisticien. Nous avons vérifié ce point pour les données de caisse des grandes enseignes. Pour d'autres types de transaction, en revanche, ce phénomène peut poser problème. Si les remboursements de la SNCF lors des grèves sont enregistrés comme des recettes en moins dans ces données de gestion, par exemple, nous ne pourrons pas construire un indice des prix à la consommation.
S'agissant du webscraping, ces données donnent l'impression d'être facilement accessibles, puisqu'elles sont diffusées sur internet. Elles permettent de bien mimer le comportement du consommateur et d'appréhender les politiques de prix sur internet. Pour les transports, nous constatons d'ailleurs que ces politiques se révèlent assez particulières, avec la volonté d'ajuster les prix en permanence pour que l'offre et la demande s'équilibrent. Or ces politiques de *yield management* posent problème au statisticien puisqu'il n'existe pas un prix unique pour un service. Pour utiliser ces données, il faut collecter une grande quantité de données. Une fois le robot développé, cependant, la collecte d'un plus grand nombre de prix n'engendre pas de coûts supplémentaires.
La technique présente néanmoins quelques inconvénients. Tout d'abord, l'accessibilité des données n'est qu'apparente. Les sites peuvent évoluer en profondeur et il faut alors reprogrammer les robots. En outre, ces données peuvent soulever des questions juridiques. Le webscraping n'affecte pas les droits des producteurs de ces bases, comme pour les données de transaction, puisque ces données sont librement accessibles sur internet. Il s'oppose en revanche aux règles et contrats d'usage des sites internet qui prohibent les robots. L'Insee a choisi d'informer systématiquement les sites de la collecte pour lesquels il entend effectuer une collecte automatisée de prix. Par ailleurs, cette collecte ne nous permet pas d'obtenir d'information sur les quantités. Comme pour les données collectées dans des points de vente physiques, nous ignorons quel site interroger et quel produit enquêter spécifiquement. Ce problème est même renforcé avec le webscraping qui peut donner l'illusion de l'exhaustivité des données. Ainsi, en collectant trop de prix, nous pouvons donner à un produit totalement marginal un poids relativement important. Enfin, il faut automatiser un certain nombre de traitements statistiques (classification des données dans la nomenclature Coicop, identification des produits, remplacement...). Sur les services formulaires et la collecte des données de transport, ces problèmes se posent cependant avec moins d'acuité. Le transport recouvre en effet un petit nombre d'opérateurs et les formulaires nous permettent d'obtenir des données relativement structurées, sur lesquelles nous procédons à un travail statistique assez classique.
Recourir à ces nouvelles sources de données semble très prometteur. L'exercice permet d'être plus précis et plus représentatif de la consommation des ménages, en particulier sur internet. Néanmoins, il pose un certain nombre de difficultés, notamment l'accès sûr et durable aux données. Il nécessite souvent des informations externes pour classer et analyser les données. Enfin, il présente quand même un coût, notamment si nous devons démultiplier les programmes d'exploitation. Pour l'indice des prix à la consommation, nous avons donc choisi la prudence. Nous n'avons pas modifié les concepts de l'indice en utilisant ces nouvelles sources de données, contrairement à certains de nos partenaires européens et nous avons testé et expérimenté avant de mettre en production ces nouvelles sources de données.

Les données de téléphonie mobile pour la statistique publique : un retour d'expérience

Benjamin SAKAROVITCH, Insee – SSP Lab
L'utilisation des données issues de la téléphonie mobile semble également très prometteuse. La littérature comporte des exemples d'exploitation de ces données pour évaluer, dans les pays en développement, les taux d'alphabétisation, la prévalence de la malaria ou la diffusion de l'épidémie, ou même les conséquences de catastrophes naturelles comme les tremblements de terre ou tsunamis.
Les données mobiles recouvrent à la fois le repérage des téléphones *via* l'antenne par laquelle est passé le signal et les contacts éventuels entre usagers. Il existe deux grands types de données mobiles :

- les données actives, qui correspondent à un événement actif de l'utilisateur (appel, message envoyé ou reçu), avec un repérage conservé à un niveau individuel par les opérateurs pour la facturation ;
- les données passives qui enregistrent la connexion d'un téléphone à l'antenne, que ce téléphone soit utilisé activement ou non, collectées de manière plus fréquente, mais agrégée par les opérateurs.

Ces données sont attrayantes, car elles nous donnent accès à une granularité temporelle et spatiale assez inédite pour la statistique publique. Cette source de données a d'ailleurs été identifiée comme prioritaire par Eurostat. Nous participons à un groupe de travail européen sur la mobilisation de cette source dans le but

de proposer un cadre et des outils communs pour l'exploitation, sachant que tous les pays disposent d'accès très différents à ces données.
L'Insee a mené des travaux dans le cadre d'un accord avec un laboratoire d'Orange Labs sur l'exploitation d'un jeu de données précis représentant cinq mois de comptes rendus d'appels pour 18 millions de clients sur l'année 2007. Ce jeu de données a été conservé à des fins de recherche, avec l'accord de la Cnil (Commission nationale de l'informatique et des libertés) et nous n'y accédons que dans les locaux d'Orange Labs.
Nous avons comparé ces données avec les productions de l'Insee pour tenter de retrouver les indicateurs que nous avions produits. Nous avons essayé de reconstituer le zonage en aires urbaines à partir des profils d'appels des antennes ou de détecter le domicile à partir des traces laissées par les individus sur le réseau en comparant cette information à la densité de population résidente que nous fournissent les sources fiscales géolocalisées. Les résultats de ces travaux se révèlent plus ou moins probants. Si nous retrouvons bien les zones urbaines, nous restons bien moins performants dans les zones rurales. Pour la population résidente, nous avons repéré des écarts parfois très importants dans des zones très denses ou, au contraire, dans celles peu couvertes par des antennes.
Il apparaît donc difficile de remplacer totalement les productions de l'Insee par cette source. Ce jeu de données date de 2007. Les usages des mobiles ont changé et les abonnements sont aujourd'hui illimités. Par ailleurs, les antennes ne sont pas distribuées de façon homogène sur le territoire. Or de leur localisation dépend notre précision spatiale. Enfin, nous n'avons travaillé que sur les données d'un opérateur dont les parts de marché sont elles aussi hétérogènes sur le territoire.
Actuellement, nous menons une étude visant à croiser les données mobiles et les données fiscales pour mesurer les phénomènes de ségrégation urbaine. Comme nous connaissons les revenus d'un quartier donné, nous pouvons allouer un revenu probable à un individu de notre base de données mobiles. Cet exercice nous permet de construire des indices de ségrégation prenant en compte la fréquence des appels entre individus habitant dans des quartiers plus ou moins riches. Sur le bassin parisien, par exemple, nos estimations correspondent bien aux faits stylisés connus. Cette étude montre que pour produire une information de qualité, nous avons besoin de croiser les sources. Par ailleurs, il ne s'agit plus seulement d'avoir accès à des données extérieures à des fins statistiques. Si nous voulons bénéficier d'un accès pérenne, nous devons partager le traitement avec l'opérateur, ce qui exige d'aller plus loin dans la coopération. Or l'idée de traitement partagé pose la question du contrôle de la qualité des indicateurs.

Échanges

Gunther CAPELLE-BLANCARD
Dans cette première partie, nous vous avons présenté des initiatives très positives d'expérimentation de ces nouvelles sources. Dans la table ronde, en revanche, nous nous montrerons un peu plus critiques.

Mireille ELBAUM
La dernière expérience me paraît très intéressante. Je m'interroge néanmoins sur la façon d'attribuer le revenu médian de la zone à une personne. Est-il possible d'aller plus loin en modélisant les revenus pour parvenir à une estimation individuelle plus fine ? Ces données de téléphonie soulèvent également une question de fond, car vous en déduisez des informations sur les interactions sociales. En fonction du nombre et du type d'appels passés par les personnes, vous décrivez des interactions sociales uniquement vues d'une façon quantitative. Or nous ignorons le contenu de ces appels.

Benjamin SAKAROVITCH
Des travaux sont en cours pour allouer un revenu de façon individuelle à partir de la distribution du revenu que nous connaissons dans chaque quartier. Ils soulèvent néanmoins des questions sur la collaboration avec Orange. Nous avons besoin d'amener des données de l'Insee dans l'infrastructure d'Orange, tout en préservant le secret statistique, pour pouvoir simuler ce revenu à un niveau individuel. S'agissant des interactions sociales, nous ne mesurons que le nombre d'appels entre les abonnés Orange, ce qui présente un certain nombre de biais. Il s'agit donc d'une étude exploratoire parcellaire.

Nila CECI-RENAUD, Dares
Sur l'utilisation des données satellitaires, vous avez évoqué un acteur privé qui convertit ses données pour qu'elles soient utilisables dans le cadre de la statistique publique. Existe-t-il un risque que votre partenaire privé se substitue totalement à l'enquête que vous cherchez à enrichir à l'aide de ces données ?

Béatrice SÉDILLOT
Pour l'instant, nous menons nos travaux avec un laboratoire de recherche, le CesBio. Nous avons eu l'occasion d'échanger avec d'autres acteurs comme Airbus Défense & Space qui a constitué des laboratoires de recherche dans ces domaines et souhaitait travailler avec nous pour qualifier ses modèles. Il existe aujourd'hui des outils de différentes natures développés par une pluralité d'acteurs. Certains travaillent en lien avec les compagnies d'assurance pour la mise en œuvre de dispositifs d'assurance

récolte. Si ces dispositifs peuvent proposer des qualifications en apparence concurrentes de celles de la statistique publique, leur usage n'est pas le même. Il importe en revanche pour le statisticien de pouvoir le cas échéant s'en saisir pour faire évoluer ses propres dispositifs.

Jean-Luc TAVERNIER
Lorsque nous avons commencé à parler des vacances d'emploi au niveau européen, les Néerlandais avaient indiqué que le recours ne pouvait pas se substituer aux outils existants pour estimer le volume global des vacances d'emploi, mais plutôt apporter des éléments de répartition sectorielle, voire géographique. Arrivez-vous à la même conclusion aujourd'hui ?

Alexis EIDELMAN
Cela reste un objectif à atteindre, avec la question des compétences. Il existe, dans les offres, des informations que nous ne captons pas dans l'enquête : la précision géographique, le profil recherché, l'expérience attendue. À ce stade, les travaux n'ont pas abouti et l'ESSnet qui a déjà travaillé pendant quatre ans sur le sujet a relancé un nouveau cycle de travail.

François BRUNET, Banque de France
À quel moment avez-vous informé les sites sur lesquels vous réalisez du webscraping ? Quelle a été leur réaction ? Certains ont-ils mal réagi ?

Marie LECLAIR
Dans le domaine du transport aérien, des accords ont été signés entre la DGAC et les sites qui donnent un accès direct à leur API (interface de programmation applicative). Des liens ont été développés entre le prestataire auquel la DGAC a fait appel et ces sites. Sur le transport maritime, nous avons d'abord cherché à concevoir des robots. Lorsque nous avons lancé une collecte régulière, nous avons envoyé des courriers sans obtenir de retour. Nous faisons de même pour le transport ferroviaire. Nous n'avons pas reçu de retour officiel de la SNCF pour l'instant. Lorsque nous réalisons une collecte internet manuelle, nous prévenons aussi le site enquêté, conformément aux bonnes pratiques de la statistique. Parfois, les sites nous bloquent. Or nous n'avons pas toujours d'interlocuteur *ad hoc* dans l'entreprise concernée. Nous n'avons pas senti d'opposition particulière, mais je doute que les entreprises puissent identifier l'Insee parmi tous les acteurs qui effectuent de la collecte internet sur leur site.

Table ronde
Gunther CAPELLE-BLANCARD
J'ai le plaisir d'accueillir Paul-Antoine Chevalier, Pierre-Philippe Combes, David Cousquer, Sylvie Lagarde et Michail Skaliotis. Cette table ronde se veut interactive, voire provocatrice dans certains cas, pour approfondir ces questions de big data. Rappelons tout d'abord quelques chiffres. 90 % des données qui existent aujourd'hui ont été créées, paraît-il, au cours des deux dernières années. On dit aussi que 90 % de ces données ne sont pas utilisées. Je retiens que personne n'est capable de citer de sources sur la quantité de données et l'exploitation qui en est faite. L'arbitrage entre qualité et quantité apparaît donc très difficile à faire.
Cette profusion des données pose un problème de gouvernance. L'État producteur des années 1970-1980 s'est progressivement transformé en État régulateur. N'assistons-nous pas au même phénomène dans les données ? Après avoir produit des données, l'État pourrait se concentrer sur leur organisation. Quel est le rôle d'Etalab (mission chargée de la politique d'ouverture et de partage des données publiques du gouvernement) dans ce domaine ?

Paul-Antoine CHEVALIER, Etalab
Historiquement, l'État a constitué le premier producteur de données statistiques, comme l'avait rappelé notamment Alain Desrosières. Aujourd'hui, il existe de nombreux autres producteurs. Je considère qu'il existe trois axes pour qualifier les données. Le premier de ces axes tient à la nature du producteur. Outre l'État, producteur historique, il faudrait citer aujourd'hui le secteur privé mais aussi les communs numériques contributifs, comme OpenStreetMap, Wikidata ou Open Food Facts sur l'alimentation. L'État ne détient plus le monopole de la production de la donnée et il est certain que nous nous dirigeons maintenant vers une régulation de l'usage de la donnée.

Gunther CAPELLE-BLANCARD
L'Insee se voit-il comme un organisateur de la production de données ?

Sylvie LAGARDE, Insee
Comme l'a indiqué Mireille Elbaum en introduction, la production des instituts nationaux de statistiques s'organise autour de l'exploitation de données mixtes, avec des enquêtes, des données administratives et des données massives de différentes natures. La statistique publique joue de multiples rôles. Nous devons

tout d'abord identifier à quel type de problème il nous faut répondre et en déduire les sources à mobiliser pour y répondre. Selon les problématiques, nous utiliserons des sources différentes. Autrefois, nous concevions une enquête pour répondre au concept que nous souhaitions mesurer. Aujourd'hui, nous changeons de paradigme : nous ne contrôlons plus du tout l'amont. Nous partons de données existantes, très diverses, qui n'ont pas été conçues pour répondre à la question posée et nous cherchons la manière de les utiliser. Or lorsqu'il s'agit de partir d'une problématique et essayer d'y répondre de la meilleure façon, le rôle du statisticien me paraît essentiel.

Gunther CAPELLE-BLANCARD
L'Insee représente donc un acteur parmi d'autres. J'éprouve encore des difficultés à appréhender la façon dont la gouvernance va s'organiser. Eurostat a peut-être plus de capacité d'interagir avec les acteurs privés pour assurer cette organisation.

Michail SKALIOTIS, Eurostat
La gouvernance s'opère à différents niveaux, avec une composante nationale et internationale. Nous ne pouvons pas parler d'une gouvernance des données sans clarifier la situation au niveau de la protection de la vie privée et des données industrielles sensibles. Je prendrais comme point de départ le rôle des statisticiens publics. Je pense que ce rôle reste inchangé pour l'essentiel. Les statisticiens publics doivent toujours informer le public sur tous les sujets importants. Il nous faut cependant savoir comment jouer ce rôle. Nous avons une responsabilité envers la société. Face à de nombreux acteurs, y compris privés, nous sommes tenus d'exprimer un avis professionnel et d'indiquer que d'autres acteurs communiquent des informations à la société. LinkedIn, par exemple, a publié des rapports sur l'emploi. Nous sommes obligés de donner notre avis sur ces évolutions.
S'agissant de la gouvernance des données entre les instituts statistiques des pays européens, nous devons déterminer l'organisation à mettre en place dans les différentes phases du cycle de vie des données. Pour les emplois et les postes vacants, je ne crois pas que tous les instituts doivent développer le webscraping, par exemple, car il existe des spécificités liées à la langue notamment. Les instituts statistiques nationaux peuvent encore jouer un rôle de ce point de vue. Nous évoquerons cette question de la gouvernance dans le contexte de notre société numérique de façon très formelle lors d'une prochaine réunion, en octobre.

Gunther CAPELLE-BLANCARD
Trendeo produit des données en dehors de la statistique publique. Souhaitez-vous conserver une grande liberté dans cette production des données ? Verriez-vous au contraire un intérêt dans l'organisation de ce nouveau secteur par les autorités publiques ?

David COUSQUER, Trendeo
Nous avons commencé à collecter des données sur l'emploi et l'investissement en France en 2009, comme de nombreux acteurs le font. Plutôt que de diffuser des articles sélectionnés en fonction de thématiques, nous avons décidé de doter chacun de ces articles d'un certain nombre d'attributs permettant de les agréger ensuite. Cette démarche nous différencie des acteurs qui effectuent une veille presse plus basique et nous rapproche davantage de la statistique. Du fait de ce travail, devons-nous être soumis à un contrôle méthodologique de nos données ? Pourquoi pas ? L'utilisation qui est faite de nos données constitue un autre sujet. Lorsque des journalistes publient nos données, par exemple, je pense que la balle est dans leur camp. Nous pourrions interdire tout simplement la statistique privée, mais je doute que ce soit une bonne solution.

Gunther CAPELLE-BLANCARD
Pouvez-vous nous dire quelques mots sur votre business model ? Un nouveau service se développe et remet en cause le monopole d'État.

