N° 1312

ASSEMBLÉE NATIONALE

CONSTITUTION DU 4 OCTOBRE 1958

SEIZIÈME LÉGISLATURE

Enregistré à la Présidence de l'Assemblée nationale le 1er juin 2023.

RAPPORT D'INFORMATION

DÉPOSÉ

en application de l'article 146 du Règlement

PAR LA COMMISSION DES FINANCES, DE L'ÉCONOMIE GÉNÉRALE ET DU CONTRÔLE BUDGÉTAIRE

*sur l'***accès** *aux* **données privées** *: une nouvelle ressource pour l'***Institut national de la statistique et des études économiques** *?*

ET PRÉSENTÉ PAR

M. MICHEL SALA,
rapporteur spécial

SOMMAIRE

—

Pages

SYNTHÈSE

L'Insee, acteur de référence en matière de production d'études et de statistiques publiques en France, voit ses prérogatives définies par un cadre juridique particulièrement exigeant visant à limiter les risques inhérents à la divulgation de données personnelles, parfois sensibles. Le traitement des données que réalise l'Insee s'est perfectionné au fil du temps, afin d'être en mesure d'appréhender à la fois le volume de données disponibles et ses différentes formes (échantillon collecté par enquête de terrain, données de nature administrative et données de nature privée).

La politique d'*open data*, impulsée tant au niveau européen que national, s'est traduite par l'ouverture des données de l'Insee. Cette ouverture a rendu nécessaire une évolution des pratiques de l'Institut afin de rendre les données publiques utilisables par les acteurs de la société civile mais également pour garantir un bon niveau de compréhension du public. Le corollaire d'une telle ouverture des données est un phénomène d'attrition des recettes de l'Insee qui, auparavant, revendait une partie de ses données à des tiers.

Aujourd'hui, l'activité de l'Insee intervient dans un contexte largement transformé par l'émergence du *Big Data*. L'encadrement normatif actuel n'apparaît plus pleinement pertinent pour garantir une utilisation optimale de la variété des données disponibles. Entre lourdeurs et protections des données de nature privée, l'Insee tâtonne pour faire évoluer ses pratiques, alors que le cadre juridique fait actuellement l'objet de perspectives de révision.

L'absence d'un cadre clair et lisible définissant l'utilisation par les services de statistique publique des données de nature privée engendre des complexités voire des crispations dans la mesure où la statistique se retrouve bridée dans ses potentialités. Pour surmonter ces limites, depuis la crise sanitaire, l'Insee a entrepris des démarches partenariales pour avoir accès à de nouvelles données. Des partenariats avec des établissements bancaires ou encore avec des opérateurs de téléphonie mobile ont démontré l'intérêt et l'utilité d'avoir accès à de nouvelles données.

Pour faciliter l'accès aux données détenues par les acteurs privés, il apparaît nécessaire de faire évoluer le cadre législatif applicable en matière de production statistique, en vue de fluidifier et d'intensifier les flux de données des acteurs privés à destination des acteurs publics, sans dégrader la qualité de la protection entourant les données. Le constat est clair, l'avenir de la statistique publique se situe, aujourd'hui notamment dans l'exploitation des données privées.

LISTE DES RECOMMANDATIONS DU RAPPORTEUR SPÉCIAL

Recommandation n° 1. Établir une définition législative large de la notion de donnée d'intérêt général et de ses modalités d'utilisation à des fins de production statistique, afin de permettre un accès facilité à ces données pour l'Insee et les services statistiques ministériels.

Recommandation n° 2. Permettre, pour les données en lien avec les politiques publiques environnementales, une expérimentation d'une définition législative large de la notion de donnée d'intérêt général.

Recommandation n° 3. Modifier la rédaction actuelle de l'article 3 *bis* de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.

Recommandation n° 4. Recenser, au cas par cas, avec les acteurs concernés, dans le cadre de partenariats de gré à gré, les données que les acteurs privés pourraient communiquer à l'Insee et plus largement aux acteurs de la statistique publique.

Recommandation n° 5. Proposer des garanties renforcées aux acteurs privés concernés par les échanges de données.

Recommandation n° 6. Prévoir un dispositif de réquisition administrative des données privées sous certaines conditions.

Recommandation n° 7. Sanctionner les refus d'accès systématiques et injustifiés d'accès aux données privées.

Recommandation n° 8. Uniformiser la notion de secret statistique afin de renforcer la sécurité juridique attachée à cette notion.

Recommandation n° 9. Développer la démarche de labellisation de séries statistiques.

Recommandation n° 10. Favoriser des concertations afin d'anticiper les besoins en matière de protection des données personnelles.

INTRODUCTION

Lointain héritier du Bureau de statistique institué en 1800 par Lucien Bonaparte et du service de la statistique générale de France [(1)] mis en place par Adolphe Thiers, l'Institut national de la statistique et des études économiques (Insee) assure, dans la continuité de sa fonction historique, un rôle d'éclaireur pour les décideurs et les citoyens. Étranger à la malversation dans l'utilisation des données dont peuvent être coutumiers les régimes autoritaires, comme l'évoque Ismaël Kadaré dans son *Palais des rêves* [(2)], l'Institut agit aujourd'hui comme un constructeur de confiance.

Dès 2003, et avec davantage d'acuité depuis 2016, l'Insee assure, dans le cadre de la politique d'*open data*, un rôle de diffuseur de données. Cette politique s'entend comme la mise à disposition de données à titre gratuit, de manière fiable et sécurisée au public. La France, largement engagée dans cette politique, a obtenu pour les années 2021 et 2022 la première place au classement *Open data maturity report* établi par la Commission européenne qui souligne l'implication des États membres dans la mise en œuvre de l'ouverture numérique.

L'Insee assure traditionnellement une triple fonction : fournir des données conjoncturelles directement disponibles et accessibles pour les différents acteurs de la société ; produire des études pour orienter les acteurs sur des politiques structurelles ; centraliser et traiter une masse importante de données disponibles dont la nature est hétérogène (données issues d'enquête de terrain, données administratives voire données issues de la sphère privée).