David COUSQUER
Nous ne présentons pas notre activité comme de la production de statistiques privées. Nous devons nous assurer de la qualité de ce que nous produisons avant d'affirmer que nous vendons des statistiques. La notion même de statistiques emporte des exigences de qualité, d'exhaustivité et de pérennité des méthodes. Pour l'instant, nous ne sommes qu'un petit acteur et nous découvrons la qualité de nos données en avançant. Avec neuf ans d'existence, nous pouvons quand même constater une corrélation de nos données avec les données trimestrielles et surtout annuelles de l'emploi de l'Insee par exemple. Certaines de nos données correspondent à des faits stylisés. Ainsi, les données que nous enregistrons sur le solaire ont fait apparaître la disparition des emplois dans cette filière en 2011, après la modification des tarifs de rachat.
Dans dix ans, nous pourrons peut-être affirmer que nous faisons de la statistique privée, car nous l'aurons prouvé. Il faut faire preuve de modestie au départ. Nos clients utilisent à la fois des données agrégées et des données granulaires. Ils ne souhaitent pas forcément connaître le nombre d'emplois perdus dans

l'année dans une région, mais plutôt être en mesure d'identifier au jour le jour une usine qui va fermer pour proposer des aides, racheter les locaux, etc. Premise Data s'est lancée dans la statistique privée sur les prix, en particulier dans les pays en développement, avec du webscraping et une méthode de collecte de données assez basique consistant à prendre des photographies dans les étals. Cette entreprise fondée par Google Ventures est composée de personnes comme Larry Summers et Hal Varian qui possèdent les moyens nécessaires pour s'engager directement dans la statistique privée. Nous ne nous trouvons pas dans cette situation, même si nous avons ajouté à notre base de données sur la France, depuis deux ans, une base de données recensant les investissements industriels dans le monde, avec l'idée d'enregistrer au jour le jour les annonces de création d'usines, centres de R&D (Recherche et développement), de logistique ou de production d'énergie et, à terme, de pouvoir nous présenter comme des producteurs de statistiques. En France, de nombreuses fédérations professionnelles réalisent leur propre enquête de conjoncture. Au niveau mondial, en revanche, cette démarche reste moins développée.

Sylvie LAGARDE
Face à ces sources multiples, la statistique publique se positionne pour produire de façon nouvelle. Je ne vois pas la statistique publique évoluer uniquement vers la labellisation de données produites par d'autres. L'intégration de ces différentes sources renouvelle la production de la statistique publique. Nous l'avons bien vu sur les données administratives. D'autres acteurs produisent aussi des données administratives et les mettent à disposition comme la Cnam (Caisse nationale d'assurance maladie) pour les données de santé. La statistique publique joue un rôle spécifique dans ce cadre, puisque ces organismes nous demandent de labelliser certaines de leurs séries pour leur donner un label « statistique publique ». Or nous attribuons ce label à la demande des organismes, ce qui démontre son intérêt. Nous devrons sans doute nous demander si la statistique publique doit, de la même manière, « certifier » les données produites par les acteurs privés, sans que ceux-ci en soient forcément demandeurs. Nous n'avons pas encore suffisamment travaillé la question, mais nous devrons y réfléchir à l'avenir.

Gunther CAPELLE-BLANCARD
Après la gouvernance, je vous propose d'aborder la qualité des données. J'étais assez réservé quant à l'utilisation de ces données, car j'ai, en tant que chercheur, été confronté aux données massives dans le domaine de la finance depuis plusieurs années. Les données de marché sont enregistrées à la milliseconde. Or nous n'avons pas pour autant amélioré fondamentalement notre connaissance des marchés boursiers. Quelle est l'utilité de ces données ? Vont-elles améliorer sensiblement notre perception ?

Pierre-Philippe COMBES, Université de Lyon
Je suis chercheur au CNRS en économie géographique et économie urbaine et je m'intéresse principalement aux inégalités entre territoires. Ce thème est revenu sur le devant de la scène avec la disparition de la taxe d'habitation.
En tant que chercheurs, nous sommes libres de récolter des données là où elles existent, sous réserve de leur qualité naturellement. Nous utilisons de façon très importante les données de la statistique publique. Je rappelle que l'Insee a été le premier acteur à faire du big data. Le panel des DADS (Déclarations annuelles de données sociales) utilisé pour étudier les disparités de revenus nominaux remonte à 1976. À l'époque, nous disposions déjà d'un suivi de 1/24^{e} des salariés du privé. Dans les années 1990, ces données sont devenues exhaustives, pour le secteur privé comme pour le secteur public. Ainsi, depuis plus de vingt ans, nous connaissons les salaires de tous les salariés français. Ces bases de données nous ont permis d'effectuer des analyses très précises sur les disparités de revenus nominaux.
Pour aborder les inégalités des territoires, il faut confronter ces revenus au coût de la vie. Or sur ce sujet la statistique publique peine un peu. Sur le coût du logement, en particulier, les informations restent rares en France. Les notaires étaient très réticents à diffuser les informations dont ils disposaient. L'intervention de l'État va permettre aux chercheurs d'accéder à ces données sur les logements anciens. Les permis de construire nous renseignent quant à eux sur les logements neufs. En revanche, nous n'avions strictement rien sur les loyers à des niveaux géographiques fins. Depuis quelques années, nous utilisons donc des sites web pour obtenir les compléments d'information que la statistique publique ne nous offre pas. De ce point de vue, le big data me paraît très utile en ce qu'il complète des pans entiers d'information. Dans un autre domaine, l'Insee a longtemps refusé de communiquer des indices de prix locaux, faute de points de collecte en nombre suffisant. L'exercice devrait désormais être facilité par l'utilisation des données de caisse.
De la même manière, le big data offre des données sur les aménités de consommation, ce que la statistique publique ne permet pas d'obtenir non plus de façon exhaustive. Nous souhaiterions par exemple connaître le nombre de bars, de squares, d'arbres, etc. à l'échelle du pâté de maisons. L'Insee tient une base de données des biens publics qui recense les écoles et les hôpitaux, mais pas le nombre d'arbres ou de toboggans. Néanmoins, OpenStreetMap, une base de données alimentée par tout un chacun, nous apporte des données complémentaires pour calibrer des modèles économiques et géographiques et réaliser des estimations économétriques. Dans quelques années, si nous arrivons à prendre en compte les différences de revenus nominaux, de coût de la vie (logement et prix des autres biens consommés), et d'accès à toutes

les aménités locales, nous disposerons d'une vue assez précise et fiable des inégalités territoriales grâce à la complémentarité entre données publiques et privées.
Désormais, les chercheurs utilisent aussi le big data pour le passé. Nous avons par exemple travaillé sur les cartes d'état-major du XIX^{e} siècle et nous disposons de 42 milliards de pixels pour la France entière qu'il faut coder en termes d'usage du sol. Nous pourrons même remonter à la carte de Cassini de 1735 et suivre l'urbanisation dans le temps jusqu'à aujourd'hui, grâce aussi à des images aériennes des années 1950. Pour les déterminants des prix des logements actuels, nous essayons d'identifier la façon dont, voilà un siècle, l'occupation des sols environnant les zones urbanisées a affecté les prix. Par exemple, nous pourrons ainsi constater qu'un village entouré de forêts affiche un prix du logement plus élevé qu'un autre village entouré de champs faciles à urbaniser. Cette conversion de cartes historiques en données analysables via des logiciels de statistique constitue un exercice de big data pur.

Gunther CAPELLE-BLANCARD
Je ne doute pas de l'intérêt du big data pour la recherche, mais la question se pose de l'utilisation de ce big data au-delà de la recherche. Etalab est plus orienté vers le grand public. Quel est votre public ? Avez-vous des retours ?

Paul-Antoine CHEVALIER
Etalab a été créé en 2011 pour concevoir le portail data.gouv.fr et diffuser l'open data de l'administration française. La mission lui a ensuite été confiée d'aider l'administration à exploiter ses données de manière innovante pour améliorer l'action publique. Depuis 2014, nous accompagnons donc les administrations à utiliser au mieux leurs données. L'administration sait parfaitement produire de la connaissance à partir des données. Elle peut dénombrer les chômeurs, mais elle peine à produire des outils pour aider les chômeurs à retrouver un emploi. Nous essayons d'introduire cette démarche nouvelle à travers nos différents programmes d'accompagnement. Nous nous adressons donc à la fois aux chercheurs, aux citoyens et aux administrations.
Plus la donnée est utilisée, plus il existe de chances qu'elle soit de bonne qualité. Plus elle est exposée à un public, plus il existe de chance qu'elle soit corrigée. Tel est le cas sur OpenStreetMap par exemple. Autrefois, nous vivions dans un monde où l'État produisait des données, vraies ou fausses, et nous ne pouvions rien faire. Désormais, les boucles de rétroaction permettent à tout citoyen de corriger la donnée qui lui paraît erronée. La notion d'usage se révèle essentielle pour penser la qualité de la donnée. 400 000 nouvelles adresses apparaissent chaque année et il s'avère très difficile pour les administrations comme l'IGN, la DGFiP (Direction générale des Finances publiques) ou La Poste de mettre à jour leurs données. Les contributeurs d'OpenStreetMap viennent compléter l'information des administrations et contribuent ainsi à la qualité des données. Nous constatons également ce phénomène sur Wikipédia.

Gunther CAPELLE-BLANCARD
J'entends bien cette idée d'interaction entre l'utilisateur et le producteur de données pour améliorer la qualité, mais je reste persuadé que, pour le grand public, le big data constitue plus une crainte qu'une opportunité. Au-delà des comparateurs de prix, le grand public n'a pas encore appréhendé l'utilité du big data. Quels sont les utilisateurs d'Eurostat ?

Michail SKALIOTIS
Les exemples restent encore rares. Nous avons terminé la première phase d'engagement d'un dialogue collectif avec le système statistique européen. 22 membres ont participé à l'ESSnet évoqué dans la session précédente. Nous n'avons pas d'autre produit en production au-delà des prix. Il existe, en la matière, un long historique. Les producteurs ont en effet accès aux données de scan depuis 20 ans déjà. Nous nous sommes engagés dans une phase de mise en œuvre avec les compteurs intelligents. Dans les prochaines années, nous verrons apparaître des exemples concrets dont les usagers pourront juger de l'utilité. Sur les vacances d'emploi, nous sommes également passés à la mise en œuvre malgré les difficultés que nous avons constatées. Nous suivons les systèmes d'identification automatisés pour les navires avec des informations sur les chargements et d'autres informations importantes qui peuvent être utilisées pour les statistiques. Nous sommes encore en phase de recherche.

Gunther CAPELLE-BLANCARD
L'Insee est-il sollicité par le public sur ces questions de big data ?

Sylvie LAGARDE
Nous le sommes par le Cnis aujourd'hui.
Je souhaiterais revenir sur les questions de qualité des big data. Nous n'avons pas encore établi de cadre d'assurance qualité sur les données massives, car les instituts nationaux de statistiques ne travaillent pas depuis longtemps sur celles-ci. Pour les données administratives, d'ailleurs, nous travaillons encore source

par source, au niveau national. L'exercice apparaît d'autant plus exigeant sur les données massives que celles-ci sont diversifiées et ne soulèvent pas toutes les mêmes problématiques en termes de qualité.
Des travaux expérimentaux sont toutefois menés par les instituts européens sur le sujet. Comme nous ne maîtrisons pas la conception des données, nous devons tout d'abord nous interroger sur l'origine des données, leur producteur et sa pérennité, la structure de ces données et l'existence de métadonnées associées permettant de les comprendre, le respect de la confidentialité des données, etc. Nous devons nous poser des questions sur les données elles-mêmes avant d'aborder leur qualité intrinsèque.
Les données massives posent avant tout un problème de sélection. Nous devons donc nous interroger sur les biais éventuels des statistiques que nous pourrions en tirer. Le premier de ces biais peut tenir à la couverture des données ou au champ de la population d'intérêt. Sur les données de téléphonie mobile, par exemple, si nous ne disposons que des données d'Orange, nous devons déterminer si nous pouvons tirer des informations sur l'ensemble de la population. Les questions de double compte sont quant à elles illustrées par les offres d'emploi. Une même offre peut être publiée sur différents sites, mais pas tout à fait de la même façon. Nous devons donc identifier ce problème et le traiter. À cela s'ajoute la question des concepts ou de la proximité entre ce que nous récupérons dans les données massives en termes d'unités d'observation ou de variables. Sur la téléphonie mobile, un abonné Orange peut couvrir plusieurs utilisateurs, lorsqu'un parent a pris un abonnement pour ses enfants, ou une société pour ses salariés. Nous devons aussi passer d'un utilisateur Orange à une personne résidant dans un lieu donné. Or passer de l'unité d'observation à l'unité statistique n'est pas forcément évident et cela nécessite l'extraction de signaux qui peuvent générer de l'incertitude ou des erreurs.
La pérennité des données et leur stabilité dans le temps constituent d'autres critères de qualité statistique qu'il faut chercher à contrôler. Sur les données d'offres d'emploi par exemple, une multitude de sites voit le jour. Ainsi, Facebook a récemment créé une application permettant aux petites entreprises de poster des offres d'emploi et aux personnes de candidater en ligne. Il faut être capable de donner des résultats comparables dans le temps alors que les dispositifs sont en constante évolution. Enfin, nous ne pouvons pas parler de qualité des données *in abstracto*. Il faut voir cette qualité en fonction de l'usage. Les données actives de téléphonie mobile ont pour objet de mesurer des communications entre les personnes. Il s'agit d'une utilisation assez directe. En revanche, mesurer la population résidente avec ces données représente une utilisation moins directe qui nécessite des travaux complémentaires. La qualité de la source diffère selon l'utilisation que l'on souhaite faire de cette source.

Gunther CAPELLE-BLANCARD
Il existe effectivement un certain nombre de difficultés liées aux données massives. Quelle est, pour vous, la plus grande difficulté ? Pour moi, il s'agirait de l'appariement.

Pierre-Philippe COMBES
Nous allons croiser les sources et les types de données. Les données du recensement carroyées nous permettront justement de redresser les problèmes d'échantillonnage. Je pense qu'il faut absolument pousser cette complémentarité pour tirer parti du meilleur des deux mondes tout en conservant toujours les usages à l'esprit. En tant que chercheur, je ne peux pas communiquer le salaire réel dans une commune de 500 habitants comme pourrait le souhaiter le maire. Mais je veux et peux donner des estimations des effets moyens de telle ou telle variable à partir de sources différentes. Il est certain que les contraintes sur la qualité des données ne sont pas du tout les mêmes. Il importe donc de garder ces usages en tête pour rechercher la meilleure complémentarité entre la production statistique et ces données. L'Insee est un acteur suffisamment important pour se retourner vers les producteurs de données privées et les inciter à modifier la structure de leurs données pour améliorer la qualité de l'information. Je suis assez optimiste sur le sujet.
Par exemple, sur les prix des logements anciens et des loyers, nous ne nous attendons pas à obtenir une corrélation de 1 entre les différentes séries, mais une corrélation négative nous interpellerait. Avec l'Insee, nous partions du principe que les données étaient de qualité. Avec le big data venu du web, nous devons rester attentifs en permanence à ce sujet, mais l'on n'a pas de raison de penser que les données sont nécessairement biaisées.

Paul-Antoine CHEVALIER
Je suis gêné par le terme de données massives, compte tenu de la grande diversité de celles-ci.
Il faut se poser trois questions. Qui est le producteur de la donnée ? Il peut s'agir de l'administration, des acteurs privés ou des communs contributifs. Quel est le mode de collecte de la donnée ? La donnée peut venir de mesures réalisées par des professionnels, du crowdsourcing actif – lorsque l'utilisateur corrige la donnée sur le site, ou du crowdsourcing passif – lié à toutes les données renseignées dans l'utilisation des services numériques. Enfin, quelle est la nature des données ? Pendant longtemps, la donnée était tabulaire. Or les images constituent aussi de la donnée et nous pouvons, grâce au machine learning, en extraire de l'information et la segmenter pour identifier des objets ou des personnes. Le texte et la voix

peuvent aussi constituer des sources de données. Lorsque nous employons le terme de données massives, nous masquons un peu toute cette diversité.
Les décisions de justice sont à la fois massives et non structurées. Il faut apprendre à exploiter cette source pour pouvoir produire des statistiques sur les mesures prononcées par les tribunaux par exemple. Je pense que ces trois questions permettent de mieux s'orienter. L'exploitation de décisions de justice pour en faire de la donnée constitue un exercice bien différent de la récupération des données de trajet sur Google Maps pour en déduire le comportement de déplacement des citoyens.

Gunther CAPELLE-BLANCARD
J'en suis tout à fait d'accord. Les présentations de ce matin ont montré la grande hétérogénéité des données. Mireille Elbaum a évoqué, au-delà des difficultés d'appariement, de fiabilité et de pérennité, le sujet encore plus complexe du phénomène « data driven ». Je citerai une phrase que l'on attribue souvent à tort à Einstein : « *ce qui compte ne peut pas toujours être compté* ». Plus nous disposons d'informations et plus nous entrons dans une illusion de contrôle. Comment appréhendez-vous ce problème d'une recherche un peu trop « data driven » ? Sur les paradis fiscaux, par exemple, il n'existe presque aucune information et je ne vois pas comment le big data peut nous aider.

Pierre-Philippe COMBES
Lorsque j'ai rédigé ma thèse, voilà vingt ans, nous ne faisions que de l'économie théorique. Nous avons commencé à utiliser des données pour tester les modèles. Aujourd'hui, nous nous trouvons dans un monde où les entrées sont directement empiriques. Mais je pense que nous disposons d'un tel volume de données que nous allons justement revenir à l'économie théorique. Sans cela, nous ne seront pas capable de comprendre les corrélations que nous mettrons en valeur. En comparaison des autres sciences sociales, l'économie tire justement sa force de l'économie théorique qui permet de dire ce que l'on explique par quoi. Un débat s'est ouvert sur la différence entre l'inférence, la prédiction et la causalité. Les chercheurs se demandent à l'heure actuelle si le big data peut nous permettre d'obtenir des conclusions en termes de causalité. La politique publique en faveur des quartiers défavorisés permet-elle de réduire le chômage, par exemple ? Sans une utilisation sérieuse de l'économie théorique, nous ne pouvons pas répondre à cette question, et cela restera le cas, voire s'amplifiera avec les données du big data.