Ces fonctions traditionnelles connaissent toutefois une importante métamorphose sous l'effet du *Big Data*, caractérisé par un accroissement massif des données de nature privée disponibles. Aujourd'hui, les données auxquelles a accès l'Insee se diversifient et soulèvent de nouveaux enjeux. Des enjeux relatifs à l'accessibilité et à la bonne communication de l'information d'abord, dans la mesure où l'accroissement des ressources peut être un facteur complexifiant faisant perdre en lisibilité. Des enjeux d'adaptation du cadre juridique ensuite, puisque celui-ci, pour permettre une exploitation optimale de la nouvelle étendue des données, doit être à même de s'adapter. Des enjeux d'ouverture enfin, puisque l'Insee doit désormais sortir des sentiers battus pour s'orienter vers l'exploitation des données privées dont la richesse augmente à mesure que les sociétés se numérisent.

Les auditions conduites par le rapporteur spécial dans le cadre de ce rapport d'information ont été l'occasion d'explorer un domaine technique, parfois peu connu, souvent incompris, mais dont les enjeux apparaissent fondamentaux. Le

(1) Appellation qui durera près d'un siècle.

(2) Ismaël Kadaré, Le Palais des rêves, 1981.

rapporteur spécial a ainsi fait le choix de retenir un domaine d'évaluation étendu permettant d'appréhender cette politique publique dans sa globalité.

Le rapport d'information montre que si l'Insee s'impose comme un acteur de référence dans la production statistique (I) il n'en demeure pas moins corseté par un cadre normatif complexe et concurrencé par des acteurs privés disposant de toujours davantage de données (II). Un tel constat impose aujourd'hui de réfléchir aux moyens de fluidifier et d'intensifier les flux de données, au service de la statistique publique (III).

I. ACTEUR MAJEUR DE LA POLITIQUE D'*OPEN DATA*, L'INSEE ASSURE, PAR LA QUALITÉ DE SES TRAVAUX ET DE SES INDICATEURS, UNE ACTION PUBLIQUE ET DES CITOYENS ÉCLAIRÉS

L'Insee est un acteur de référence pour la connaissance de la situation économique et sociale du pays. La diversification de ses domaines d'étude au fil des années lui confère une crédibilité institutionnelle forte (A). La politique d'*open data* entraîne toutefois une transformation des pratiques de l'Insee, à la fois sous la forme d'obligations juridiques de mise à disposition des données et sous la forme d'information des publics (B).

A. L'INSEE APPARAÎT COMME L'ORGANE DE RÉFÉRENCE DE LA PRODUCTION STATISTIQUE EN FRANCE ET BÉNÉFICIE À CET ÉGARD D'UN ACCÈS ÉTENDU AUX DONNÉES.

1. L'activité statistique de l'Insee est délimitée par un cadre normatif ancien et exigeant

a. Un cadre juridique exigeant

L'activité contemporaine de l'Insee a été établie afin de consacrer, en France, une institution unique chargée de recueillir les données (1).

Le législateur a circonscrit l'activité de l'Institut aux domaines de la statistique, de la mécanographie, de la documentation et des études relatives aux problèmes économiques. L'Insee a également été consacré comme l'organe chargé de coordonner les méthodes, les moyens et les travaux statistiques des administrations publiques et d'organismes privés (2).

Cette activité, un temps limitée au niveau national, s'est progressivement élargie à l'échelon européen, afin d'accompagner le développement du système statistique européen et pour rendre possible les comparaisons économiques entre les États membres. Au début des années 2000, la collecte de données en vue de la participation française à l'établissement des statistiques européennes prend une part prépondérante dans le programme annuel national de la statistique publique. Un règlement du Parlement européen et du Conseil de 2009 (3) vient pérenniser cette collaboration et prévoit désormais des échanges nombreux et récurrents entre l'Insee et Eurostat. Ainsi, chaque année, l'Institut est amené à transmettre environ une soixantaine d'indices économiques conjoncturels à Eurostat.

L'Insee bénéficie par ailleurs d'une importante indépendance professionnelle dans la conception, la production et la diffusion de statistiques

(1) Loi n° 46-854 du 27 avril 1946.

(2) Décret n° 46-1432 du 14 juin 1946 relatif à l'Institut national de la statistique et des études économiques en métropole et la France d'outre-mer.

(3) Règlement n° 223/2009 du Parlement européen et du Conseil du 11 mars 2009.

publiques, ce qui est fondamental pour s'imposer comme un acteur crédible et favoriser la confiance du public. L'Autorité de la statistique publique (ASP) instituée par le législateur en 2008 (1) veille à cette indépendance en assurant un rôle de vigilance quant au respect des principes du code européen des bonnes pratiques par l'Insee.

Si les prérogatives de l'Insee en matière d'exploitation des données sont étendues, elles demeurent strictement encadrées lorsqu'il s'agit de réaliser des enquêtes statistiques. La loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques, qui régit la statistique publique, prévoyait, dans sa rédaction originelle, que l'Insee réalise ses productions à partir d'enquêtes et de fichiers administratifs uniquement. Aujourd'hui, pour tenir compte de la numérisation, l'Insee s'ouvre aux données privées.

Invariables toutefois, les productions de l'Insee doivent respecter un principe cardinal de son activité : **le secret statistique**. Une telle notion se décline en une triple contrainte couvrant l'entièreté des données susceptibles d'être connues par l'Institut.

L'article 6 de la loi de 1951 prévoit l'encadrement des données récupérées par l'intermédiaire d'enquête statistiques. À cet égard, les informations ayant trait à la vie personnelle, familiale et d'une manière générale, aux comportements privés sont revêtues d'une interdiction de communication dont le Comité du secret statistique (2) assure le respect.

L'article 7 *bis* de cette même loi, quant à lui, encadre l'utilisation à des fins de statistiques publiques des informations recueillies par l'administration dans le cadre de ses missions, après avis du Conseil national de l'information statistique (Cnis), à l'exclusion des données relatives à la vie sexuelle.

Enfin, l'article 3 *bis* de cette même loi délimite le recours aux données collectées auprès des organismes de droit privé. La possibilité de recourir à de telles données est récente puisqu'elle a été introduite par la loi pour une République numérique du 7 octobre 2016 (3).

(1) La loi n° 2008-776 du 4 août 2008 de modernisation de l'économie institue, par son article 144, l'Autorité de la statistique publique.

(2) Décret n° 2009-318 du 20 mars 2009 relatif au Conseil national de l'information statistique, au comité du secret statistique et au comité du label de la statistique publique.

(3) Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.

La statistique publique pendant la seconde guerre mondiale

En 1998, à la demande de Jean Claude Milleron (ancien directeur général de l'Institut) formulée en 1992, les historiens Jean-Pierre Azéma et Béatrice Touchelay ainsi que Raymond Lévy-Bruhl (Inspecteur général honoraire de l'Insee) ont remis un rapport portant « Mission d'analyse historique sur le système statistique français de 1940 à 1945 ».

Cette étude avait deux missions principales : i) critiquer et analyser les diverses sources relatives à l'histoire de la statistique pendant la seconde guerre mondiale ; ii) apporter une appréciation sur le rôle joué pendant le conflit par les organismes qui avaient précédé la naissance de l'Insee.

Une telle étude a révélé l'ambivalence du rôle de la statistique durant la période de la seconde guerre mondiale et de l'occupation.

Les responsables du Service national statistique (SNS) de l'époque se sont montrés vigilants à ce que l'information ne facilite pas les réquisitions allemandes notamment dans le recensement des jeunes personnes mobilisables au titre du STO (service du travail obligatoire). Toutefois, comme le relèvent les rédacteurs du rapport : « *le SNS a appliqué sans grands états d'âme la législation antisémite vichyssoise* ».

Le rapport concluait que les données disponibles de l'époque, notamment celles relatives à l'identité et au recensement de la population, avaient été mises à disposition des services de police pour permettre d'identifier une population spécifique.

À cette aune, l'enseignement de l'histoire rappelle que l'utilisation des données doit toujours être réalisée avec, comme ligne directrice, un respect profond des principes de la statistique publique, dont le secret est aujourd'hui la pierre angulaire.

Aujourd'hui, le cadre juridique entourant la protection des informations relatives à la vie privée à des fins de production statistique est particulièrement étendu.

b. Une production statistique riche et diversifiée

L'Insee assure une production d'études statistiques diversifiées. Ces enquêtes sont d'abord diversifiées dans leur nature et dans les domaines d'activités concernés. Certaines grandes enquêtes sont ainsi régulièrement utilisées par les acteurs publics et les acteurs de la communication afin d'appuyer leurs propos sur des données de référence [(1)]. À cet égard, quelques grandes études peuvent être mentionnées.

L'enquête relative au recensement de la population est assurée, depuis 1801, par un service de statistique en France. Dès le XIX^{e} siècle cette étude est employée afin de connaître le nombre d'administrés du royaume. Aujourd'hui, elle permet de déterminer les populations légales de la France et de ses circonscriptions administratives de manière annuelle. Le recensement est rendu possible par un partenariat entre les communes ou les établissements publics de coopération

(1) L'article L. 321-4 du code des relations entre le public et l'administration (CRPA) définit cette notion de donnée de référence et renvoie à un décret en Conseil d'État le soin de définir les critères permettant de qualifier comme telle une donnée.

intercommunale (EPCI) et l'Insee. La loi du 27 février 2002 relative à la démocratie de proximité a fait de l'Insee l'acteur chargé de la collecte des données [1]. Si cette enquête s'avère relativement discrète, elle n'en demeure pas moins incontournable : près de 350 articles de lois ou de codes se référant directement à la population légale des circonscriptions administratives.

Peut également être mentionnée l'Enquête sur l'emploi, le chômage et l'inactivité (EEC**)**. Cette enquête permet d'observer le marché du travail sur le plan conjoncturel et structurel. Elle est la seule étude répondant aux critères définis par le Bureau international du travail (BIT) en matière d'appréciation du chômage. Elle est donc indispensable aux comparaisons internationales. Produite de manière annuelle, elle est constamment enrichie par les services de l'Insee en vue d'accroître l'échantillon d'étude et d'obtenir ainsi des indicateurs plus représentatifs.

Enfin, l'Enquête logement (ENL), permet de décrire les conditions de logement des ménages et de comprendre les dépenses d'habitation. Elle permet ainsi d'apprécier les conditions de logement en France, de qualifier le mal-logement ou encore d'apprécier la diversité du patrimoine immobilier.

Pour être fiables et représentatives, les enquêtes réalisées par l'Institut doivent viser une volumétrie d'enquêtés suffisante. Ainsi, afin de recueillir des informations statistiques fiables sur la situation économique, l'Institut a enquêté, en 2021, auprès de 20 000 entreprises tous les mois. Les ménages sont également sollicités ; c'est ainsi près de 2 000 ménages qui ont été entendus chaque mois afin de recueillir leur ressenti sur la conjoncture économique.

Au-delà des enquêtes, l'Insee produit des indicateurs conjoncturels permettant d'obtenir des informations clefs sur la situation du pays.

Certains indicateurs sont utiles à l'action publique pour pouvoir agir sur une situation donnée. À cet égard, l'Indice des prix à la consommation (IPC) permet de rendre compte de l'augmentation généralisée des prix des biens et des services à caractéristiques constantes. Pour la construction d'un tel indicateur, l'Institut examine un grand nombre de données disponibles.

(1) Articles 156 à 158 de la loi n° 2002-276 du 27 février 2002 relative à la démocratie de proximité.

La fabrique de l'Indice des prix à la consommation (IPC)

D'abord, l'Insee définit un échantillon de produits sélectionnés pour suivre les différentes formes de vente sur l'ensemble du territoire. Ces biens ou services sélectionnés font l'objet d'un suivi selon un rythme *a minima* mensuel, afin de permettre l'actualisation des métadonnées et de délivrer, en bout de chaîne, un indicateur cohérent chaque mois.