David COUSQUER
Nous ne faisons pas de big data. Nous entrons trente informations par jour auxquelles nous accrochons une vingtaine d'attributs ou de métadonnées. Grâce à cette base de données, nous avons observé la remontée des emplois industriels en temps réel, le pouvoir de création d'emplois des entreprises de taille intermédiaire. Nous parvenons à identifier les meilleures régions françaises en termes de productivité et de performance, comme avec les données publiques. Nous pouvons appréhender 40 % de ce qui se passe dans l'économie française en temps réel, avec un va-et-vient entre qualitatif et quantitatif qui me paraît intéressant.
Cet exercice ne s'apparente pas à du big data. Il en est de même pour les 190 000 prix pour l'indice des prix. Le big data constitue l'une des technologies numériques existantes, comme l'intelligence artificielle, les moteurs de recherche, etc. Premise.com combine ces technologies : il collecte les données mobiles et utilise l'intelligence artificielle pour apporter de la pertinence à l'analyse. Nous employons 2,5 personnes pour saisir des données sur l'industrie en France à temps plein et nous pouvons apporter des compléments aux statistiques publiques, mais la statistique publique reste la seule référence. En mobilisant 60 personnes, nous pourrions obtenir la même qualité d'information sur l'économie mondiale. Sur deux ans, nous savons que l'Asie pèse pour 50 % de l'investissement industriel et, derrière ce chiffre agrégé, nous connaissons les noms. Il reste encore des biais, mais ce n'est pas du big data. Le big data vient aussi du fait que les appareils photo et les vidéos ont une meilleure définition. Le volume est accru. Néanmoins, la quantité d'information n'est pas forcément plus importante. Nous sommes fascinés par la quantité des données, mais elle n'est pas nécessairement pertinente.

Sylvie LAGARDE
Je rejoins Pierre-Philippe Combes. Il ne faut pas penser que les données vont parler d'elles-mêmes, mais identifier la question à laquelle elles vont permettre de répondre. Il importe de prioriser les sources nouvelles que nous pourrions exploiter. Nous sommes sollicités pour des exploitations qui peuvent se révéler extrêmement coûteuses. Nous devons nous organiser au mieux pour le faire en fonction des problématiques qui se posent à nous et le Cnis peut nous aider à identifier les sujets. Nous devons également avancer ensemble au niveau européen pour nous répartir les problématiques d'intérêt et les sources à utiliser. Tel est l'enjeu de l'ESSnet Big data.
Les demandes adressées à la statistique publique comportent des contradictions importantes. On nous demande de produire des données de qualité toujours plus vite, à des niveaux de plus en plus fins et comparables au niveau international. Ces données massives donnent l'impression que nous pourrons

répondre à toutes ces demandes, mais il faut nous aider à prioriser les sujets sur lesquels nous devons travailler, en lien avec les autres instituts de statistiques.

Michail SKALIOTIS
Votre question de conclusion est excellente, mais je pense qu'elle ne s'applique pas aux statistiques officielles. Nous partons du principe qu'il s'agit d'une grande base. Or ce serait formidable si cette base existait. Tel n'est cependant pas le cas. L'un des obstacles majeurs consiste à convaincre les unités dans le monde des affaires d'accueillir le big data. Il paraît effectivement nécessaire de fixer des priorités. Notre programme européen est très sélectif. L'occasion m'est donnée de poser la question de l'après big data. J'ai le sentiment que nous allons revenir à une phase de conception de la statistique. Nous avons commencé avec le concept des smart statistiques ou statistiques intelligentes. Lorsqu'il se produit quelque chose, il faut que des statistiques soient construites.
Avec l'instauration d'une protection des données personnelles, nous devons aussi nous emparer de la question de la confiance. Il va exister une « black box » comme dans tous les systèmes intelligents et la situation sera difficile à appréhender pour le citoyen moyen. Dans les mois ou années à venir, il nous faudra travailler sur ce concept des statistiques dans lesquelles nous pouvons avoir confiance. Des projets sont menés actuellement sur le sujet, notamment dans le cadre de l'ESSnet.
Ces nouvelles sources de données constituent un véritable défi, mais elles nous ramènent aussi aux origines des statistiques. Pour que ces systèmes puissent apporter les performances qu'ils prétendent offrir, il faudra en repenser le design. Nous pouvons concevoir des échantillons avec le big data comme nous le faisions avec les enquêtes. Je pense que nous devons vraiment nous poser cette question.

Gunther CAPELLE-BLANCARD
Je propose que nous ouvrions le débat à la salle.

Gérard LANG, Société française de statistiques
Je pense que la statistique publique contribue de manière spécifique à l'organisation et la régulation des données sur deux aspects : les nomenclatures et le triptyque formé par le numéro de Siren (Système d'identification du répertoire des entreprises), le numéro de sécurité sociale et le code officiel géographique. Aujourd'hui, le code des relations entre l'administration et l'État comporte une notion de données de référence avec neuf fichiers, dont le code officiel géographique et le numéro de Siren. Quarante ans après la création de la Cnil, l'objectif de l'Insee en 1973 de faire en sorte que le numéro de sécurité sociale représente le numéro unique d'identification du citoyen dans ses relations avec l'administration est quasiment réalisé avec les données de santé dans la sphère de l'administration. Les deux autres numéros sont également devenus prégnants dans les données privées. Il reste en revanche une certaine résistance vis-à-vis du NIR (Numéro d'inscription au répertoire des personnes physiques).
Dans la période précédente, la production des grands fichiers relevait effectivement de l'État. Je mentionnerai toutefois la première carte de l'organisation du territoire, la carte de Cassini, qui constitue une carte privée, nationalisée par la Révolution en 1790.
J'en viens à l'explosion des données et l'illusion du contrôle. Voilà un mois, le Journal officiel de l'Union européenne a publié un règlement qui donne la liste des données que les établissements de crédit doivent transmettre à l'Agence bancaire européenne pour lui permettre de vérifier la validité de la modélisation interne visant à calculer une estimation des taux de réserve obligatoire au sens de Bâle 3. Ce règlement prend 6 200 pages du Joue (Journal officiel de l'Union européenne). Ce règlement de 6 200 pages met à jour un règlement précédent paru voilà cinq ans de 1 400 pages. Où ce genre de textes nous conduit-il ?

Paul-Antoine CHEVALIER
L'État a un rôle à jouer sur des jeux de données pivots pour décrire l'économie, en particulier le répertoire Sirene (répertoire national des entreprises et de leurs établissements), le code officiel géographique et le répertoire des associations. Plus elles circuleront, plus ces données deviendront des standards de fait pour toute l'économie. L'économie de la donnée peut se structurer autour de ces jeux de données pivots.

David COUSQUER
La question des nomenclatures me paraît importante, notamment en lien avec le big data. Les nomenclatures sont en grande partie conventionnelles. Or je crois que les big data ne savent pas bien jouer avec les conventions. Nous utilisons les nomenclatures de l'Insee pour les secteurs, mais nous procédons à des reclassements. Lorsqu'elles s'installent en France, des entreprises étrangères peuvent créer un bureau commercial. Elles sont alors classées en conseil, alors qu'elles fabriquent des photocopieurs ou des appareils photo par exemple. Nous essayons donc de les reclasser dans leur secteur. De la même manière, les logiciels sont classés à la fois dans les secteurs 58 et 62, selon qu'il s'agisse de l'édition ou de logiciels à façon. Ces conventions sont très mal traitées par le big data. Nous avons également essayé d'utiliser l'intelligence artificielle sur des articles de presse pour les classifier automatiquement. Or l'intelligence artificielle ne classait ces articles qu'avec 80 % de fiabilité lorsqu'ils parlaient de créations ou suppressions

d'emploi et ne pouvait les classer dans le bon secteur qu'avec 60 % de fiabilité. Ces résultats ne sont pas compétitifs par rapport à nos analystes.

Gunther CAPELLE-BLANCARD
Il existe un grand écart culturel difficile à concilier entre l'esprit un peu libertaire du big data et le caractère très normé, encadré et rigoureux de la statistique publique.

Christine CHOGNOT, Union nationale interfédérale des œuvres et des organismes privés sanitaires et sociaux (Uniopss)
J'ai été interpellée par l'intervention de Mireille Elbaum sur le renvoi à la société civile. Comment prendre en compte le citoyen et la société civile ? Dans le champ social, même un expert est confronté à un grand nombre de données difficiles à comparer. N'est-il pas urgent que la gouvernance s'intéresse à ce problème de l'utilisation et la comparaison de données multiples, des controverses qu'elles recouvrent ? Le rôle du Cnis doit-il être accentué dans ce domaine ? De grands médias publient des articles à partir d'une seule source, alors que d'autres sources auraient pu donner des informations différentes. La gouvernance va-t-elle répondre à ce problème pour le débat public ?

Sylvie LAGARDE
Il s'agit d'une question complexe qui renvoie à la multiplicité des données. Je faisais appel au Cnis sur les problématiques prioritaires à identifier. Je pense qu'il revient aux différentes commissions de déterminer, en fonction de ces problématiques, la façon dont le système d'information se construit pour y répondre de la meilleure façon en intégrant les différentes sources de données qui existent. Dans la sphère de la statistique publique, le Cnis doit jouer un rôle de pédagogie dans la construction d'un système d'information statistique le plus efficient.

Roxane SILBERMAN, Représentante des chercheurs au Cnis
Les instituts statistiques ne sont pas tous dans la même situation face à ces nouvelles sources de données. Ceux dont le périmètre est très restreint à de la pure production de données sont plus fortement exposés à cette situation de concurrence que d'autres où, comme c'est le cas pour l'Insee, l'activité d'analyse (comme l'indique son nom) est importante. De leur côté, les données administratives apparaissent comme une source nouvelle dans un pays comme la France où la statistique reposait essentiellement sur les enquêtes alors qu'elles constituent la base de la production statistique dans les pays dit à registre. Ces questions de gouvernance se posent ainsi très différemment selon les situations.

Mireille ELBAUM
J'ai lu ce week-end la position des instituts statistiques européens sur les statistiques après 2020. Les lignes traditionnelles rejetant l'analyse sont en train d'évoluer pour répondre aux demandes des utilisateurs et fournir, face à cette multitude de données, des éléments d'analyse de base. Dès lors, le modèle un peu spécifique de l'Insee, qui pouvait être parfois rejeté pour des questions de principe, semble revenir sur le devant de la scène.
Le rôle d'interaction avec les utilisateurs me semble effectivement très important pour identifier les données qui répondent aux bonnes questions. Nous observons quand même une petite déviation du débat sur les fake news. Les instituts statistiques doivent-ils se positionner sur ces fake news ? Il ne faut pas non plus répondre par un gadget à des questions de fond. Quand les fake news proviennent d'une mauvaise utilisation de données produites par la sphère publique, il paraît très important de réagir. Courir après tout ce qui se dit avec de la fausse statistique produite de tous côtés constitue un autre défi pour les instituts statistiques publics. Où devons-nous nous arrêter ?

Gunther CAPELLE-BLANCARD
Aujourd'hui, c'est surtout la statistique publique qui est malheureusement décrédibilisée. Dans ce contexte, la production de données privées apparaît parfois comme une réponse au monopole de l'État qui manipulerait les données.

Adam BAÏZ, Service de la donnée et des études statistiques (SDES)
Vous avez rappelé qu'il est nécessaire d'éprouver les théories économiques à l'aide de données et d'éclairer ces données massives à l'aide de ces mêmes théories. Dans cette boucle un peu fermée, qu'est-ce qui permettrait de valider la production et l'expansion de connaissance ?

Pierre-Philippe COMBES
Je répondrai la pluralité. Nous disposons d'une palette de modèles économiques pour tester différentes théories. De ce point de vue, les données historiques présentent un grand intérêt. Dans le cadre de notre projet autour des cartes historiques entre 1735 et 2015, par exemple, nous faisons des allers et retours permanents entre la théorie et les données. La source unique d'information et le modèle théorique unique

constituent des risques. En ce sens, le fait qu'il existe une certaine concurrence entre les acteurs privés et les acteurs publics pour la production de données n'est pas forcément une mauvaise chose. C'est grâce à la pression extérieure que les chercheurs ont pu accéder à certaines données. C'est en constatant que s'ils refusaient de nous communiquer leurs données, nous pourrons obtenir des données similaires par d'autres biais, et encore une fois sous pression de l'État, que les notaires ont envisagé une ouverture. La pluralité me semble essentielle, de même que la réplicabilité des études. Toutes les revues de recherche exigent d'ailleurs désormais la mise à disposition des données qui sous-tendent les analyses.

Sylvie LAGARDE
Tous les instituts statistiques s'interrogent sur le comportement à tenir vis-à-vis des fake news. Je doute que nous puissions nous engager dans une course effrénée à la réponse à ces fake news. Nous devons plutôt travailler sur le long terme, vis-à-vis des jeunes notamment, pour qu'ils prennent conscience de la validité des différentes sources de données. Nous travaillons au niveau national, mais aussi au niveau européen et en lien avec le ministère de l'Éducation nationale pour tenter de développer la littératie statistique des jeunes par le biais de moyens de communication adaptés.

De la salle
Les instituts statistiques européens ne devraient-ils pas veiller à la définition dynamique du périmètre des biens communs ? Dès lors que Waze diffuse des données qui permettent aux gens d'embouteiller des quartiers, au regard de la régulation de la circulation dans les villes ces données ne relèvent-elles pas au moins en partie du bien commun ? Comment ces données peuvent-elles être gérées ? Cela vaut pour la circulation, la santé, etc.

Gunther CAPELLE-BLANCARD
Il existe bien évidemment des obstacles techniques à l'utilisation de ces nouvelles sources de données, mais l'obstacle le plus important tient à la gouvernance et la régulation autour des biens communs. Nous avons besoin de cette gouvernance.

Jean-Luc TAVERNIER
Vous évoquiez le modèle économique des sociétés privées qui produisent des données s'apparentant à de la statistique publique. Existe-t-il un grand nombre de sociétés dont le modèle économique permet de concurrencer la statistique publique ?

David COUSQUER
Je peux répondre pour mon cas. Nous sommes à peine rentables et nous avons très peu de clients. Le marché de la statistique privée n'est pas aussi vaste que cela. Certains acteurs produisent des statistiques à des fins de publicité et de communication. Il est vrai que nous bénéficions d'une couverture de presse totalement disproportionnée par rapport à nos moyens et nos revenus. Cette appétence actuelle pour le chiffre est utilisée par de nombreux groupes et nous avons vu passer des statistiques totalement farfelues.
Les grands acteurs de la donnée privée peuvent vendre un produit de conjoncture bien précis, comme Markit et son indice d'activité. D'autres se concentrent sur la donnée d'entreprise, à l'instar de Dun & Bradstreet qui, avec le DUNS Number, possède un équivalent international du numéro SIREN. Les acteurs multinationaux qui constituent un système d'identifiant mondial peuvent percer. Quelques acteurs sont en mesure de bâtir un système de comptabilité économique mondial, mais je m'interroge sur l'intérêt économique d'une telle démarche. Premise.com vient de lancer un indice de prix, mais je doute que l'exercice se révèle très rentable. La société vend plutôt des systèmes d'analyse des données locales. Pour l'instant, la dimension générale qui est le propre de la statistique publique reste peu explorée, mais il n'est pas exclu qu'avec le développement de la technologie certains acteurs se positionnent sur cette niche.

Paul-Antoine CHEVALIER
La statistique privée n'est pas rentable en soi, mais elle pourra constituer un « by product » de business rentables. Facebook dispose par exemple d'une connaissance beaucoup plus fine de l'opinion des personnes que beaucoup d'instituts de sondage.

Gilles SAMSON, Chambre de commerce et d'industrie
La statistique publique devient finalement un utilisateur des données produites par de grandes multinationales dont le business model est orienté vers le croisement de données. Quels sont les moyens pour rétablir les équilibres ?

Jan- Robert SUESSER
J'ai apprécié la façon dont Sylvie Lagarde nous propose de réfléchir à ce débat : à quoi la statistique publique doit-elle naturellement répondre et au-delà, à quoi pense-t-elle pouvoir répondre ? L'utilisation des

nouvelles sources de données est pertinente tant pour les productions classiques de la statistique publique que pour la part que nous avons vocation à couvrir en fonction des débats qui se déroule dans la société.
Sur ce deuxième aspect, je me souviens d'initiatives portées par le milieu de la statistique publique avec la montée d'offres politiques dans des pays développés qui faisaient fi de l'information apportée par les données (i.e. Italie des années 1990). Cela a été à l'origine des grandes rencontres « Measuring the Progress of Societies, World Forum on Statistics, Knowledge and Policy » animées par Enrico Giovannini. Quinze ans après, le phénomène des fake news marginalise encore davantage l'information porteuse de connaissances que produit la statistique publique, sans que cette dernière ait eu les moyens d'y répondre.
Pourtant, s'est également développé le fac-checking, dont nous constatons les forces et les extrêmes limites. Avec qui, quand et comment agir pour éclairer le débat public, dans un climat de plus en plus souvent délétère et violent ? Il s'agit d'une question complexe, mais importante pour la statistique publique dont la mission dans des sociétés démocratiques est d'apporter de la connaissance afin d'éclairer les débats publics. L'adversité ne diminue pas la responsabilité qu'elle porte.
La diversité des producteurs d'informations compétents n'est pas une nouveauté. Je me souviens il y a quarante ans de cela d'études sectorielles détaillées produites par une entreprise privée qui utilisait la statistique publique, les données des entreprises cotées et enrichissait tout cela pour ses utilisateurs. Aujourd'hui, nous sommes confrontés massivement à ce phénomène de multiplicité de producteurs compétents, eux comme nous noyés dans un univers de grand n'importe quoi. Cela nous laisse toujours avec le sujet classique du champ sur lequel la statistique publique doit avoir un apport aujourd'hui.

Sylvie LAGARDE
S'agissant des échanges avec les opérateurs privés qui peuvent être organisés mondialement, nous travaillons pour l'instant avec un laboratoire de recherche au sein d'Orange. Cette démarche me semble intéressante, car en travaillant avec les chercheurs de ces grandes entreprises, nous pouvons peut-être, avec cet allié de l'intérieur, avancer sur les données. Pour l'instant, nous nous sommes intéressés à des données actives anciennes, mais le laboratoire souhaiterait aujourd'hui travailler sur des données plus récentes ou des données passives. Nous devrions donc pouvoir nous appuyer sur eux, voire sur Eurostat – puisque nous avons signé une convention tripartite – pour élargir notre démarche au-delà de la France. Nous pouvons avancer sur le sujet, car nous avons les mêmes intérêts.

Gunther CAPELLE-BLANCARD
Je doute que nous puissions répondre pleinement à cette question des fake news. Je remercie une fois encore les participants de cette table ronde.