Ensuite, l'Insee réalise une collecte de la donnée dont la volumétrie doit être suffisamment importante pour être représentative et fiable. Pour ce faire, l'Institut collecte les données suivantes :

– environ **150 000 relevés de prix, chaque mois**, dans plusieurs dizaines de milliers de points de vente ;

– environ **500 000 relevés de prix sur internet**. Certains sont réalisés manuellement par le personnel de l'Insee. La plupart sont réalisés par moissonnage automatisé (*webscraping*) sur internet ;

– les données de caisses des enseignes de la grande distribution, leur transmission ayant été rendue obligatoire depuis un arrêté du 13 avril 2017 (1) ;

– des **relevés de « tarifs »** disponibles sur internet ;

– des enquêtes statistiques par méthode de sondage auprès des ménages pour la mesure de l'évolution des loyers (enquête Loyers et Charges, enquête sur les Loyers auprès des Bailleurs sociaux).

Enfin, l'IPC est publié chaque mois au Journal Officiel, permettant d'indexer de nombreux contrats.

Cette fabrique de l'information par l'Insee est rendue possible par un accès large à une importante variété de données disponibles.

L'Insee n'est toutefois pas le seul service public de statistique à tirer profit d'un accès large aux données, notamment de nature privée.

À cet égard, les services de la direction de l'animation de la recherche, des études et des statistiques (DARES) qui produit des analyses et des études statistiques sur les thèmes du travail, de l'emploi, de la formation professionnelle et du dialogue social, pourraient être favorables à une telle extension.

De la même manière, la direction de la recherche, des études, de l'évaluation et des statistiques (DRESS), qui apporte son expertise dans le domaine de la santé, du grand âge ou encore du handicap, profiteraient utilement des ouvertures de données privées.

(1) Arrêté du 13 avril 2017 rendant obligatoire la transmission de données par voie électronique à des fins de statistique publique.

2. L'Insee reçoit des données variées et perfectionne ses méthodes de collecte

a. Le caractère diversifié des données collectées

La variété des données dont bénéficie l'Insee et l'ouverture récente aux données privées permise par la loi pour une République numérique (LPRN) offrent aux services de statistique des perspectives nouvelles. Le choix de la donnée idoine est tributaire de considérations budgétaires, du temps nécessaire pour la traiter et du niveau de détail souhaité pour les statistiques à publier.

D'abord, l'Insee s'est historiquement appuyé sur **les données en provenance des enquêtes par échantillonnage consistant en une collecte directe**.

Aujourd'hui, les enquêteurs de l'Insee sont essentiellement concentrés au sein des quinze directions régionales. Si la collecte de données à partir du terrain – via des questionnaires réalisés au domicile de l'interrogé ou auprès des entreprises par les enquêteurs de l'Insee – peut apparaître comme surannée, compte tenu du développement du *multimode* et des questionnaires par internet, elle n'en demeure pas moins fondamentale sur le plan méthodologique.

En premier lieu, l'enquête de terrain permet de pallier certaines limites de la méthode par échantillonnage dans la mesure où toute la population ne bénéficie pas d'un accès pérenne à internet, sans compter les personnes frappées par le phénomène d'illectronisme [(1)].

En second lieu, les enquêtes en physique sont moins sujettes aux biais cognitifs, ce qui permet de garantir la fiabilité de l'information recueillie.

Le rapporteur spécial tient à souligner, après avoir auditionné les syndicats Sud-Insee et CGT-Insee, que le phénomène de compression des effectifs des enquêteurs – pour l'essentiel composés d'enquêtrices contractuelles – devrait être stabilisé afin de ne pas précariser davantage leur situation et ne pas se priver des ressources collectées sur le terrain.

Ensuite, l'Insee a fait de **l'utilisation des données administratives** une ressource privilégiée pour ses travaux statistiques.

Si cette pratique n'est pas nouvelle, elle tend à s'accroître à mesure que la dématérialisation des procédures administratives progresse. À cette aune, l'Insee s'appuie par exemple sur les déclarations fiscales des ménages et des entreprises, sur la déclaration sociale nominative (DSN) remplie par les employeurs ou encore sur les données de prestations sociales et familiales.

---

(1) M. Raymond Vall, Rapport de la mission d'information relative à la lutte contre l'illectronisme et pour l'inclusion numérique », n° 711 (session 2019-2020), Sénat, 17 septembre 2020.

Toutefois, conformément aux principes de nécessité et de minimisation, seules les données utiles à l'élaboration de statistiques peuvent être transmises à l'Insee.

Ces données administratives nécessitent des exploitations spécifiques coûteuses pour l'institution afin de produire des résultats statistiques de qualité. En effet, la couverture des sources peut ne pas être parfaite, il peut exister des double-comptes et les variables renseignées peuvent correspondre à des données de gestion dont la nature diffère des concepts que le statisticien cherche à mesurer. Par exemple, dans les DEFM − demandeurs d'emploi en fin de mois −, ce n'est pas le concept de chômage au sens du BIT qui est mesuré.

Ces données administratives sont toutefois privilégiées par l'Insee dans la mesure où elles permettent de limiter les coûts de dépenses de personnel inhérents à une enquête de terrain. **Sur le plan budgétaire, la ressource administrative, lorsqu'elle est adaptée à la nature de l'étude, est donc privilégiée** (1).

Le rapporteur spécial tient à souligner, s'agissant de ces données administratives, une culture du secret persistante. Le Conseil d'État le relevait dans son étude sur l'intelligence artificielle : « *La politique de la donnée s'apparente encore à un exercice imposé […] les administrations sont encore trop nombreuses à omettre, par inertie ou à rechigner, par culture du secret, par précaution face à l'irréversibilité d'une communication excessive de données* » (2). Cette culture du secret nuit à la bonne exploitation des données publiques.

Enfin, l'Insee a la possibilité d'exploiter, depuis la loi du 7 octobre 2016 pour une République numérique, des **données de nature privée**.

Cette possibilité offerte, restée largement inexplorée entre 2016 et 2020, a été développée avec la crise du Covid-19, lorsqu'il a fallu pallier la suspension temporaire des enquêtes de terrain.