SESSION 2 - LE DILEMME ENTRE INTÉRÊT GÉNÉRAL ET PROTECTION DES DONNÉES PRIVÉES

Antoine BOZIO, Institut des politiques publiques
Je suis maître de conférences à l'École des hautes études en sciences sociales, chercheur à l'École d'économie de Paris et je dirige l'Institut des politiques publiques. Cette rencontre constitue un moment privilégié pour débattre entre acteurs mobilisant des données statistiques à des fins de recherche, d'évaluation et de réflexion. Cette session consacrée au dilemme entre intérêt général et protection des données privées comportera deux parties. Après les interventions de nos trois invités, nous organiserons une table ronde pour tenter de résoudre ce dilemme.

Le système national des données de santé (SNDS)

Javier NICOLAU, Direction de la recherche, des études, de l'évaluation et des statistiques (Drees)
Je vous présenterai le contenu de la loi sur l'accès aux données de santé. La loi de santé du 26 janvier 2016 met en place les conditions d'un accès ouvert aux données de santé avec la création du système national des données de santé (SNDS) qui rassemble les différentes bases médico-administratives. Elle crée aussi l'Institut national des données de santé (INDS) pour faciliter l'accès et les traitements et veiller à la qualité des données. Enfin, elle met en cohérence le régime d'autorisation de la Cnil pour toutes les opérations de recherche et d'évaluation.
Le SNDS recouvre cinq flux. Les flux relatifs aux remboursements de l'assurance maladie, aux informations des hôpitaux et à la base des causes médicales de décès sont d'ores et déjà opérationnels. Les flux concernant les données sur le handicap en provenance des MDPH (Maisons départementales des personnes handicapées) et centralisées par la CNSA (Caisse national de solidarité pour l'autonomie) et un échantillon des données en provenance des organismes complémentaires d'assurance maladie sont en cours de constitution. Le SNDS forme ainsi la plus grande base au monde de données de santé. Les données de tous les citoyens y sont déversées, soit plus d'un milliard de feuilles de soins par an et

11 millions de séjours hospitaliers. La base ne comporte ni nom ni prénom, mais un pseudonyme nous permet de suivre toutes les consommations d'un individu donné durant vingt ans.
Les informations sur les patients restent réduites. Dans la base il y a le mois et l'année de naissance, le lieu de résidence, la mise en CMU-C (Couverture maladie universelle complémentaire) ou le diagnostic d'ALD (Affectation de longue durée), ainsi que, le cas échéant, la date et les causes de décès. La base recouvre toutes les prestations remboursées en ville. Chaque passage chez un médecin ou en pharmacie et chaque examen sont enregistrés. Toutes les informations sur les séjours hospitaliers, tous champs confondus, des courts séjours à la psychiatrie, sont également retracées. Chaque passage dans un établissement donne lieu à l'établissement d'un résumé. Sur les pathologies, enfin, le système comprend les diagnostics de mise en ALD et d'hospitalisation, ainsi que les médicaments traceurs des pathologies.
Cette base présente néanmoins des limites. Nous ne connaissons en effet pas les motifs de recours au soin, les résultats des examens, les antécédents familiaux ou personnels. La base comporte également peu de données sociales ou environnementales. Pour réduire ces limites, la loi nous offre la possibilité d'apparier ces données à d'autres bases existantes. Face au manque de données sociales, par exemple, l'Insee et la Drees (Direction de la recherche, des études, de l'évaluation et des statistiques) projettent d'apparier le SNDS avec l'échantillon démographique permanent.
La loi a défini les conditions d'accès à ces données. Ainsi, tout projet envisageant l'utilisation de ces données doit répondre à un intérêt public examiné par l'Institut des données de santé. Des finalités sont interdites : les données ne peuvent servir à la promotion des produits de santé ou la sélection des risques. La loi introduit aussi un principe un peu nouveau de transparence. En contrepartie de l'accès aux données, les personnes doivent publier, à la fin de leur étude, la méthodologie utilisée et les résultats. Enfin, un référentiel de sécurité a été rédigé pour définir la façon dont ces données doivent être stockées.
La loi prévoit deux modalités d'accès et introduit un principe d'open data pour tous les jeux de données qui ne présentent pas de risque d'identification. Les organismes exerçant une mission de service public disposent d'un accès permanent. Cette disposition concerne 100 à 130 organismes (ARS, Santé publique France, Inserm, etc.) qui comptent 2 000 utilisateurs pour lesquels aucune démarche Cnil n'est requise. Le périmètre d'accès dépend cependant des missions de l'organisme concerné. Cet accès comporte aussi des contreparties. Ainsi, les organismes sont censés enregistrer tous les traitements qu'ils réalisent et identifier précisément les personnes habilitées.
Tout autre accès doit être autorisé par la Cnil, projet par projet. Cette procédure, classique dans le milieu de la santé, fait intervenir trois organismes. L'INDS joue le rôle de guichet unique ; il vérifie la complétude du dossier et l'intérêt public de la demande, puis transmet celle-ci au Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé (Cerees) qui donne un avis sur la méthodologie utilisée et le besoin d'accès aux données de santé pour répondre aux objectifs énoncés. Cet avis accompagne le dossier jusqu'à la Cnil qui autorise le traitement en s'appuyant sur l'avis de ces deux organismes. Après six mois de mise en œuvre, une cinquantaine de projets a été acceptée, tant publics que privés, et les premiers projets privés commencent à accéder aux données de santé.

Jean-Louis JANIN, Académie de l'Eau
Quel est l'intérêt du SNDS pour le grand public ?

Javier NICOLAU
La loi définit cinq finalités. Par exemple, ces données peuvent être utilisées par les organismes publics pour améliorer le parcours des patients. C'est sur cette base par exemple que l'Assurance maladie et l'ANSM (Agence nationale de sécurité du médicament et des produits de santé) a mis en évidence les effets néfastes du Médiator sur la santé. L'Assurance maladie les utilise quant à elle pour établir les comptes de la santé. Des projets privés ont également été lancés pour aider les patients à respecter l'observance ou éviter les interactions médicamenteuses. L'utilisation de ces données a donc des retombées directes sur les patients.

Roxane SILBERMAN
L'IDS (Institut des données de santé) précédent n'était pas, me semble-t-il très sensiblement différent dans son rôle et organisation quant à l'examen des dossiers de l'INDS. En quoi la nouvelle institution va-t-elle favoriser un plus large accès aux données ?

Javier NICOLAU
La loi ouvre l'accès aux données de santé à l'ensemble d'acteurs, public ou privés. L'INDS a été créé sur l'ancien Institut des données de santé. La grande différence tient au changement de son périmètre d'intervention. Aujourd'hui, les missions de l'INDS sont élargies. Il est en effet chargé de veiller à la qualité des données de santé, de faciliter l'accès aux données en accompagnant les demandeurs durant le processus d'accès, de vérifier l'intérêt public. Ce suivi complet n'existait pas auparavant.

Antoine BOZIO
Nous accueillons Philippe Lemoine, membre de la Commission nationale de l'informatique et des libertés (Cnil). Essayiste et entrepreneur, il a publié en mai 2018 « *Une révolution sans les Français ?* » et va évoquer la malédiction des données.

La malédiction des données

Philippe LEMOINE, Commission nationale de l'informatique et des libertés
J'avais suggéré une ouverture sur le concept de malédiction des données sur lequel j'ai publié un article dans la revue *Esprit* du mois de juin, car, dans mon optique et celle de la Cnil, il importe de ne jamais adopter une vision seulement instrumentale de l'informatique. L'informatique est reliée à un ensemble d'évolutions économiques et de société.
Les enjeux qui se dégagent autour des applications informatiques diffèrent selon les périodes de l'histoire. Nous avons connu des cycles de 25 ans avec des enjeux mouvants. De 1936 et le premier article de Turing sur le concept de machine universelle à 1960 et la mise sur le marché des premiers ordinateurs, des débats extrêmement aigus s'étaient fait jour sur le ressort de l'intelligence humaine. Entre 1960 et 1984, l'informatique a trouvé son grand domaine de prédilection, l'informatique de gestion. Le débat consistait à déterminer qui sortirait gagnant de la rationalisation très importante des processus : les États comme on le pensait en début de période ou les entreprises. De 1984 à 2008, nous avons connu l'étape des réseaux et de l'informatisation de la société, avec l'affrontement de logiques locales et globales, notamment dans la réforme des grands systèmes sociaux d'énergie, de transport, d'éducation, de santé, etc.
Depuis 2008, nous sommes entrés dans une période de transformation numérique. C'est à cette époque que le terme numérique (ou digital en anglais) est apparu. La mise sur le marché des smartphones et tablettes par Apple a changé véritablement la donne. Un nouvel acteur entre dans la danse et mène le jeu, les personnes. En très peu de temps, deux milliards de personnes s'équipent et changent, inventent, participent à l'innovation. Ce n'est pas Airbnb qui invente la colocation ou Uber qui invente l'autopartage, ce sont les personnes qui inventent de nouvelles façons de produire, consommer, communiquer, se loger. Certaines entreprises récupèrent ces innovations de la société pour en faire des modèles d'affaires. Dans la période actuelle, les grandes organisations, notamment publiques, doivent faire face à ce nouvel acteur, mais aussi à une source de richesse nouvelle, les données. L'équilibre entre les deux représente le grand enjeu de la période dans laquelle nous sommes entrés depuis la fin des années 2000 et la crise économique de 2008.
La malédiction des données représente la malédiction qui attend les entreprises, les organisations, qui attend les nations qui ne voudraient retenir que la richesse des données. Or les données en tant que telles se révèlent moins importantes que leur rapport avec l'humain. La réglementation actuelle vise justement à définir un meilleur équilibre entre les uns et les autres. Nous pouvons parler de malédiction des données au même titre que la malédiction de la rente pétrolière qu'évoquent les économistes. Les pays qui possèdent les gisements de pétrole les plus grands ne sont pas forcément mieux dotés que les autres, car ils souffrent des effets de volatilité, de taux de change, de non-incitation à produire. Ils se trouvent donc handicapés par leur richesse. Il en serait de même pour une nation qui ne croirait qu'aux données sans tenir compte des relations que celles-ci entretiennent avec le reste.
Les anciennes législations ont très fortement évolué en 2018 pour maintenir pour l'avenir un bon équilibre entre les données et les personnes. Dans le paquet de réglementations adoptées au cours des dernières années, vous êtes peu concernés par la directive de 2016 sur les traitements effectués par les services de police et de justice. Vous l'êtes bien plus par le règlement général pour la protection des données du 27 avril 2016, dit RGPD. Ce texte est entièrement fondé sur l'idée de lever le pied sur les mécanismes d'autorisation préalable par les autorités administratives indépendantes comme la Cnil pour renforcer la responsabilisation des personnes dans l'utilisation des technologies de l'information.
Le RGPD renforce le droit des personnes sur leurs données et responsabilise les responsables de traitement. Il s'applique à tous les traitements mis en œuvre par les organismes publics ou privés et à l'ensemble des secteurs d'activité, y compris la recherche et la statistique. Le texte ne constitue pas pour autant une révolution absolue, puisqu'il reprend tous les principes qui existaient déjà en France dans la loi Informatique et Libertés : licéité, loyauté, limitation des finalités, minimisation et proportionnalité entre les traitements mis en œuvre et la finalité, exactitude, sécurité, limitation de la durée de conservation, interdiction de traitement des données sensibles, etc. De nouveaux droits sont également institués pour les individus, dont le droit de portabilité qui permet à une personne d'obtenir le transfert de ses données dans un format réutilisable d'un responsable de traitement vers elle ou vers un autre responsable de traitement. Ce droit nouveau et le droit à l'oubli introduit par la jurisprudence de la Cour européenne de justice viennent s'ajouter aux droits qui perdurent (droit d'accès, droit d'autorisation préalable, droit d'opposition, etc.).
Le règlement renforce les droits des personnes concernées par les traitements. En effet, il indique clairement que le droit s'applique à tous les traitements touchant les citoyens européens, quel que soit le lieu physique de ces traitements, dès lors que ces derniers visent à fournir des biens ou des services aux résidents européens ou à les cibler. Autrefois, nous débattions avec les Gafa (Google, Apple, Facebook et

Amazon) sur l'application du droit européen. Aujourd'hui, le règlement est clair : dès lors que le traitement a pour objectif de vendre des prestations ou d'étudier de façon ciblée les citoyens européens, le droit européen s'applique à l'opérateur, quelle que soit sa nationalité. À cela s'ajoute une augmentation très sensible du plafond des sanctions pécuniaires. Ainsi, les amendes administratives peuvent s'élever de 10 à 20 millions d'euros et, pour les entreprises, jusqu'à 2 à 4 % du chiffre d'affaires mondial, le montant le plus élevé des deux étant retenu.

Le texte impose aussi de nouveaux devoirs pour le responsable de traitement. La logique de responsabilisation se traduit par la disparition de la majorité des formalités préalables, mais aussi par l'obligation de réaliser des analyses d'impact. Dès lors qu'il existe un risque sensible élevé pour les droits et libertés des personnes concernées, le responsable de traitement doit procéder à une analyse d'impact et la Cnil s'est attachée très tôt à diffuser des modèles d'analyse. Ainsi, les acteurs doivent s'interroger en profondeur avant de mettre en œuvre un traitement sur les incidences que peut avoir celui-ci sur la vie des personnes concernées, ainsi que sur les moyens à déployer pour éviter la réalisation de risques identifiés. Cette exigence repose sur un principe de « privacy by design ». Le traitement, dans sa structure même, doit être organisé pour consolider le respect de la vie privée des personnes. Cette obligation d'analyse d'impact constitue l'élément nouveau le plus important. Elle s'applique aussi dans le domaine de la statistique et de la recherche.

Obligation est faite aux autorités et organismes publics de procéder à la désignation d'un délégué à la protection des données (ou DPO), dont les missions vont bien au-delà de celles du correspondant informatique et libertés, puisqu'il est chargé de cartographier les traitements de données à caractère personnel, s'assurer de leur conformité et diffuser au sein de la structure une culture informatique et liberté. Enfin, le texte introduit une responsabilisation de l'ensemble de la chaîne de traitement, puisqu'un sous-traitement de données à caractère personnel peut également faire l'objet de sanctions en cas de violation des dispositions de la loi Informatique et libertés ou du règlement. Ainsi, le texte introduit une notion de coresponsabilité des responsables de traitement avec leurs sous-traitants.

Le RGPD offre cependant des marges de manœuvre aux chercheurs. En dehors d'un point, toutefois, ces principes étaient déjà à l'œuvre dans la loi Informatique et libertés. Le considérant 162 du règlement définit la notion de fins statistiques ; il s'agit de toute opération de collecte et de traitement de données à caractère personnel nécessaire pour les enquêtes statistiques ou la production de résultats statistiques. Dès qu'un responsable de traitement poursuit de telles finalités, il entre dans le cadre particulier suivant. Ces résultats statistiques peuvent être utilisés à différentes fins, notamment des fins de recherche scientifique. Il existe une compatibilité de finalité entre traitement statistique et traitement de recherche scientifique.

Les dérogations au cadre général de protection des données existaient déjà dans la loi française. Le règlement prévoit des aménagements directement applicables par les États membres pour les traitements statistiques ou de recherche. Selon l'article 5B, lorsqu'un traitement a été déclaré et mis en œuvre par le responsable de traitement pour une certaine finalité, le fait de réutiliser les résultats de ce traitement ou les données collectées pour des fins statistiques ou de recherche n'est pas considéré comme un détournement de finalité. L'article 5E confirme la possibilité de conserver les données au-delà de la durée nécessaire au traitement initial. Tout traitement, lorsqu'il est créé, doit être accompagné de la définition d'une durée, mais une prolongation de la durée d'utilité administrative est possible à des fins archivistiques, de recherche scientifique ou historique, ou à des fins statistiques.

L'article 9J précise que les finalités statistiques ou de recherche scientifique peuvent justifier la collecte et le traitement de données sensibles (orientations politiques, sexe, données de santé). Cette collecte constitue cependant un cas typique exigeant la réalisation d'une analyse d'impact. Dans l'affaire Cambridge Analytica, une société qui se présentait comme un organisme de recherche avait passé un accord avec Facebook pour accéder aux données déposées par les internautes et tous leurs correspondants. Aujourd'hui, de nombreux projets de recherche souhaitent utiliser les données des réseaux sociaux. La Cnil a déjà été amenée à indiquer qu'elle était prête à considérer positivement la demande s'il s'agit d'équipes de recherche connues notamment dans le monde de la recherche publique, mais qu'il faudrait systématiquement une analyse d'impact, en particulier lorsque ladite recherche a trait aux opinions politiques, orientations ou inclinations par rapport à l'extrémisme.

Dans la seule mesure où l'exercice de ces droits serait de nature à compromettre gravement la réalisation des objectifs du traitement, d'autres dérogations sont admises. L'article 14-5 du règlement prévoit ainsi que le droit d'information des personnes ne s'applique pas lorsque les données ne sont pas collectées auprès de la personne concernée et lorsque la fourniture des informations se révèle impossible ou exigerait des efforts disproportionnés. Cette hypothèse recouvre la réutilisation des données à des fins statistiques. Des projets de recherche ont également été soumis à la Cnil, comme une étude sur les réseaux sociaux de l'inclination à l'extrémisme des personnes. Il apparaît impossible de procéder à une information préalable des personnes, sauf à priver la recherche de tout intérêt. Dans ce cas, l'intérêt général donne une base légale au traitement. Enfin, la plus grande nouveauté par rapport à la législation précédente est introduite par l'article 17-3 qui prévoit le droit à l'oubli. Celui-ci ne s'applique pas non plus au domaine de la recherche et des statistiques.

Le règlement a prévu par ailleurs des marges de manœuvre que les États sont libres d'actionner dans le droit interne, ce qui n'est pas le cas de la France. À chaque exception au droit des personnes, il est

demandé que des garanties soient mises en place pour respecter tout de même les droits et libertés des personnes concernées. D'une part, le considérant 162 précise que les fins statistiques impliquent que le résultat du traitement ne constitue pas des données à caractère personnel, mais des données agrégées et que ce résultat ne soit pas utilisé à l'appui de mesures ou décisions concernant une personne physique en particulier. D'autre part, l'article 89 prévoit que les dérogations au droit des personnes ne peuvent être envisagées que dans la mesure où sont mises en place les mesures techniques et organisationnelles pour assurer le respect du principe de minimisation des données. Dès lors que la mise en place d'un droit est suspendue, le responsable de traitement doit veiller aux conditions organisationnelles, procédurales et techniques qui assurent la sécurité des traitements mis en œuvre. De ce point de vue, la Cnil s'est toujours montrée favorable à des procédures techniques comme le centre d'accès sécurisé aux données (CASD), une bonne pratique qui permet d'accéder à des données même non anonymisées. Nous essayons d'ailleurs d'étendre ces procédures dans le domaine de la santé.