La mobilisation des données privées peut relever de deux fondements : d'une part, l'article 3 *bis* de la loi de 1951 précitée ; d'autre part, des partenariats *ad hoc,* de gré à gré, avec des groupes issus de la société civile.

La diversité des données pouvant être exploitées est infinie. Il peut s'agir ainsi de données de relevés de compteurs d'électricité ou de gaz, de données de transactions par carte bleue ou encore de données de téléphonie mobile.

Une telle opportunité est toutefois un défi pour les services de l'Insee qui doivent faire face à une masse croissante de données (*Big Data*). L'utilisation de ces dernières, au-delà d'un cadre légal permettant l'exploitation et l'établissement de partenariats avec des acteurs privés, connaît plusieurs limites.

---

(1) Blog Insee : « Quels types de sources l'Insee utilise-t-il pour construire ses statistiques », publié le 15 mai 2023.

(2) Rapport du Conseil d'État à la demande du Premier ministre, « Intelligence artificielle et action publique : construire la confiance, servir la performance » adopté en assemblée générale plénière le 31 mars 2022.

L'Insee ne maîtrise pas les données privées, en ce sens que les codes sources et métadonnées ne sont pas toujours connus par les services statistiques de l'Institut – ce qui ne permet pas de garantir la qualité et l'objectivité des données utilisées.

Au surplus, une autre limite tient à l'incomplétude des données : ces dernières ne fournissent parfois qu'une information partielle voire partiale d'une situation économique ou sociale.

Enfin, les services de l'Insee se refusent à rémunérer les détenteurs privés de données afin d'y avoir accès, ce qui constitue une limite significative à leur exploitation.

b. Le perfectionnement des méthodes de collecte des données

En plus de la diversité des données dont bénéficie aujourd'hui l'Insee, la méthodologie de traitement des *data* évolue et se perfectionne pour faire face à l'inflation numérique.

Premièrement, les services de l'Insee développent, conformément aux objectifs de dématérialisation des enquêtes établis dans le document de cadrage « Horizon 2025 » (1), la méthode de collecte en multimode. Cette dernière permet d'importants gains d'efficience puisqu'elle permet de réduire le temps consacré aux enquêtes en face-à-face au profit de réponses par internet.

Le rapporteur spécial tient à souligner, comme il l'avait déjà fait à l'automne dernier, que cette transformation des tâches risque d'accroître les contraintes pesant sur les enquêteurs avec des heures supplémentaires passées au téléphone.

Le rapporteur spécial déplore que cela s'accompagne de la fermeture d'antennes d'études ou de réduction d'effectifs de l'Insee dans les territoires (2).

Deuxièmement, l'Insee s'appuie davantage sur les méthodes de *data science* et de *cloud computing* (3) (ou infonuagiques) pour parvenir à traiter l'entrée croissante de nouvelles données. À cette fin, en 2018, deux unités – le lab de la statistique publique (SSP Lab) (4) et la division innovation et infrastructure technique – ont été créées afin d'impulser et d'animer l'innovation en matière de *data science* au sein de l'Insee.

---

(1) Orientations, objectifs et actions 2016-2025 : « Horizon 2025 : Une stratégie ambitieuse pour l'Insee ».

(2) Les fermetures de certaines antennes territoriales sont consécutives à la réforme de l'administration territoriale de l'État (RÉATE), *ainsi le centre de traitement de l'informatique de l'Insee à Rocquencourt et à Aix-en-Provence ont été fermés en 2010.*

(3) Le cloud computing, *ou l'informatique en nuage, est un modèle d'accès à des ressources informatiques à distance via Internet. Ainsi, au lieu d'héberger et d'exécuter des programmes ou des fichiers depuis son propre outil informatique, le* cloud computing *permet de les stocker et de les traiter sur des serveurs distants, souvent gérés par des fournisseurs de services cloud.*

(4) Centre de ressources et d'animation pour la recherche appliquée et le développement expérimental visant à promouvoir l'innovation et la nouveauté en matière de sources de données, de technologies et de méthodes de data science, relatives aux productions statistiques du système de statistique publique (SSP).

Les effectifs de ces deux unités demeurent circonscrits puisqu'ils s'élèvent à 13 ETP fin 2022. Le développement des méthodes issues de la *data science* mobilise les technologies de l'intelligence artificielle, principalement pour de l'analyse textuelle et du traitement d'images (documents scannés, images satellitaires).

De manière récurrente, l'Insee a recours au *web scraping* (moissonnage des données en ligne) afin d'enrichir ses bases de données de référence et procéder à une contemporanéisation de celles-ci.

*

Acteur majeur de la production statistique, l'Insee bénéficie aujourd'hui d'un panel de données étendu lui permettant de mener à bien ses missions. Si l'accroissement des données disponibles bénéficie à l'Insee, l'Institut est également contraint, par la politique d'*open data*, à l'ouverture de ses bases d'information. Une telle évolution suppose une démarche pédagogique de l'institution à destination des acteurs de la société civile.

B. L'*OPEN DATA* FAVORISE L'OUVERTURE DES BASES DE DONNÉES ET INDUIT UNE ACCESSIBILITÉ ACCRUE ET UNE PÉDAGOGIE RENFORCÉE POUR L'INSEE

1. L'*open data* induit pour l'Insee un renforcement de ses pratiques mais engendre une perte sèche de recettes

*a. Aller en avant de l'*open data *: rendre les données réutilisables*

La politique d'ouverture de la donnée menée par l'INSEE sous l'impulsion de son directeur général s'inscrit dans les deux premières orientations du plan « Horizon 2025 » précédemment mentionné. La politique d'*open data* de l'Insee est antérieure à la loi pour une République numérique. En effet, dès 2003, le comité de direction a décidé de diffuser gratuitement sur internet tous ses résultats statistiques. Aussi, la politique d'*open data* fait partie intégrante de l'ADN de l'Institut.

Désormais, toutes les informations et données de l'Insee peuvent être réutilisées librement et gratuitement, y compris à des fins commerciales.

La loi prévoit en effet une utilisation des données particulièrement étendue. L'article L. 321-1 du code des relations entre le public et l'administration (CRPA) dispose que : « *Les informations publiques figurant dans des documents communiqués ou publiés par les administrations […]* ***peuvent être utilisées par toute personne qui le souhaite à d'autres fins que celles de la mission de service public*** *pour les besoins de laquelle les documents ont été produits ou reçus* ».

Les données comportant des informations à caractère personnel ne sont réutilisables qu'après anonymisation (article R. 322-3 du CRPA).

Les limites à ces réutilisations se matérialisent par des obligations de « bonnes pratiques » pour les utilisateurs secondaires de données. L'article L. 322-1 du CRPA prévoit que *« la réutilisation des informations publiques est soumise à la condition que ces dernières ne* ***soient pas altérées, que leur sens ne soit pas dénaturé et que leurs sources et la date de leur dernière mise à jour soient mentionnées*** ».

Le rapporteur spécial comprend le caractère vertueux d'une telle réutilisation des données, mais déplore le fait que la réutilisation commerciale ne s'apparente parfois qu'à un faux « fait maison », consistant pour les utilisateurs diffuseurs de données à modifier celles-ci à la marge, entraînant un pur effet d'aubaine pour les sociétés privées.

L'enjeu pour l'Insee n'est donc pas tant celui de l'ouverture des données que celui de la qualité de leur mise à disposition. Les besoins et les goûts du public en matière d'information évoluent vers des accès multiples aux données (fichiers Excel, téléchargements de masse et, de plus en plus, par interface de programmation d'application (API) (1)).

Pour répondre à ces besoins épars, la loi pour une République numérique a créé le Service public de la donnée (SPD), établi de concert entre l'Insee et la Direction interministérielle du numérique (Dinum), concentrant l'offre en matière de données. Le plus haut niveau de qualité dans la mise à disposition des données est toujours recherché. Par exemple l'ouverture de la base Sirene, historiquement réalisée par l'Insee, via l'interface de programmation d'application API Sirene, permet aux systèmes d'information des usagers d'intégrer les données Sirene sans intervention humaine extérieure, et favorise ainsi des usages professionnels beaucoup plus performants. L'Institut n'oppose aucun frein au développement de l'*open data* ; au contraire, cette diffusion étendue des données permet de s'imposer comme « standard de fait » (2) dans certains domaines de diffusion.

La mise à disposition des résultats et bases de données s'accompagne systématiquement de l'ensemble des métadonnées, y compris de la documentation, nécessaires à leur bonne utilisation, à rebours de la pratique organisée par certaines entreprises privées ne délivrant pas nécessairement leurs métadonnées.

Aujourd'hui, l'Insee continue de compléter son offre en matière de données par l'ouverture de trois nouvelles API depuis 2020 (3).

Le rapporteur spécial appelle toutefois à ce que les API soient moins limitatives. En effet, la possibilité de formuler, par exemple pour l'API Sirene,

---

(1) Une interface de programmation d'application, dite API, constitue une interface logicielle qui permet de connecter un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.

(2) Rapport au Premier ministre de l'administrateur général des données 2016-2017, « la donnée comme infrastructure essentielle ».

(3) L'API Données locales (statistiques pour tous les zonages géographiques allant de la commune à la France entière), l'API Banque de données macroéconomiques et l'API métadonnées.

uniquement trente appels par minute pour un demandeur de données, restreint le champ des informations qui sont effectivement disponibles.

Toutefois, si la loi de 2016 n'a pas eu pour l'Insee une incidence comparable à celle qu'elle a eue pour d'autres administrations (1) en raison de l'ouverture précoce de ses données, cette politique a tout de même eu pour conséquence de mettre en extinction les redevances pour réutilisation des données que percevait auparavant l'Institut (2).

b. L'ouverture des données de l'Insee engendre une perte sèche de recettes

Auparavant l'Insee monétisait un certain nombre de services rendus et bénéficiait ainsi de redevances constituant des ressources propres. Mais la loi pour une République numérique a prévu une réutilisation gratuite des données des administrations, prohibant la possibilité de vendre des données à des opérateurs privés.

La politique d'*open data* a constitué, pour l'Insee, une perte sèche de recettes qui a été compensée entièrement par l'État.

Les redevances précédemment perçues, en provenance pour l'essentiel de la vente de données provenant du fichier Sirene, représentaient une contribution au budget de fonctionnement de l'Insee (hors dépenses de titre II). La suppression des redevances a toutefois été assortie d'une compensation budgétaire à hauteur de 11 millions d'euros, prévue en loi de finances pour 2017.

Par ailleurs, les attributions de produits (3) au profit de l'Institut connaissent, depuis 2017, une trajectoire baissière qui s'explique par la diminution de conventions payantes signées entre l'Insee et ses partenaires institutionnels.

Aussi, bien que relative et limitée, la politique d'*open data* a eu comme effet secondaire de réduire les ressources propres perçues par l'Institut.

*c. L'***open data** *est toutefois une réussite nationale*

Le rapporteur spécial a souhaité mettre en avant la réussite que constitue la politique d'*open data*. La France est très régulièrement dans le *trio* de tête des classements européens en matière de mise en œuvre de cette politique.

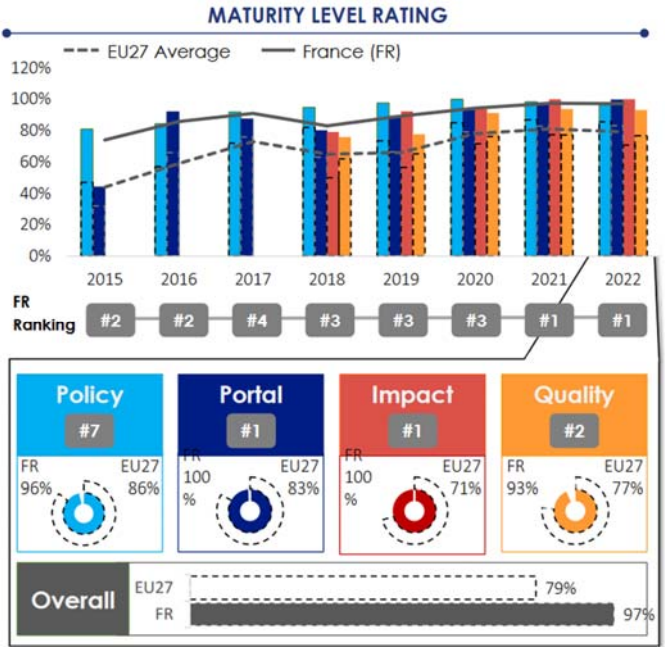
Pour 2021 et 2022, la France est même première au classement établi dans l'*Open data maturity report* réalisé par la Commission européenne depuis 2015.