Gérard LANG
Est-il toujours pertinent et légitime pour la Cnil, 40 ans après sa création, de poursuivre l'une de ses actions fondatrices qui consiste à s'opposer à un projet qui date de 1941 ? Nous avons créé le numéro de sécurité sociale ou NIR pour en faire le numéro d'identification unique des citoyens dans leurs relations avec l'administration. Si je prends à la lettre la loi de 1978 et le RGPD, que je considère la République française comme une personne morale unique et que je m'adresse au Premier ministre pour faire valoir mon droit d'accès et de rectification sur l'ensemble des données personnelles que la République française détient sur moi, que se passera-t-il ?

Philippe LEMOINE
Ce numéro a été mis en place en 1941 et diffusé dans les administrations locales en échange de tickets de rationnement. Pour survivre, les personnes avaient l'obligation de s'identifier. Ce numéro apparaît donc comme une mesure de répression, d'autant qu'il constituait un numéro signifiant. Nous nous trouvons dans un cas de figure dans lequel l'informatique aurait prolongé des procédures totalement perverses. La Cnil a privilégié l'existence de numéros sectoriels. Nous avons cependant étendu le numéro d'identification du répertoire dans le domaine de la santé qui n'avait pas été en mesure, comme les impôts, de constituer un identifiant spécifique. Nous voyons aujourd'hui apparaître, à travers différents dispositifs qui ne reposent pas sur des identifiants signifiants et des systèmes extrêmement souples, la possibilité de mettre en relation des bases les unes par rapport aux autres, de façon ponctuelle et sous le contrôle des personnes, sans comporter les mêmes dangers. J'estime que l'existence d'un fichier unique de population aux Pays-Bas a constitué l'une des sources permettant de décimer la population juive, ce qui n'a pas été le cas en France.

Antoine BOZIO
Je vous propose d'accueillir Jacques Fournier, le directeur général des statistiques à la Banque de France.

Un nouveau gisement pour les statisticiens et les économistes : les données de la Banque de France

Jacques FOURNIER, Banque de France
À l'automne 2016, nous avons ouvert toutes les données de la Banque de France à la recherche, à la suite notamment de travaux du Cnis. Nous avons ainsi ouvert un domaine extrêmement vaste. Nous proposons trois niveaux de publications statistiques dans le domaine financier et monétaire. À un niveau très agrégé, *Les Stat Infos* paraissent tous les deux jours sur la balance des paiements, l'accès au crédit, l'épargne, la conjoncture, etc. à destination du grand public et des médias. Nous mettons par ailleurs à disposition, *via* l'outil WEBSTAT, 40 000 sériés semi-agrégées. Enfin, nous donnons aux chercheurs un accès à 800 millions de séries (qui comportent parfois une seule donnée), par le biais de notre Open Data Room. Nous offrons tout cela à la recherche de façon gratuite, avec un service de soutien.
Les données de la Banque de France concernent tout d'abord le marché monétaire et les transactions interbancaires. Nous gérons d'ailleurs les statistiques de toutes les transactions hebdomadaires sur le marché interbancaire pour le compte de l'Eurosystème. Nous produisons aussi des statistiques prudentielles sur les banques, les compagnies d'assurance, les entreprises d'investissement et tous les intermédiaires financiers. Certains aspects ne sont cependant pas publics, notamment sur la lutte contre le blanchiment. Nous disposons par ailleurs d'informations sur les ménages, puisque nous tenons les fichiers de surendettement, de chèques impayés et de crédits impayés, ainsi que le droit au compte. Enfin, nous avons des informations sur les entreprises. En appui de la cotation, nous tenons un fichier bancaire des entreprises permettant d'effectuer une analyse financière des bilans. Au total, nous disposons donc de 25 milliards de données.
Dans l'Open Data Room, les données sont anonymisées. Nous sommes tenus de le faire, car nous dépendons comme toutes les banques centrales des pays européens d'une réglementation spécifique qui ouvre l'accès à la recherche scientifique pour des besoins de publication académique sous cette réserve.

Nous pouvons apparier des données avec d'autres bases provenant d'autres sources. Nous promouvons l'utilisation du LEI (Legal Entity Identifier), le code mondial d'identification des entreprises qui se développe.
Pour accéder à ces données, le chercheur doit venir dans nos locaux, 37 rue du Louvre. Pour l'instant, nous n'avons pas mis en place d'accès à distance compte tenu des impératifs de sécurité. Cet accès devrait néanmoins être disponible en 2019. Depuis novembre 2016, nous avons accueilli près d'une cinquantaine d'équipes de recherche françaises et internationales. Nous ouvrirons aussi prochainement des écrans à New York pour les chercheurs américains. Le dispositif fonctionne bien. Les équipes de recherche accèdent aux données gratuitement et bénéficient à la fois d'une aide informatique et d'un soutien logistique. Enfin, un professionnel du métier – que nous allouons à cette tâche à temps partiel – permet aussi de s'y retrouver dans ces séries très souvent codées, sans intervenir bien sûr dans les résultats des recherches.

Jean-Pierre JEANTHEAU, Agence nationale de lutte contre l'illettrisme (ANLCI)
Au regard de tous les droits offerts aux personnes pour contrôler les données collectées, avez-vous réfléchi à l'accès au droit ? Aujourd'hui, ce droit est peu utilisé, car les personnes estiment que cette démarche reste trop compliquée. Les sites comporteront-ils un jour un bouton permettant d'actionner ce droit à l'oubli ? Par ailleurs, j'ai compris que ce droit était suspendu lorsque les données font l'objet d'une étude. Les organismes qui détiennent ces données n'ont-ils pas un intérêt à voir se développer de nombreuses recherches pour pouvoir suspendre indéfiniment ce droit ?

Jacques FOURNIER
Nous détenons des données sur les personnes physiques, mais elles n'intéressent pas beaucoup les financiers car ce sont les ménages qui ont des incidents de paiement. Ce fichier dispose d'un statut officiel et relève du cadre exposé par Philippe Lemoine. Les données sont donc conservées pour la durée prévue dans le texte *ad hoc* et les personnes ont accès aux données et peuvent les rectifier. Ce cas de figure se produit assez rarement. Dans la vie réelle, la personne qui a effectué un chèque impayé souhaite avant tout sortir de cette situation en régularisant son découvert au compte, plutôt qu'en vérifiant que la dernière information la concernant est exacte.
Les autres données relèvent beaucoup de fichiers ou concernant des personnes morales. Nous détenons aussi des données sur les chercheurs qui, si vous permettez ce sourire, cherchent assez peu le droit à l'oubli. Si l'un d'eux souhaitait néanmoins accéder aux données le concernant, nous le ferions très volontiers. Nous ne rencontrons donc pas de problème particulier sur le sujet.

Philippe LEMOINE
Il est évident que la mise en œuvre de ces droits suppose un effort de pédagogie très important de la part des autorités administratives indépendantes et des acteurs concernés pour que ces droits soient utilisés. Dans certains pays où il n'existait pas d'autorité administrative indépendante, comme aux États-Unis, les personnes se mobilisaient parfois de façon encore plus forte face aux problèmes de libertés. Lorsque des problématiques aussi importantes que l'affaire Snowden apparaissent, une autorité ne suffit pas ; il faut que les personnes elles-mêmes se sentent concernées. La Cnil peut faire œuvre de pédagogie et demander que dans le design de l'information donnée sur les sites, notamment sur les conditions générales d'utilisation, des consentements spécifiques soient exigés. Il faut en passer par une grille d'analyse qui repose presque sur le design des écrans.

Table ronde

Antoine BOZIO
Outre les trois intervenants de la session, nous accueillons Chantal Cases qui, après avoir dirigé l'Irdes (Institut de recherche et documentation en économie de la santé), est aujourd'hui directrice des statistiques démographiques et sociales à l'Insee. Son expérience au sein de l'Ined (Institut national d'études démographiques), l'Irdes et au Haut conseil de santé publique nous aidera dans ce débat. Nous a également rejoints Bruno Ricard, conservateur général du patrimoine, sous-directeur de la communication et de la valorisation des archives au service des Archives de France et membre de la Commission d'accès aux documents administratifs. Enfin, José Bardaji qui, après un passage à la Dares, à l'Insee et au Trésor, travaille actuellement comme directeur des études économiques et statistiques à la Fédération française de l'assurance. Merci à tous de nous avoir rejoints pour cette table ronde.
La question de la table ronde est largement issue d'un groupe de travail du Cnis faisant suite à la loi pour une République numérique sur l'accès aux données administratives à des fins de recherche scientifique. Ce groupe nous avait permis de dresser un état des lieux du droit et des problématiques pratiques soulevées par cet accès. Un point sous-jacent à cette réflexion portait sur l'arbitrage entre la nécessaire protection des données privées et l'intérêt général à exploiter ces données pour améliorer le bien commun.
Pour amorcer ce débat, je souhaitais revenir sur celui qui s'était fait jour lors de la mise en place de l'impôt sur le revenu en 1914. À l'époque, ce débat ne portait pas sur le caractère progressif de l'impôt ni sur son impact potentiel sur l'activité économique. Il touchait essentiellement à l'accès aux données. L'impôt sur le

revenu nécessite la collecte d'informations au niveau des foyers, ce qui constitue une intrusion intolérable dans notre intimité. Néanmoins, la mise en place de cet impôt sert l'intérêt général, puisqu'elle poursuit certaines visées économiques et sociales. À l'extrême, certains économistes estimaient que la capitation représentait le seul impôt respectant la vie privée. L'accès aux données fiscales avait été limité à l'administration en charge de mettre en place cet impôt, puis des exceptions ont été instaurées par les lois récentes pour donner accès à ces données dans un cadre très sécurisé pour réaliser des exploitations en vue de l'intérêt général. La loi numérique a étendu ce type de dispositif à l'ensemble des données administratives et favorisé la possibilité de mettre en place des appariements entre ces données, renouvelant ainsi la question entre protection des données et intérêt général.
La loi numérique permet de favoriser un certain nombre d'appariements. Quels sont les points de tension entre les possibilités offertes par la loi et les risques ? Comment la loi a-t-elle tenté de circonscrire ces risques ?

Javier NICOLAU
Les données de santé sont régies par la loi Santé. Cette nouvelle loi permet l'appariement des données grâce au NIR, qui peut être utilisé comme toute autre information à caractère personnel. Jusqu'à présent, un décret en Conseil d'État s'avérait nécessaire pour une telle utilisation. Désormais, une autorisation de la Cnil suffit. Deux finalités restent néanmoins interdites et un référentiel de sécurité strict est mis en place pour encadrer l'utilisation de ces données. Des problèmes de gouvernance demeurent, car chaque base est régie par le producteur des données. Pour accéder aux données, il faut donc connaître le responsable de chacune de ces bases, ce qui complique la tâche.
Dans notre démarche pour compléter les données de santé avec l'échantillon démographique permanent, nous constatons que la Cnil fait de la loi une lecture qui n'est pas, à notre sens, compatible avec l'esprit des législateurs. La Cnil, en effet, ne permet pas de constituer d'entrepôt avec les données de santé du SNDS. Ainsi, nous ne pouvons pas réaliser d'appariements pérennes. La démarche doit répondre à une finalité précise et être circonscrite dans le temps.

Antoine BOZIO
Vous avez évoqué le cas du Médiator. Quels sont les potentiels de l'exploitation de ces données dont la finalité d'intérêt général pourrait s'avérer importante ? Quel type d'exploitation pourrions-nous faire de ces données ?

Javier NICOLAU
Les données sont utilisées pour la régulation sanitaire, pour la planification ou le contrôle des dépenses par les établissements de santé. Dans le milieu de l'épidémiologie, elles permettent à Santé publique France, l'Agence du médicament et l'Inca (Institut national du cancer) de décrire l'évolution de la pathologie, des patients, des traitements adaptés. Grâce à ces données, nous pouvons disposer d'une connaissance du milieu de la santé qu'il aurait été impossible d'obtenir jusqu'à présent. La base se révèle coûteuse, mais elle l'est bien moins que la réalisation d'études *ad hoc*. Aujourd'hui, nous pouvons identifier les causes de décès un mois après un infarctus par exemple, ou les chances de réussite d'un traitement après un cancer.

Antoine BOZIO
Pouvez-vous nous en dire plus sur l'appariement des données de santé avec l'échantillon démographique permanent ? Quelles sont les potentialités et les difficultés d'un tel exercice ?

Chantal CASES, Insee
La santé constitue un domaine encore mal documenté en matière de politique de santé et de connaissance de l'état de santé des populations. Nous savons, à travers un grand nombre de travaux partiels, que les inégalités sociales de santé sont fortes en France, et même plus fortes que partout ailleurs en Europe de l'Ouest, mais nous ne pouvons guère aller plus loin. Pour élaborer des politiques de lutte contre ces inégalités de santé, il faut pouvoir les documenter et les évaluer.
Une méthode relativement simple techniquement consisterait à apparier de grands fichiers de données sociales avec les fichiers de l'assurance maladie. Tel était l'objectif de départ de la Drees et de l'Insee. L'échantillon démographique permanent existe depuis 1967. Il s'est enrichi au fil du temps en intégrant, outre les données du recensement et les données d'état civil, les déclarations de données sociales des employeurs, ainsi que les données fiscales et sociales. Cet échantillon couvre un peu plus de 4 % de la population française et comporte toutes les données sociales systématiquement collectées dans le système statistique public, soit à travers des collectes d'enquêtes, soit à travers des données administratives. Ce réservoir de données, qui comprend 3,6 millions de trajectoires de personnes mortes ou vivantes, nous permet de disposer d'un modèle réduit de la population duquel nous pouvons tirer des échantillons pour des études et recherches. Cette démarche est déjà très répandue dans le cadre sécurisé du CASD. Cet échantillon permet enfin au système statistique public de mener des travaux méthodologiques afin, en

comparant les différentes sources de données, d'en évaluer la qualité. Nous travaillons par exemple actuellement sur l'évaluation des doubles comptes dans les recensements.
L'appariement de l'échantillon démographique permanent avec le SNDS, avec la richesse des parcours de soins des populations et un consentement très large à fournir ces données, nous offrirait des informations mettant en regard des données sociales, notamment les diplômes, les catégories sociales et les revenus et des données de santé. Nous pourrions enfin analyser à travers différentes variables sociales les variations des parcours de soins entre les personnes que nous savons très différenciées selon les catégories et les lieux d'habitation, mais pour lesquelles nous ne disposons pas de diagnostic général, et évaluer des politiques qui seraient menées pour lutter contre les inégalités sociales de santé.
Ce projet motivant se heurte néanmoins au fait qu'il existe des silos de données. Les données de santé sont régies par des textes spécifiques tandis que les données sociales relèvent de la loi statistique et de la loi numérique. Or ces silos se parlent difficilement, car le droit d'accès à ces données est organisé de manière différente. Les comités autorisant les chercheurs à travailler sur ces données – comité du secret statistique d'un côté, Cerees de l'autre – sont eux aussi différents. Enfin, les référentiels de sécurité sont spécifiques à chacune de ces catégories de données. Ainsi, pour fonctionner en cohérence, il faut vérifier l'ensemble de ces dispositifs. De plus, nous avons constaté récemment que ce projet d'appariement, que nous souhaitons pérenne pour travailler sur des données au long cours, devient un réservoir de données qui pourrait être redevable d'une nouvelle loi santé. Nous réfléchissons avec la Cnil à la façon de travailler sur ce sujet.
Ce projet nous paraît indispensable pour les chercheurs, mais aussi pour les usagers du système de soin. En matière de santé, la démarche d'arbitrage entre protection de la vie privée et intérêt général s'est révélée exemplaire. Les usagers ont joué un rôle moteur dans ce domaine. Le résultat du premier débat initié en 2010 n'a pas varié au fil des concertations, y compris lors de l'élaboration de la dernière loi santé. Nous avons besoin que ces données soient traitées, car elles nous permettront de découvrir des éléments que nous ne connaissons pas encore, mais elles doivent l'être dans la plus grande sécurité et la plus grande traçabilité pour que les personnes concernées n'y voient aucun inconvénient. C'est un peu dans cette optique que nous souhaiterions réaliser cet appariement. J'espère donc que nous trouverons des solutions pour le faire.

Antoine BOZIO
Les données de santé constituent un bon exemple de cet arbitrage

Philippe LEMOINE
Au sein du collège de la Cnil, nous avons tous des responsabilités sectorielles et la santé ne relève pas de mon ressort. Je vous donnerai donc plutôt mon sentiment personnel. Militant de longue date des sujets informatique et liberté, j'ai participé à la rédaction de la loi de 1978. À chaque évolution de la législation, nous avons ajouté des articles concernant spécifiquement le domaine de la santé et comportant un nombre de cas spécifiques importants. Or cette multiplication d'articles n'a en rien tari l'insatisfaction de la plupart des acteurs.
Le règlement général adopté en 2016 et entré en vigueur en mai dernier constitue un texte d'harmonisation contraignant au niveau européen dont la jurisprudence sera elle aussi harmonisée, puisqu'un mécanisme d'aller et retour est prévu entre les Cnil européennes. Pour autant, l'article 9-4 de ce texte prévoit que les États membres peuvent maintenir ou introduire des conditions supplémentaires, y compris des limitations concernant le traitement des données génétiques, biométriques ou des données relatives à la santé.
De fait, la loi de transposition française maintient l'ensemble des dispositifs antérieurs, avec quelques petits ajustements. Depuis 2006, la Cnil a introduit la notion de méthodologies de référence qui permet, pour les cas qui ne posent pas de problèmes spécifiques, de mettre en œuvre des autorisations simplifiées. Or un projet comme celui-ci ne relève pas des méthodologies de référence. Néanmoins, quelques assouplissements ont été prévus dans le cadre du RGPD. Le règlement pose le principe que si, au bout de deux mois, la Cnil n'a pas pris position sur une demande d'autorisation, cette autorisation est réputée acquise, sauf dans le cas où avait été rendu préalablement un avis qui n'était pas entièrement favorable par les organismes qui jugent de la demande au fond.
Sur le fond, personne ne peut contester qu'il existe un intérêt à rapprocher les bases de santé de données qui concernent la trajectoire sociale et les conditions de vie et de logement des personnes. Ce projet se situe néanmoins très certainement dans un domaine dans lequel les conditions procédurales se révéleront extrêmement importantes. Nous pouvons sans doute critiquer les exigences de la loi Informatique et libertés. Cependant, dans les pays qui n'ont pas adopté une telle loi comme les États-Unis, par exemple, les détournements de données sont considérables. La sophistication des techniques d'intrusion est telle que nous ne pouvons pas écarter totalement ce risque. Je crois que c'est moins sur la finalité même du rapprochement que sur les conditions dans lesquelles il sera effectué, notamment les conditions de sécurité, que la Cnil se concentrera.