---

(1) Le rapport du Conseil d'État précédemment cité estime que « les administrations françaises peinent à basculer entièrement, résolument et irréversiblement dans l'ère de la donnée. La politique de la donnée s'apparente encore à un exercice imposé ». Par ailleurs, le rapport conjoint Insee/Dinum intitulé « l'évaluation des besoins de l'État en compétences et expertises en matière de données » estime nécessaire d'acculturer les administrations à la donnée afin que celles-ci favorisent un partage plus ouvert des données.

(2) Selon l'article L. 324-6 du CRPA : « La réutilisation des informations publiques produites par le service statistique public […] ne peut donner lieu au versement d'une redevance. »

(3) Ces dernières sont constituées pour l'essentiel de la gestion de l'identifiant d'entité juridique (LEI) et de rémunérations de services rendus par l'Insee.



Source : Commission européenne, Open data maturity report, 2022.

Toutefois, le rapporteur spécial relève que des progrès sont possibles pour la mise à disposition de jeux de données [(1)]. En effet, les jeux de données mis à disposition par la France connaissent une trajectoire stagnante en volume (entre 150 000 et 200 000 jeux de données sur la période 2020 à 2023), à l'inverse de l'Allemagne qui continue de mettre à disposition toujours plus de jeux de données (volume compris entre 200 000 et 600 000 jeux sur la même période).

2. L'accessibilité des données de l'Insee et leur bonne compréhension deviennent des objectifs cardinaux

a. Des investissements répondant aux enjeux d'accessibilité des données

L'Institut assure aujourd'hui une véritable politique didactique visant à rendre la donnée accessible à tous les publics. Comme l'énonçaient les représentants de l'Insee à l'occasion des auditions, l'objectif est désormais de « *produire du signal et pas seulement du bruit* ».

L'Insee garantit ainsi un important rôle de diffusion de la donnée qui vise tant à réduire les phénomènes d'anti-sélection qu'à informer un public, lequel peut être non-averti.

---

(1) Un jeu de données, également appelé ensemble de données ou dataset *en anglais, fait référence à une collection structurée ou non structurée de données qui sont regroupées en tant qu'entité distincte. Un jeu de données peut contenir des informations de différents types, tels que des chiffres, du texte, des images, des vidéos, des enregistrements audio, etc.*

Le site internet de l'Institut, dont le fonctionnement est assuré par les crédits ouverts à l'action 08 – *Information économique, démographique et sociale* du programme 220 de la mission *Économie*, connaît un succès certain avec plus de 30 millions de visites en 2022 (en hausse de 16 % par rapport à 2021). Les informations disponibles sur le site, retracées par thèmes et échelles géographiques, rendent l'information aisément disponible.

L'Insee insiste, ces dernières années, sur l'accessibilité de la donnée en tant que ressource exploitable par les utilisateurs, au-delà de la seule production statistique. Par le biais de focus groupes et d'enquêtes auprès des utilisateurs, le site internet se transforme, notamment avec le développement de la « datavisualisation » et l'amélioration de l'accessibilité et de la lisibilité des statistiques (amélioration du repérage des informations et de la navigation, mise à disposition de versions html intégrales pour toutes les publications en complément des versions pdf, données intégralement accessibles en téléchargement).

L'Insee s'adresse également à des utilisateurs experts et potentiellement diffuseurs de données. Ce faisant, il veille à constamment progresser dans l'échelle de l'*open data*. Les interfaces de programmation d'application (API) remplissent alors pleinement leur rôle en permettant une interopérabilité « machine-machine » encourageant une propagation plus fluide des données.

L'Insee tente par ailleurs de toucher de nouveaux publics, en étendant son canal de diffusion des données.

b. Un renforcement de la pédagogie de l'Insee dans l'exploitation des données

La création, en février 2020, du « blog de l'Insee » répond à cet objectif de renforcer la pédagogie. Ce canal de communication, moins institutionnel, permet d'avoir un ton plus didactique permettant d'exposer les avantages et limites de l'utilisation des données. En outre, ce canal est également vertueux puisqu'il entend lutter contre la propagation de fausses informations.

Dans le même sens, l'ouverture en juin 2021 de la première application mobile pour smartphone « Insee Mobile » permet également de découvrir les données grâce à une sélection de chiffres-clés et d'actualités déclinée en différentes thématiques.

Pour autant, cette digitalisation de la pédagogie n'est pas la solution unique. Les représentants syndicaux de l'Insee auditionnés ont fait valoir que l'*open data* constitue plutôt une marche arrière en matière de pédagogie, notamment en raison de la disparition de certains livrables écrits permettant de présenter clairement les travaux.

Cette ouverture des données publiques n'est néanmoins pas totale dans la mesure où révéler certaines informations pourrait conduire à rompre avec le principe même de secret statistique qui guide l'activité de l'Insee.

Des acteurs de la société civile, tels que les associations, peuvent pourtant être intéressés par une ouverture plus complète de l'*open data*. À titre d'illustration, l'association Ouvre-boîte milite pour une « *libération des documents de l'administration* » et spécialement des bases de données et codes sources.

L'Institut a notamment été saisi d'une demande de mise à disposition des données et méthodes de construction de l'indice des prix à la consommation (IPC). Or, l'Insee a été confronté à un dilemme. Si cette demande a pu apparaître comme légitime et conforme au droit applicable, il s'est avéré que révéler certains indices élémentaires risquait de conduire à la possibilité d'identification de points de vente en raison du caractère fin de la donnée fournie (précision géographique et chiffre d'affaires de l'établissement dans un espace donné) rompant ainsi avec le secret statistique.