Chantal CASES
Nous y sommes prêts.

Antoine BOZIO
Il existe des finalités interdites, en particulier l'exploitation commerciale de ces données. Quelle est la position des assureurs sur ce type de données et sur leur utilisation restreinte à des fins d'intérêt général ?

José BARDAJI, Fédération française de l'assurance
J'ai quand même l'impression que la profession d'assurance est totalement alignée sur le sujet. Le métier d'assureur est marqué par la mutualisation, dans la mesure où il existe un risque qui va toucher un certain pourcentage de la population, et par l'inversion du cycle de production. Le métier de l'assureur se révèle assez spécifique. En effet, nous appelons la prime dans un premier temps et nous indemnisons dans un second temps. Entre les deux, il peut se passer un laps de temps assez important durant lequel l'état de santé de l'assuré peut évoluer.
Pour les sociétés d'assurance, la donnée constitue donc un élément essentiel, indispensable même à l'exercice du métier. Nous utilisons cette donnée pour bien percevoir le risque et déterminer la prime. Lorsque le risque se matérialise, nous intervenons pour indemniser et réparer le préjudice. La donnée sera encore plus importante demain, compte tenu du développement des algorithmes, des objets connectés, de l'intelligence artificielle, etc. Sur le risque de santé, nous utilisons les données non pas pour connaître la personne à laquelle elles sont attachées, mais pour percevoir et quantifier le risque auquel nous faisons face. Nous n'utilisons pas les données de santé pour apprécier le fonctionnement du contrat de l'assuré.
En assurance, nous parlons de dommages corporels, qui recouvrent la santé et la prévoyance (incapacité, invalidité, dépendance, accident corporel, etc.). Il existe des limites. La profession d'assurance est régie par une réglementation importante. Toutes les données relatives à l'état de santé sont couvertes par le secret professionnel. Dès lors qu'une personne souhaite être assurée, nous devons faire intervenir un médecin-conseil ou une personne dûment habilitée. En matière de complémentaire santé, à la souscription, le contrat peut être qualifié de solidaire ou non. Si nous ne prenons pas en compte l'état de santé de la personne, le contrat dit alors « solidaire » est affecté d'une taxe moins élevée (13 % contre 20 % pour un contrat non solidaire). Enfin, l'article 6 de la loi Évin du 31 décembre 1989 interdit de modifier le contrat de santé au cours de la vie de ce contrat, et ce, même si l'état de santé de la personne évolue.
S'agissant de l'accès aux bases de données de santé, la profession de l'assurance bénéficiait déjà d'un accès à des fins d'études. L'individualisation ne nous intéresse absolument pas.

Antoine BOZIO
Nous pourrions décliner cette question avec les données de transactions financières. Quelles sont les exploitations de ces données qui pourraient bénéficier à tous ? Quels sont les risques propres à ce type de données ?

Jacques FOURNIER
Nous veillons bien sûr à protéger les données individuelles, mais je trouve cette distinction entre intérêt général et intérêt particulier un peu artificielle. La situation me paraît en effet plus complexe que cela. Il peut exister une contradiction entre différents aspects de l'intérêt général par exemple. De la même manière, l'intérêt général peut parfois servir un intérêt particulier.
La question s'est posée sur les données prudentielles des banques. Traditionnellement, ces données étaient peu accessibles. Nous pensions que si une information sur une banque ou un ensemble de banques faisait apparaître une difficulté particulière ou laissait planer un doute sur la situation financière de ces établissements, le public retirerait ses dépôts et que, de ce fait, la crise que nous cherchons à éviter surgirait, les banques ne pouvant plus servir les déposants et ne pouvant plus consentir de crédits. À l'inverse, si nous ouvrons les données bancaires au public, même sur un secteur, nous éveillons la prise de conscience, nous permettons des recherches qui peuvent créer une bonne émulation avec des recherches internes et nous favorisons *in fine* une meilleure analyse du cycle économique et financier et donc la stabilité financière. À l'issue de ce débat entre intérêt général et intérêt général, que faut-il faire ? Nous avons considéré que le premier cas de figure restait assez hypothétique et qu'il existait par ailleurs des moyens d'y faire face. Nous avons donc opté résolument pour le second.
De même, l'intérêt général peut très bien servir l'intérêt particulier. Sur les sujets d'inclusion bancaire, nous avons demandé à l'ensemble des banques de repérer les personnes en situation de fragilité financière dans leur clientèle et de nous communiquer des informations sur la tarification qui leur est appliquée. 3,6 millions de personnes ont été identifiées. Pour cela, il a fallu réaliser un recensement individuel des personnes et constituer des fichiers. La question a été posée par certains intermédiaires financiers de savoir si cette démarche ne porterait atteinte à la vie personnelle. Au plan juridique, il n'y a pas de problème de cet ordre, puisque l'exercice est permis par la loi. Ces données nous sont envoyées de manière anonymisée. Nous cumulons ces données et nous en déduisons des préconisations pour aider les personnes en situation de fragilité financière et développer une offre qui leur soit adaptée. *In fine*, nous aidons donc ces personnes dans la gestion de leur budget face à des tarifs aujourd'hui trop élevés. (NB : c'est grâce à ces travaux et à leur analyse statistique que le ministre de l'économie et des finances a annoncé le 3 septembre 2018 un

double plafonnement des frais d'incident pour la population financièrement fragile souscrivant l'offre légale dite spécifique).
La dichotomie entre intérêt général et particulier présente donc des limites significatives. Bien sûr, cela ne veut pas dire qu'il ne faut pas protéger l'intérêt individuel. Pour cela, outre les dispositions juridiques, nous sommes grandement aidés par la technologie d'anonymisation. Voilà encore quelques années, l'anonymisation reposait sur un processus informatique relativement lourd et il fallait beaucoup de temps pour obtenir des résultats. Aujourd'hui, les systèmes de hachage permettent de mettre ces données à disposition très rapidement, tout en préservant la vie personnelle. Enfin, nous assurons la protection de tout ce qui peut être publié.
Nous sommes extrêmement attachés à la protection de la vie personnelle, mais il ne faudrait pas brider de façon un peu excessive les analyses qui visent à protéger chaque membre de la population.

Philippe LEMOINE
Rechercher un équilibre entre accès aux données et protection de la vie privée a beaucoup de sens. Néanmoins, nous tenons beaucoup en France à la notion d'informatique et libertés. Or les libertés dont il est question excèdent largement la protection de la vie privée. Le concept d'autodétermination informationnelle du droit allemand, selon lequel les personnes conservent la maîtrise des informations qui organisent l'évolution de leur vie, a fortement inspiré le RGPD. De ce point de vue, l'accès aux données et les libertés font meilleur ménage qu'on peut l'imaginer.
Dans le domaine financier, outre les évolutions extrêmement positives comme l'Open data room, je pense qu'il existe aussi un débat entre des intérêts privés. Bien souvent, certains acteurs se saisissent de l'argument de défense de la vie privée pour défendre en réalité le pré carré informationnel dans lequel ils se trouvent. La France reste l'un des seuls pays au monde dans lequel il n'existe pas de fichier positif en matière de crédit. Jamais les grands acteurs étrangers du crédit, notamment américains n'ont réussi en France, car ils ont l'habitude de travailler avec un accès très large à l'information. Cette limitation a profité aux acteurs nationaux. La Cnil n'a pas été favorable à la constitution de ce fichier, car il nous a été démontré que la seule manière de le créer consistait à recourir à ce fameux NIR. Cet argument peut prêter à sourire. Nous devons toujours veiller à ce que les textes ne soient pas utilisés pour préserver des intérêts particuliers ou des situations de rente. Tel n'est pas en effet l'objectif de la loi Informatique et libertés. Nous devrions nous montrer moins sourcilleux à l'avenir.

Jacques FOURNIER
Nous favorisons aussi l'accès aux données parce qu'à défaut, cet accès s'opérerait en dehors de la sphère statistique de service public. Les personnes donnent aujourd'hui volontairement –voire involontairement– leurs données à de nombreux organismes privés. Grâce à l'équivalence accordée par le RGPD aux États-Unis, ces données peuvent y être exploitées dans un régime censé équivalent au régime européen. Si nous ne proposons pas une telle ouverture dans la sphère de la statistique publique, nous laisserons des acteurs privés exploiter ces données et nous courrons le risque que la mauvaise donnée chasse la bonne.
Quant au fichier positif, je trouve que c'est une fausse bonne idée et suis donc assez opposé à sa création. Il ne s'agit pas pour la Banque de France de protéger le pré carré des banques ou des assurances. La concurrence bancaire est d'ailleurs extrêmement forte en France. Nous le voyons dans les taux des crédits à l'habitat, qui sont nettement inférieurs au taux d'usure. Mais nous ne souhaitions pas que les ménages relativement aisés soient assaillis de sollicitations pour emprunter, ce qui aurait été la conséquence immédiate d'un fichier positif. En outre, les expériences étrangères notamment en Belgique ont montré que le nombre de surendettés est plus élevé lorsqu'il existe un fichier positif.

Antoine BOZIO
Nous avons parlé à plusieurs reprises du droit à l'oubli et des durées de conservation. Le rapport du Cnis sur l'accès aux données avait abordé les questions d'archivage. Beaucoup d'entre nous pensent que ces questions concernent les historiens, mais l'exploitation dans plusieurs décennies de données produites par ailleurs apparaît importante. Le groupe de travail a constaté, dans l'interprétation des injonctions de la Cnil sur les délais de conservation des données administratives et les conditions d'archivage à des fins de conservation patrimoniale, scientifique ou dans l'intérêt de la Nation, des lectures et des pratiques différentes. Pouvez-vous nous présenter ce constat et l'enjeu de la question ?

Bruno RICARD, Archives de France
De nombreux participants à ce groupe de travail ont découvert que nous pouvions conserver les données à l'issue des délais inscrits dans les délibérations de la Cnil. Pour les archives, la question du droit à la mémoire se révèle essentielle. Elle constitue le cœur de notre logiciel. Nous conservons des documents datant de l'époque mérovingienne et il nous paraît fondamental que les documents que nous collectons aujourd'hui puissent être transmis aux générations futures sur des centaines, voire des milliers d'années. La justification est double ; cette conservation des archives poursuit une finalité historique et une finalité probatoire. Nous pouvons perdre un titre de propriété. Il est important qu'il soit conservé par l'administration.

C'est parce que nous avions conservé après la Seconde Guerre mondiale les dossiers constitués par l'administration de Vichy que nous avons pu indemniser les familles juives spoliées durant la guerre et ré-instruire ces dossiers dans les années 1990-2000.
Le droit à la mémoire constituait, jusqu'à présent, le droit principal. Nous avons cependant assisté à un basculement avec la jurisprudence de la Cour de justice de l'Union européenne, la loi Lemaire et les « directives », qui permettent aux personnes de déterminer le sort de leurs données après leur décès, ainsi qu'avec le RGPD qui renverse le raisonnement. Ainsi, le droit à la mémoire devient dans certains cas un droit d'exception. Les archivistes doivent pouvoir conserver au moins une partie des documents et données à l'issue de la période de conservation dans le traitement initial. Or dans la plupart des administrations, nos interlocuteurs sont étonnés de l'existence d'une disposition de la loi Informatique et Libertés, son article 36, prévoyant la conservation des données au-delà de la durée indiquée par les déclarations/autorisations Cnil, à des fins de recherche historique, scientifique et statistique. Le RGPD a repris cette disposition dans son article 5. C'est sur cette base, par exemple, que nous conservons les recensements de la population française depuis le XIX^{e} siècle, presque sans lacune.
Bien évidemment, ce droit à la mémoire est associé à des contreparties. Le RGPD exige en effet des conditions et garanties appropriées pour déroger au droit à l'oubli. J'en citerai deux qui me paraissent essentielles. Selon le principe de la sélection, nous ne conservons pour l'éternité que ce qui se révèle strictement nécessaire. Nous sommes loin de conserver 100 % de ce qui est produit dans la sphère publique pour laquelle l'administration des archives contrôle les éliminations. À l'issue de la durée de conservation du traitement initial, assimilable à la notion archivistique de « durée d'utilité administrative », nous détruisons 90 à 95 % des dossiers papier et des données. Cela correspond chaque année à l'élimination de 800 kilomètres linéaires de dossiers. Il existe par ailleurs un droit à l'oubli temporaire. Le Code du Patrimoine impose en effet un droit à l'oubli temporaire, lié aux délais de communicabilité. Tout un chacun ne peut accéder à tous les documents. Dans le domaine du secret statistique, ce délai se révèle très long ; il atteint 75 ans et même 100 ans pour les mineurs. Le Code du Patrimoine prévoit toutefois une dérogation : sur justification, après accord du service producteur et avis du comité du secret statistique, les chercheurs peuvent avoir accès à des données encore confidentielles. Cet accès est néanmoins entouré de conditions de sécurité suffisantes, notamment offertes par le CASD. L'économie globale du dispositif nous permet d'avoir atteint un point d'équilibre entre la protection des secrets, le droit des personnes et le droit à la recherche.

Antoine BOZIO
Chantal Cases, vous avez souligné tout à l'heure l'importance de la profondeur des données pour pouvoir apprécier des effets sur le long terme. Pouvez-vous revenir sur ce point ?

Chantal CASES
Dans le domaine de la santé, cette importance se manifeste tout particulièrement pour certaines maladies professionnelles qui se déclarent parfois longtemps après la fin d'activité. Les panels longs permettent aussi de réaliser des travaux intergénérationnels. Pour étudier la transmission de la pauvreté d'une génération à une autre, par exemple, nous devons disposer de données sur très longue période. Ces démarches restent assez classiques. Elles nous permettent d'élaborer, le cas échéant, les politiques nécessaires et de les évaluer.

Antoine BOZIO
Vous êtes-vous interrogé sur les délais de conservation des données ?

Chantal CASES
Nous archivons très soigneusement nos données. Du moins le pensions-nous. La protection du secret représente un peu une seconde nature chez les statisticiens. Or nous avons découvert que nous n'opérions pas toujours comme il le fallait le traitement des données nominatives. Nous archivons nos enquêtes lorsque nous avons retiré l'identité des personnes, mais nous devons discuter avec les archives pour la conservation des fiches adresses par exemple, en dehors de la constitution des panels longs. Nous devons garantir aux enquêtés que la confidentialité est bien respectée. Nous sommes donc preneurs d'un mode d'emploi très détaillé sur l'ensemble de nos productions.

Antoine BOZIO
Dans le cadre du groupe de travail du Cnis, certaines administrations ont admis avoir détruit des données qui n'auraient pas dû l'être au motif qu'elles avaient respecté les injonctions de la Cnil. Comment réagissez-vous ?

Philippe LEMOINE
Au fil des évolutions de la législation et des problématiques, la Cnil doit mener une concertation très étroite avec d'autres institutions, dont les Archives et la CADA (Commission d'accès aux documents administratifs). Nous avons effectivement constaté que, dans la manière de formuler ses décisions, la Cnil pouvait donner l'impression que tout devait être détruit. Or un acte administratif constitue une archive dès qu'il existe. Au terme du délai, il ne doit plus figurer dans les bases actives, mais il ne doit pas être détruit. Quelques habitudes demeurent. Néanmoins, nous restons vigilants. Nous avons pu adopter, par le passé, un langage insuffisamment précis. Le fait de parler de durée de vie d'un traitement n'implique pas la destruction de la totalité des données.
Deux autres questions présentent une grande importance dans ce domaine. Les anciens textes sur les archives instauraient une opposition relativement forte entre la notion d'accès aux données et celle de diffusion et de réutilisation. L'open data, puisqu'il faut publier les données sous une forme aisément réutilisable, donne à penser que les réutilisations sont permises d'avance. Or ce n'est pas parce que le droit organise la publicité des bans des mariages à une certaine fin que n'importe qui peut utiliser cette information pour des publicités pour des listes de mariage par exemple. La notion de diversité d'utilisations et de finalités existe, même dans le monde de l'open data.
À cela s'ajoute la question de l'indexation qui renvoie au droit à l'oubli. Il ne faut pas qu'en tapant un nom sur un moteur de recherche, nous accédions en premier lieu à une information non établie, concernant l'intimité éventuelle de la personne. Le droit à l'oubli ne vise pas à détruire des archives ou faire disparaître des articles de presse, même si ceux-ci sont erronés. La Cnil reste très attentive à ce sujet. Certains cas se révèlent complexes, notamment dans le domaine de la recherche citoyenne. En lien avec les commémorations sur la guerre de 1914-1918, des outils archivistiques avaient été mis en place avec la possibilité pour les personnes d'indexer ou contribuer à alimenter des dossiers numérisés de façon brute. Le cas des fusillés avait alors été soumis à la Cnil. Parmi eux, certains avaient déserté ou commis un larcin, voire un crime. Le fait de pouvoir accéder immédiatement à la totalité de l'information, ou obtenir le nom des membres d'un peloton d'exécution peut soulever un certain nombre de questions pour les héritiers, même cent ans après. Comment créer sur internet un minimum de recueillement dans l'accès à des informations très sensibles sur l'histoire des familles ? Nous n'avons pas encore trouvé toutes les réponses, car il s'agit de problèmes lourds. Il faut accéder à l'information, mais pas de n'importe quelle façon.

Antoine BOZIO
Avec ce groupe de travail du Cnis, j'ai craint que nous détruisions plus d'informations aujourd'hui que par le passé. L'archivage très poussé des données permet aux historiens de travailler sur cette période passée. Il conviendrait que les générations futures puissent faire de même. Il me paraît donc important que la Cnil porte ce message avec les Archives pour changer les pratiques dans ce domaine.
L'accès aux données privées soulève une question de protection du droit des affaires. La demande émerge parfois d'un renforcement de cette protection. Or l'exploitation de ces données peut présenter un intérêt général. Comment appréhendez-vous le possible accès aux données des sociétés d'assurance ?

José BARDAJI
Nous sommes favorables à l'utilisation des données à des fins de recherche, mais la profession a défendu l'introduction d'une réserve liée au secret des affaires dans l'utilisation des données personnelles. La frontière entre les données personnelles et le modèle de tarification, qui constitue le secret industriel de la société d'assurance, reste peu claire. Des travaux d'économie comportementale montrent que l'individu possède une capacité de concentration limitée. Dès lors, il faut trouver les bonnes questions à poser à l'individu pour pouvoir déterminer le risque. Les réponses à ces questions constitueront les réponses les plus discriminantes afin d'établir le risque et définir la tarification. De ce fait, donner la possibilité d'accéder aux données permet en quelque sorte de donner des informations sur le modèle lui-même. Pour une assurance automobile, par exemple, la couleur apparaît plus discriminante que la cylindrée de la voiture. Ainsi, les données personnelles renseignent sur le modèle. Communiquer largement sur ces données, c'est transmettre une information sur le modèle lui-même.

Jacques FOURNIER
Avec d'autres institutions publiques, notamment l'Insee et la direction générale des entreprises du ministère de l'Économie et des finances, nous réalisons la mesure de la compétitivité externe et la balance des paiements. Pour ce faire, nous interrogeons de nombreuses entreprises sur leur fonctionnement et leurs transactions internationales. Dans les grands groupes, nous entrons vite dans le secret des affaires. Une protection apparaît donc légitime. Nous devons assurer cette protection pour les entreprises, mais aussi pour la qualité de la statistique publique. Si nous n'offrons plus cette protection, en effet, les entreprises ne nous communiqueront plus ces informations et nous ne pourrons plus établir des comptes exacts.
Il existe plusieurs façons de contourner la difficulté. Outre l'anonymisation, nous pouvons constituer des comités d'accès aux données. Il en existe dans le domaine économique et financier notamment. Ce comité répond en général positivement à toutes les demandes de recherche. Il rejette en revanche les demandes

des entreprises qui, par ce biais, cherchent à obtenir des informations sur leurs concurrents. Ce système apparaît utile lorsque nous en avons besoin pour éviter des contournements. Une autre solution réside dans la régulation. Dans certains domaines, une utilisation des données sans régulation peut faire des exclus. Dans le secteur de l'assurance, la convention Aeras (S'assurer et emprunter avec un risque aggravé de santé) permet aux personnes gravement malades de bénéficier quand même d'une assurance, la réglementation fixant un plafond. Sans ce dispositif conventionnel, certaines personnes ne seraient plus assurées. Ce dispositif de solidarité publique conçu de façon partenariale avec les compagnies d'assurance se révèle très utile. On pourrait s'interroger sur la mise en place de tels mécanismes dans d'autres secteurs.

Francis JUDAS, Confédération générale du travail
Aujourd'hui, nous manquons de données sur les expositions professionnelles. Les fiches d'exposition dans les entreprises ont été supprimées pour la plupart des produits dangereux. La médecine du travail connaît par ailleurs de grandes difficultés et ses prérogatives diminuent. Il semblerait nécessaire d'utiliser le SNDS pour mieux repérer les maladies professionnelles, avec les limites et les difficultés que l'exercice pose en termes de secret et de protection des individus. En constatant une concentration de certaines pathologies dans une zone déterminée, par exemple, nous pourrions détecter des problématiques particulières et leur apporter une réponse publique. Dans le même temps, ces informations pourraient être utilisées dans le système assurantiel ou bancaire. Ainsi, en cumulant des risques importants, ces personnes pourraient être victimes de mesures négatives si le secret que la Cnil doit faire respecter était remis en cause.

Javier NICOLAU
Aujourd'hui, le SNDS ne comporte aucune donnée environnementale. Nous ne connaissons que la commune de résidence et la commune de soin. Néanmoins, Santé publique France suit depuis cinq ans une cohorte de travailleurs pour lesquels elle dispose de l'historique professionnel, la cohorte COSET. Ce programme permettra d'identifier les expositions en fonction de la profession et de les mettre en évidence, comme l'avait fait l'Inserm (Institut national de la santé et de la recherche médicale) pour l'amiante à partir d'enquêtes. L'approche retenue ici consiste à utiliser une cohorte et suivre les personnes de façon passive. Il en est de même avec la cohorte Constances. Les personnes qui intègrent ces cohortes sont suivies par le biais de questionnaires annuels et d'examens biologiques. Les données du SNDS permettent, en parallèle, de corriger les éventuels oublis. Les informations que nous recueillons apparaissent correctes pour ce type d'analyses.

Chantal CASES
Ces dispositifs permettent de réaliser des études de portée générale sur les conséquences sur la santé d'un certain nombre de parcours professionnels. Lorsqu'il s'agit de rechercher un lieu donné ou des personnes données pour apprécier les suites de leur exposition professionnelle, nous ne pouvons plus utiliser des fichiers généraux, nous devons constituer une cohorte et l'apparier avec les données de santé. Or ces dernières ne sont, en principe, conservées que durant vingt ans, ce qui pose des difficultés pour identifier les soins prodigués après l'exposition.

Javier NICOLAU
La loi prévoit que les données sont vivantes durant vingt ans, puis doivent être archivées durant dix ans.

De la salle
Je m'interroge sur la régulation entre intérêt général et intérêt marchand. Dans le champ de la santé, qu'est-ce qui garantit que ces bases de données ne permettent pas de réaliser un ciblage des bons et mauvais risques ? Comment les citoyens sont-ils protégés des effets néfastes de la constitution de méga-bases de données qui permettent de les cibler lorsqu'ils consomment en ligne ? Ce dispositif est-il régulé ou en passe de l'être ?

Javier NICOLAU
L'intérêt général ne s'oppose pas à l'intérêt marchand. Toute demande d'accès à des données de santé doit être motivée par l'intérêt général. Aujourd'hui, un acteur ne peut pas récupérer les données du SNDS pour constituer son propre fichier. L'accès aux données est réglementé. L'intérêt public est validé en amont par l'Institut national des données de santé et le Cerees évalue la méthodologie, les objectifs et l'adéquation des informations demandées par rapport à ces objectifs. Enfin, la Cnil autorise le projet. Les données restent confinées ; les personnes travaillent directement sur un portail à la Cnam et n'ont accès qu'aux données que le Cerees a jugé nécessaires pour répondre aux objectifs. Un système d'audit et de contrôle a été mis en place par la Cnil et un comité d'audit accompagne la mise en place du SNDS. Ce sujet ne soulève donc pas de problématique.
Lors des débats publics sur l'accès à ces données, deux points été mis en avant. Certains se demandaient pour quelle raison les industriels pourraient réaliser des bénéfices avec nos données et considéraient que le public serait pénalisé si les assureurs accédaient à ces données et établissaient des profils. Face à ces

craintes, le législateur a défini deux finalités interdites. D'une part, les industriels peuvent utiliser ces données, mais ils ne peuvent pas les utiliser pour vendre des médicaments ou des dispositifs médicaux auprès des professionnels ou des établissements de santé. D'autre part, les assureurs ne peuvent pas utiliser ces données pour faire du profilage. Ces deux catégories d'acteurs ne peuvent accéder aux données s'ils n'ont pas montré en amont à la Cnil qu'ils ne peuvent pas mettre en œuvre ces finalités interdites. À défaut, ils n'accéderont à ces données que par le biais d'un bureau de recherches ou bureau d'études.

Chantal CASES
Cette notion d'intérêt public ne va pas de soi. Dans la convention constitutive du système national des données de santé, un comité d'experts indépendants a été mis en place pour juger de l'intérêt public d'un certain nombre d'études afin d'élaborer une jurisprudence. Déterminer cet intérêt ne se révèle pas aussi simple que cela. Le travail de ce comité nous fournira donc un retour intéressant sur la réflexion autour de la protection de ces données.

José BARDAJI
Il faut distinguer l'assurance dans des conditions standards, l'assurance dans des conditions dégradées et la non-assurance. Les assureurs souhaitent mieux appréhender le risque pour assurer leurs clients dans des conditions standards. S'il ne parvient pas à mesurer le risque convenablement, l'assureur peut appliquer une surprime ou refuser d'assurer une personne. L'Aeras (S'assurer et emprunter avec un risque aggravé de santé) joue dans le cadre de l'assurance emprunteur. Lorsque vous demandez un crédit immobilier, par exemple, vous devez souscrire une assurance emprunteur et remplir un questionnaire. Selon votre état de santé, vous entrez dans l'un de ces trois cas de figure. La convention Aeras est allée plus loin et a utilisé des recherches pour cibler certaines pathologies et instituer une sorte de droit à l'oubli. Le futur emprunteur n'est pas tenu d'indiquer à son assureur qu'il a souffert d'un cancer dix ans plus tôt (ou cinq ans pour les personnes atteintes d'un cancer avant 18 ans). Il ne sera pas reproché à l'assuré de ne pas informer son assureur dans ce cas. La profession d'assurance pousse l'utilisation des données pour mieux appréhender le risque et réduire autant que possible les cas de non-assurance ou de surprime.
Nous avons vu apparaître certains fantasmes sur les données génétiques et leur utilisation potentielle par les sociétés d'assurance. Or je tiens à rappeler que cette utilisation est interdite par la loi au travers de cinq codes (Code de la santé publique, Code pénal, Code des assurances, Code de la sécurité sociale et Code de la mutualité).

Roxane SILBERMAN
Un nouveau paysage se dessine, avec un accès aux données plus ouvert pour la recherche et la possibilité d'appariements. Néanmoins, les procédures d'accès peuvent être différentes selon les acteurs, avec des critères qui peuvent être différents et des modes d'accès également différents et qui pourraient se multiplier. On peut du reste penser que des acteurs privés pourraient en proposer également. Cela pose évidemment des problèmes lorsqu'il faut travailler sur des données disponibles dans des environnements différents. Tout regrouper à terme n'est sans doute pas réalisable ni souhaitable en termes de protection des données si l'on pense à l'immense masse de données confidentielles. Comment voyez-vous l'organisation de ces accès ?

Chantal CASES
Il me semble qu'il faut organiser des passerelles entre ces silos, comme nous essayons de le faire autour de l'EDP-Santé (Échantillon démographique permanent- santé), surtout en termes de procédures. Nous pourrions considérer que lorsqu'un comité a statué sur le sujet, l'autre ne devrait intervenir qu'en complément, sur des points particuliers. Quant aux lieux d'accès, la question des référentiels de sécurité n'est pas totalement résolue. Nous nous heurtons aussi à la spécialisation des chercheurs sur certains types de données. Utiliser à la fois des données fiscales et des données de l'assurance maladie exige deux vies de travail. Dans notre projet, nous envisageons de créer deux versions de l'appariement en simplifiant pour partie l'un ou l'autre des jeux de données pour faciliter l'interdisciplinarité.

Jacques FOURNIER
Dans le domaine des données financières ou de banque centrale, nous travaillons au niveau national avec l'Insee et le CASD afin de surmonter sur le plan opérationnel les différences juridiques et rendre les appariements plus faciles pour les chercheurs. À l'international, nous avons créé Inexda, un réseau de banques centrales qui relie les ouvertures de données de différentes banques centrales. La Banque de France en assure le secrétariat. Dans une première étape, nous recenserons tout ce qui peut exister dans le monde, puis nous verrons si nous pouvons aller plus loin et donner accès en même temps à plusieurs bases géographiquement séparées. Il ne faut cependant pas sous-estimer les difficultés opérationnelles dont la résolution prendra du temps.

Antoine BOZIO
Le groupe de travail du Cnis avait insisté sur la nécessité de coordination. Ces procédures se révèlent coûteuses. La coordination permettrait d'instaurer un dispositif moins coûteux pour la société tout en garantissant la sécurité des données et en assurant un meilleur accès aux données.

Roxane SILBERMAN
J'ai été rassurée de constater que la mise en œuvre du RGPD n'a pas remis en question ce qui avait fait l'objet d'une clarification dans le groupe de travail sur les rôles respectifs de la Cnil et des Archives quant à la question de l'archivage des données confidentielles. En discutant avec des collègues étrangers, je me suis demandé s'il en était de même dans les autres pays.

Bruno RICARD
Afin de fluidifier le système, les Archives nationales ont signé une convention avec le CASD pour la communication des données archivées. Nous avions recensé de nombreuses demandes d'accès à distance auxquelles nous ne pouvions pas répondre techniquement. Cette convention a donc permis de créer une passerelle entre le CASD et le secteur des archives.
S'agissant du RGPD, nous avons essayé de garantir, durant la négociation de 2012 à 2016, puis dans la préparation de son implémentation en droit national l'an dernier, que les dérogations instaurées par le texte européen soient bien inscrites dans le droit français. Les dérogations à certains droits des personnes concernées par les données étaient effectivement proposés à titre optionnel (art. 89 du RGPD) et tous les États ne les ont pas confirmés dans leur droit positif. La France fait partie des pays qui ont inscrit le plus grand nombre de dérogations visant à accorder des droits aux chercheurs. Néanmoins, les grands États européens, comme l'Allemagne, le Royaume-Uni, l'Espagne et l'Italie ont fait à peu près de même. Nous devrions donc observer une très forte convergence.

Antoine BOZIO
Je tiens à remercier tous les intervenants de cette table ronde pour leur contribution, ainsi que les organisateurs de cette journée.

.III CLÔTURE

Jean-Luc TAVERNIER
En votre nom à tous, je remercie les organisateurs, les participants et les animateurs des tables rondes. J'évoquerai principalement la façon dont le service statistique public, l'Insee et les services statistiques ministériels s'adaptent à ces nouvelles données. J'aborderai également la production de données statistiques par des entreprises privées et les usagers.
Dès son intervention liminaire, Mireille Elbaum nous a rappelé qu'il faut poser la question avant de trouver la réponse. Or les données massives présentent le risque d'examiner la donnée, puis de se demander quelle était la question. Comme Pierre-Philippe Combes l'a souligné, nous avons besoin d'un sous-jacent théorique, d'un cadre cohérent et des données qui documentent un phénomène social ou économique que nous avons identifié ou donc nous avons perçu l'intérêt *a priori.* Je crois que ce réflexe est encore plus vrai dans la sphère statistique publique qu'ailleurs, du fait de notre double casquette initiale d'économiste et de statisticien. Nous voulons raconter l'histoire, donner l'explication, trouver la causalité, le modèle.
Nous tenons aussi beaucoup à maîtriser la donnée de base. Tel était le cas pour les deux premiers piliers que constituent les enquêtes et les données administratives. Dans ces deux situations, nous connaissons la donnée de base. Dans le cas des enquêtes, nous l'avons construite nous-mêmes. Le traitement de l'information relève ensuite du système statistique public. Nous maîtrisons donc toute la chaîne. C'est aussi, je crois, ce que nous avons cherché à faire avec les données d'entreprises dans le cadre juridique de la loi pour la République numérique. Nous demandons aux entreprises, dans ce troisième pilier de la statistique publique formé par les données structurées des bases de données d'entreprises, de la donnée élémentaire que nous traiterons nous-mêmes. De là, j'en déduis une certaine défiance vis-à-vis des données traitées y compris par les Gafa, qui dépendraient d'un algorithme un peu « boîte noire », de l'agrégation de données venant de sources d'origines différentes et des conditions dans lesquelles le webscraping peut donner des résultats intéressants, qui se révèlent assez exigeantes.
Il existe une certaine homogénéité dans le traitement des données administratives et des enquêtes conduites auprès des ménages. Sylvie Lagarde a fortement insisté sur le fait que le big data comporte des données de sources très diverses. La pertinence pour traiter les sujets varie grandement. De la même manière, la manière dont la qualité est garantie dans la sphère statistique sera fonction de la question posée et du traitement retenu. À chaque fois, il nous faut donc trouver un nouveau process, un nouveau cadre qualité. C'est aux fins d'expérimenter tout cela que nous avons créé, au sein de l'Insee, le SSP Lab qui, pour le compte de l'ensemble du service statistique public, veille à toutes ces possibilités d'utiliser des données massives.

En qualité de gestionnaire, j'ai noté une bonne nouvelle. Certes, ces données s'avèrent très exigeantes en termes de cadre juridique, de capacité informatique, de capital humain et de compétences, mais il me semble que ces éléments figuraient déjà dans notre ADN et s'inscrivent dans la continuité de ce que nous faisons pour les données administratives qui se révèlent parfois un peu massives. Toutefois, comme je l'avais souligné lors de notre précédente rencontre sur l'économie numérique, ces nouvelles données représentent en général du travail supplémentaire. Ces données ne se substituent pas aux dispositifs existants ; elles les complètent. S'agissant des vacances d'emploi, je doute que nous puissions remplacer le dispositif actuel de veille sur la dynamique des offres d'emploi par le webscraping. En revanche, ce dernier permettra de recueillir plus d'informations sur la structure et la répartition de ces offres. Je ne vois pas d'exemple dans lequel nous pourrions désarmer un processus existant grâce à ces nouvelles données.
Lorsque nous avons tiré les conclusions du groupe de travail sur les objectifs de développement durable et leur suivi par le système statistique public français dans le cadre du Bureau du Cnis, nous avons relevé une demande toujours très importante d'enquêtes structurées auprès des ménages sur le mal-logement, la grande pauvreté, l'illettrisme, le handicap, les séparations familiales, les discriminations, etc. Ces sujets nécessitent un protocole d'enquête relativement lourd et ne peuvent être traités par l'exploitation d'informations parcellaires, même massives. Je ne crois pas non plus que nous trouvions des solutions dans le big data pour pallier la suppression annoncée de la taxe d'habitation. Les données massives ne constituent pas si naturellement un facteur d'efficacité. Elles peuvent en créer un peu, mais elles ne sont pas faites pour cela.
S'agissant des statistiques privées, nous avons vu l'exemple de Trendeo avec David Cousquer ce matin. Il a cité lui-même quelques acteurs. Nous aurions également pu évoquer une entreprise spécialisée dans la production de fiches de paie qui, à ce titre, publie des indicateurs très avancés d'emploi aux États-Unis et en France. Il arrive parfois que ces données fassent d'ailleurs la une des journaux. Je me suis intéressé à la question du modèle économique de ces acteurs. En général, la statistique ne peut constituer la finalité de ces sociétés ; elle représente plutôt un produit fatal. Il suffit parfois d'une petite valeur ajoutée supplémentaire pour pouvoir proposer un produit d'appel dans la presse. Dans ces conditions, il me semble que seule une petite partie de la production de la statistique publique se trouve concurrencée par de la production privée qui trouverait son modèle économique, mais cette position apparaît peut-être un peu optimiste.
Ce matin, la question a été posée de savoir si le système statistique public peut jouer un rôle de régulation face à ces productions statistiques. Les procédures d'étalonnage et de labellisation en vigueur au sein du Cnis ont été rappelées. Elles ont été principalement utilisées par des organismes publics ou parapublics jusqu'à présent et très peu par des sociétés privées, avec plus ou moins de fortune. Nous pouvons faire mieux connaître cette offre de service. Je demande régulièrement dans différentes instances si nous devrions aller plus loin et proposer de manière plus proactive d'étalonner ces statistiques, en faisant éventuellement savoir de manière publique les réussites et les échecs. J'ignore si nous devons nous engager dans cette voie. La démarche se révélerait lourde en moyens et pousserait très loin le rôle de l'administration. En tout été de cause, ces statistiques n'entrent dans le paysage public que lorsqu'elles sont diffusées par les relais d'opinion et il me semble qu'il revient à ceux-ci de vérifier que ces données reposent sur une méthodologie et une démarche scientifique avant de leur donner un écho. Nous devons donc réfléchir au rôle de la statistique publique, mais aussi à celui des relais d'opinion en la matière.
Quant aux usagers, j'ai été sensible à l'intervention de la représentante de l'Uniopss ce matin. Nous publions des statistiques d'emploi sous différentes formes dans la statistique publique et nous éprouvons déjà des difficultés à nous y retrouver. Nous devons progresser, comme nous nous y sommes engagés auprès de l'Autorité de la statistique publique, pour aider les chercheurs à identifier les statistiques d'emploi les plus utiles en fonction de l'usage qu'ils entendent en faire. Si nous ajoutons les données d'emploi produites par les entreprises privées, nous brouillons considérablement le paysage et il devient extrêmement difficile pour l'usager de s'y retrouver. L'étalonnage et la labellisation de ces statistiques privées constitueraient une réponse à cette problématique. Une autre réponse consisterait à faire prendre conscience aux usagers des arbitrages nécessaires entre qualité, rapidité de publication et granularité. Aujourd'hui, il nous est demandé à la fois des données de qualité, publiées rapidement et à un niveau géographique fin. Or pour que le modèle de la statistique publique tienne, nous ne pouvons pas nous contenter d'une donnée « *quick & dirty* ». L'arbitrage existe et l'arrivée des données massives ne le modifie pas. Données massives ou pas, pour obtenir une évaluation de qualité des inégalités une année donnée, il faut attendre l'exploitation des données fiscales. Il en est de même pour l'évaluation de la situation financière des entreprises ou du chômage au niveau local, sauf à doubler, tripler ou quadrupler le coût de l'enquête Emploi pour former un échantillon représentatif au niveau du bassin d'emploi. Il n'existe aucune solution alternative à cela et les données massives ne peuvent apporter de réponse à cet arbitrage. Nous pouvons être émerveillés par la capacité à obtenir des données rapides et très désagrégées grâce aux données massives, mais nous rappelons souvent au niveau européen que la qualité n'est pas forcément au rendez-vous.
Enfin, pour faire écho au débat sur le droit, je crois que la statistique publique doit conjuguer différentes exigences : la confiance des enquêtés envers la confidentialité des données qu'ils nous confient, la protection des données individuelles pour les particuliers et du secret des affaires pour les entreprises,

l'ouverture aux chercheurs et l'open data. Or je ne peux que me féliciter de la façon dont nous parlons avec le monde qui écrit la loi, non seulement le législateur, mais aussi les services juridiques des différents ministères. À travers la loi Santé, la loi pour la République numérique ou la réécriture de la loi Informatique et libertés, nous avons fait au mieux pour essayer de conjuguer ces injonctions *a priori* paradoxales. La loi pour la République numérique nous offre d'ailleurs un cadre juridique qui nous permet d'alimenter la statistique publique avec des données d'entreprises. Néanmoins, ce dispositif ne fonctionnera que si les entreprises nous accordent leur confiance, car les sanctions qu'elles encourent restent insuffisantes pour les dissuader de ne pas nous répondre. Je n'entrerai pas dans le débat entre Cnil et archivage. Ce sujet apparaît très compliqué et mérite sans doute beaucoup de communication.
Je vous remercie.

Patrice DURAN
Merci à tous pour cette très belle journée.

L'introduction de Madame Elbaum a fait l'objet de la « Chroniques » N° 16 – Les enjeux des nouvelles sources de données.

Toutes les présentations ainsi que ce compte rendu sont disponibles et téléchargeables sur le site du Cnis à la page du colloque.

Liste des participants

ADAM	Lorraine	Centre national de la recherche scientifique (CNRS)
AFSA	Cédric	Conseil national de l'information statistique (Cnis)
ANXIONNAZ	Isabelle	Conseil national de l'information statistique (Cnis)
AOUIZERATE	Thierry	Insee Info Service
AUBRY	Étienne	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques
BAÏZ	Adam	Ministère de la Transition écologique et solidaire - Service de la donnée et des études statistiques (SDES)
BARDAJI	José	Fédération française de l'assurance
BARON	Jean-François	Institut national de la statistique et des études économiques (Insee) – Direction de la diffusion et l'action régionale (DDAR)
BARRAT	Daniela	Union des Industries des Cartons, Papiers et Cellulose
BAUER	Denise	Direction régionale des entreprises concurrence, consommation, travail et emploi (Direccte)
BAYET	Alain	Institut national de la statistique et des études économiques (Insee) – Secrétariat général
BELLER	Catherine	Conseil national de l'information statistique (Cnis)
BELLOC	Brigitte	Société française de statistiques
BERTHOLON	Raphaëlle	Confédération française de l'encadrement - Confédération générale des cadres (CFE-CGC)
BISCHOFF	Pierre	Commission européenne
BONNANS	Dominique	Institut national de la statistique et des études économiques (Insee) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
BONNET-GRAVOIS	Nicolas	Fédération des promoteurs immobiliers de France - FPI
BOTON	Blaise	Université Paris Dauphine
BOUCHÉ	Geneviève	Netwatz
BOULTE	Patrick	Solidarités nouvelles face au chômage (SNC)
BOURQUIN	Jean-Claude	Union fédérale des consommateurs - Que choisir ?

BOUVIER	Gérard	Institut national de la statistique et des études économiques (Insee)
BOYADJIAN	Michaël	Institut national de la statistique et des études économiques (Insee) - Direction des statistiques d'entreprises (DSE)
BOZIO	Antoine	Institut des politiques publiques
BRIÈRE	Luc	Institut National de la statistique et des études économiques (INSEE) - Direction de la diffusion et de l'action régionale (DDAR)
BROYART	Dominique	Ministère de la Transition écologique et solidaire - Direction générale de l'aménagement, du logement et de la nature (DGALN)
BRUNET	Francois	Banque de France (BdF)
BRUNSTEIN	Daniel	Université de Corse - CNRS
CACCINELLI	Chiara	Université Paris Dauphine
CAPELLE-BLANCARD	Gunther	Université Paris 1 Panthéon-Sorbonne
CASES	Chantal	Institut national de la statistique et des études économiques (Insee)
CAZAUBIEL	Arthur	École nationale de la statistique et de l'administration économique (Ensae)
CECI-RENAUD	Nila	Ministère du Travail - Direction de l'animation de la recherche, des études et des statistiques (Dares)
CÉSARI	Vartouhie	
CHALEIX	Mylène	Institut national de la statistique et des études économiques (Insee) – Secrétariat général informatique (SGI)
CHAUVET-PEYRARD	Axelle	Ministère de la Transition écologique et solidaire - Service de la donnée et des études statistiques (SDES)
CHEVALIER	Paul-Antoine	Etalab
CHOGNOT	Christine	Union nationale interfédérale des œuvres et des organismes privés sanitaires et sociaux (Uniopss)
CLING	Jean-Pierre	Institut national de la statistique et des études économiques (Insee) - Direction générale
COMBES	Pierre-Philippe	Université de Lyon
COTE-COLISSON	Daniel	dakota
COUDIN	Élise	Institut national de la statistique et des études économiques (Insee) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
COUSQUER	David	Trendeo

COUSTEAUX	Anne-Sophie	Institut national de la statistique et des études économiques (Insee)
CREUSAT	Joël	Institut national de la statistique et des études économiques (Insee)
CROGUENNEC	Yannick	Ministère des Solidarités et de la santé – Direction de la recherche, des études, de l'évaluation et des statistiques (Drees)
CUVIER	Christian	Comité du label de la statistique publique
DARMAILLACQ	Corinne	Institut national de la statistique et des études économiques (Insee)
DARRIAU	Valérie	Institut National de la statistique et des études économiques (Insee) – Direction de la diffusion et l'action régionale (DDAR)
DAUPHIN	Laurence	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (Drees)
DE BETTIGNIES	Martin	Union sociale pour l'habitat
DE VELLIS	Caroline	Agence d'urbanisme de Bordeaux
DELAUNAY	Isabelle	Conseil départemental du Vaucluse
DEROIN	Christine	Institut national de la statistique et des études économiques (Insee)
DIALLO	Thierno	Agence nationale pour l'information sur le logement (Anil)
DRUELLE	Sylvie	Insee Île-de-France
DUBOIS	Marie-Michèle	Conseil national de l'information statistique (Cnis)
DUÉE	Michel	Ministère de l'Intérieur - Direction générale des collectivités locales - Département des études et statistiques locales
DUQUESNOY	France	CARIF OREF Pays de la Loire
DURAN	Patrice	École normale supérieure
EIDELMAN	Alexis	Ministère du Travail - Direction de l'animation de la recherche, des études et des statistiques (Dares)
ELBAUM	Mireille	Haut conseil du financement de la protection sociale
EURIAT	Michel	Société française de statistiques
FAURET	Camille	Insee Île-de-France
FAYOLLE	Jacky	

FERMANIAN	Christophe	École des hautes études en santé publique (EHESP)
FLUXA	Christine	Institut national de la statistique et des études économiques (Insee) - Unité Qualité
FOLLENFANT	Philippe	Ministère de la Transition écologique et solidaire - Conseil général environnement et développement durable
FOURNIER	Jacques	Banque de France (BdF)
FRESSON-MARTINEZ	Catherine	Ministère de l'Agriculture et de l''limentation - Service de la statistique et de la prospective (SSP)
GALIANA	Lino	Insee Île-de-France
GAO	Fei	École des hautes études en santé publique (EHESP)
GASNIER	Claudine	Autorité de la statistique publique (ASP)
GÉLY	Alain	Confédération générale du travail (CGT)
GOMOT	Éléonore	Ministère de l'Économie et des finances - Direction générale des finances publiques (DGFiP)
GROLLEAU	Christine	Direction régionale et interdépartementale de l'environnement et de l'énergie - Île-de-France
GUIMARD	Philippe	Confédération générale du travail - Force ouvrière (CGT-FO)
GUYMARC	Gaël	Institut national de la statistique et des études économiques (Insee) - Direction des statistiques démographiques et sociales (DSDS)
HANEEF	Romana	Santé Publique France
HURPEAU	Benoît	Institut national de la statistique et des études économiques (Insee) – Direction de la diffusion et l'action régionale (DDAR)
JANIN	Jean-Louis	Académie de l'Eau
JEANTHEAU	Jean-Pierre	Agence nationale de lutte contre l'illettrisme (ANLCI)
JOSSE	Élodie	Conseil Régional d'Île-de-France
JOUTARD	Claire	Insee Provence-Alpes-Côte d'azur
JUDAS	Francis	Confédération générale du travail (CGT)
LAGANDRÉ	Véronique	Fédération des particuliers employeurs
LAGARDE	Sylvie	Institut national de la statistique et des études économiques (Insee) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)

LANG	Gérard	Société française de statistiques
LAURIN	Patrice	Oppchain
LAVARDE	Françoise	Ministère de l'Agriculture et de l'alimentation - Conseil général de l'alimentation, de l'agriculture et des espaces ruraux
LAVERGNE	Pierre	Préfecture Grand Est
LE GALLO	Florian	Banque de France (BdF)
LE MINEZ	Sylvie	Haut conseil du financement de la protection sociale
LECLAIR	Marie	Institut national de la statistique et des études économiques (Insee) - Direction des statistiques démographiques et sociales (DSDS)
LECOCQ	Marie	FranceAgrimer
LECOUVEY	François	Centre d'études et de recherches économiques sur l'énergie (Ceren)
LEMERLE	Stéphanie	Ministère de l'Éducation nationale
LEMOINE	Philippe	Commission nationale de l'informatique et des libertés (Cnil)
LIE	Joa	
LIXI	Clotilde	Ministère de la Justice - Sous-direction de la statistique et des études
LOUPIAS	Claire	Ministère de l'Économie et des finances - Direction générale du trésor (DGT)
MAILLARD	Sophie	Institut national de la statistique et des études économiques (Insee) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
MAKDESSI	Yara	Conseil national de l'information statistique (Cnis)
MARCHAND	Sylvie	Institut national de la statistique et des études économiques (Insee) – Direction de la diffusion et l'action régionale (DDAR)
MARECHAL	Christian	Télécom Paris Tech
MARTINEZ	Corinne	Institut national de la statistique et des études économiques (Insee)
MAUREL	Françoise	Institut national de la statistique et des études économiques (Insee) - Direction de la diffusion et l'action régionale (DDAR)
MONTÉRÉMAL	Marion	Institut national de la statistique et des études économiques (Insee) – Direction des statistiques d'entreprises (DSE)
MORAND	Elisabeth	Institut national des études démographiques (Ined)

MORDANT	Guillaume	Insee Info Service
MOTTER	Patricia	Union des caisses nationales de sécurité sociale
MOULIOM	Michel	Banque de France (BdF)
NACITAS	Catherine	Agence nationale pour la formation professionnelle des adultes (Afpa)
NARGEOT	Rodolphe	Conseil national de l'information statistique (Cnis)
NICOLAU	Javier	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (Drees)
PAQUEL	Norbert	Canope
PEROU	Olivier	Ocirp
PETIT	Jean-Jacques	Université de Reims
PICARD	Hugues	Institut national de la statistique et des études économiques (Insee)
PORTELA	Mickaël	Haut conseil à la famille
POUILLARD	Denys	Observatoire de la vie politique et parlementaire
POULHES	Mathilde	École nationale de la statistique et de l'administration économique (Ensae)
PROKOVAS	Nicolas	Pôle Emploi
PRUVOST	Christophe	Ministère de l'Économie et des finances - Direction générale des finances publiques (DGFiP)
REDOR	Patrick	Institut national de la statistique et des études économiques (Insee)
REY	Florence	Conseil Régional d'Île-de-France
REYNARD	Robert	Institut national de la statistique et des études économiques (Insee)
RICARD	Bruno	Archives de France
RICAU	Pascale	Ministère de la Transition écologique et solidaire - Service de la donnée et des études statistiques (SDES)
RIVIÈRE	Pascal	Institut national de la statistique et des études économiques (Insee) - Inspection générale
ROBIN	Marina	Institut national de la statistique et des études économiques (Insee) - Direction de la diffusion et l'action régionale (DDAR)

ROBIN	Yoan	Union nationale pour l'emploi dans l'industrie et le commerce (Unedic)
ROTH	Nicole	Institut national de la statistique et des études économiques (Insee) – Inspection générale
SAIDANI	Christine	Autorité de contrôle prudentiel et de résolution (ACPR)
SAKAROVITCH	Benjamin	Institut national de la statistique et des études économiques (Insee)
SAMSON	Gilles	Chambre de commerce et d'industrie (CCI)
SANCHEZ GONZALEZ	Joan	Institut national de la statistique et des études économiques (Insee) - Direction des statistiques d'entreprises (DSE)
SCANNAVINO	Philippe	Individuel
SCHERRER	Philippe	Institut national de la statistique et des études économiques (Insee) – Direction des statistiques d'entreprises (DSE)
SCHILTZ	Marie-Thérèse	Individuel
SEDILLOT	Béatrice	Ministère de l'Agriculture et de l'alimentation
SELZ	Marianne Marion	Centre national de la recherche scientifique (CNRS)
SEROUSSI	Géraldine	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques
SILBERMAN	Roxane	Groupe des écoles nationales d'économie et de statistique (Genes)
SKALIOTIS	Michail	Eurostat
SUESSER	Jan Robert	
SUJOBERT	Bernard	Confédération générale du travail (CGT)
TAGNANI	Stéphane	Conseil national de l'information statistique (Cnis)
TALL	Aguibou	Ministère du Travail - Direction de l'animation de la recherche, des études et des statistiques (DARES)
TAVERNIER	Jean-Luc	Institut national de la statistique et des études économiques (Insee) - Direction générale
TRAN LE TAM	Mélanie	Ministère des Outre-Mer - Direction générale des Outre-Mer (DGOM)
TRIPIER	Alain	Sereho
TROGNON	Alain	Groupe des écoles nationales d'économie et de statistique (Genes)

VALENTINO	Julien	Ministère de la Transition écologique et solidaire - Direction générale de l'aviation civile (DGAC)
VENEZIANO	Raphaël	Institut national de la statistique et des études économiques (Insee)
VUGDALIC	Suvani	Institut National de la statistique et des études économiques (Insee) - Direction de la diffusion et de l'action régionale (DDAR)
WIRTHMANN	Albrecht	Eurostat
ZOLOTOUKHINE	Erik	Centre national de la recherche scientifique (CNRS)