L'association Ouvre-boîte avait réclamé la mise à disposition par l'Insee, en outre, de la liste des 30 000 points de vente qui alimentent les relèves. L'Institut a refusé de délivrer les informations concernées, malgré l'avis favorable de la Commission d'accès aux documents administratifs (CADA). L'association a saisi le tribunal administratif de Paris sur ce point qui, dans un jugement rendu le 9 février 2023, a reconnu que l'Insee « *a méconnu l'obligation qui lui incombe en vertu des dispositions de l'article L. 311-1 du code des relations entre le public et l'administration* », annulant sur ce point le refus de communication des documents [(1)]. Toutefois, le juge administratif a estimé, dans plusieurs considérants, que d'autres demandes de communications formulées entraient en contradiction avec le secret statistique.

Le monde de la recherche se fait également le relais des limites de l'ouverture des données. Pour Samuel Goëta, les données publiées sur les portails *open data* répondent à une logique d'offre, c'est-à-dire que les administrations choisissent les données qu'elles souhaitent ouvrir sans toujours prendre en compte les besoins de l'usager.

Ce faisant, l'*open data* revêt parfois une dimension minimale. Les stocks de données mis à disposition du public sont parfois renouvelés à des fréquences qui ne permettent pas d'apprécier la situation contemporaine. Les données relatives à la base Sirene gérée par l'Insee sont ainsi actualisées tous les 30 jours, ce qui peut apparaître comme une échelle de temps impropre à la lecture conjoncturelle de la situation des entreprises en France.

Par ailleurs, les données sont traitées de manière multiple, ce qui peut rendre parfois leur usage délicat [(2)].

---

(1) TA de Paris, 9 février 2023, Association ouvre-boîte, n° 2109576/5-2.

(2) Samuel Goëta, « Instaurer des données, instaurer des publics : une enquête sociologique dans les coulisses de l'open data », Télécom ParisTech, 2016.

Le rapporteur spécial regrette que, dans le cadre de son travail d'information, il n'ait pas disposé d'un temps suffisant pour auditionner les acteurs du monde de la recherche.

*

Si la politique d'*open data* semble constituer une opportunité pour l'Insee, elle demeure un défi. D'abord, en ce qu'elle suppose une véritable démarche pédagogique pour rendre accessible les données. Ensuite, en ce que les données administratives recueillies peuvent être minimales. Enfin et surtout, en raison de difficultés – limites de l'encadrement normatif et concurrence avec les acteurs privés – auxquelles fait face l'Institut.

II. L'ACTIVITÉ DE L'INSEE ÉVOLUE DANS UN CONTEXTE NORMATIF CORSETANT L'ACCÈS ET L'UTILISATION DE CERTAINES DONNÉES, EN CONCURRENCE AVEC DES ACTEURS PRIVÉS

L'accès et le traitement des données sont strictement encadrés sur le plan normatif. Ce cadre peut paraître vétuste, peu lisible et mouvant, ce qui constitue un frein à l'activité statistique de l'Insee (A). Surtout, dans l'accès et le traitement, l'Institut est concurrencé par des acteurs de la société civile disposant de données intéressant directement la statistique publique (B).

A. LES PRÉROGATIVES DE L'INSEE SONT CORSETÉES PAR UN CADRE NORMATIF ANCIEN ET PEU LISIBLE

1. Un cadre normatif complexe ne permettant pas de tirer entièrement profit de la multiplication des données

a. L'utilisation tant des données publiques que des données privées est affectée de lourdeur

Si la politique d'*open data* a été conçue comme une ouverture du public vers le secteur privé, le rapporteur a été étonné de constater que cette ouverture n'a pas permis de fluidifier les échanges de données publiques entre les administrations.

De manière paradoxale, le cadre juridique en vigueur limite le partage des données publiques entre les administrations.

Les administrations et l'Insee n'ont pas plus de droit, compte tenu de la rédaction de l'article 1er de la loi pour une République numérique, que le grand public dans l'accès aux données des autres administrations. Le législateur n'a pas prévu de dispositifs spécifiques permettant de simplifier et d'intensifier les échanges de données publiques entre les administrations.

Les administrations sont paradoxalement corsetées dans leur mission de service public, elles ne peuvent revendiquer le bénéfice du principe de la libre réutilisation des données prévu à l'article L. 321-1 du code des relations publiques et l'administration (CRPA). En effet, *« l'échange d'informations publiques entre des administrations, aux fins de l'exercice de leur mission de service public* » ne constitue pas une réutilisation au sens de l'article précité (article L. 321-2 du CRPA).

Au surplus, le régime général de la réutilisation des données publiques s'entend d'une manière restrictive. La communication des informations entre administrations porte sur les « documents administratifs » et non pas directement sur les données, ce qui suppose l'existence d'un support écrit, communiqué ou régulièrement publié, au sein duquel figurent des données. Il ne s'agit pas d'un accès direct aux codes sources et métadonnées, pourtant nécessaire à l'établissement de statistique publique.

Fort de ce constat, le **rapporteur spécial** partage les remarques du député Éric Bothorel dans son rapport « *Pour une politique publique de la donnée* ». Il est en effet étonnant de constater que l'administration bénéficie, pour ses missions de service public, moins de ses propres données publiques que les acteurs du secteur privé.

Alors que la politique d'*open data* constitue une véritable percée conceptuelle dans le monde de l'accès aux données, le cadre juridique applicable à l'Insee demeure fondé sur un texte ancien, la loi de 1951, parfois inadapté au contexte actuel.

Tout d'abord, les articles 6 et 7 *bis* de la loi de 1951 constituent des freins à l'accès de certaines données. Ces deux articles consacrent une définition non homogène du secret statistique en prévoyant des dérogations et des interdictions plus ou moins approfondies selon la nature de la donnée recueillie. L'absence d'uniformité de cette définition est vecteur de lourdeur pour les services statistiques qui doivent en permanence veiller à demeurer dans les canons de conformité applicables aux données traitées.

De manière plus fondamentale, les limites posées par l'article 3 *bis* de la loi de 1951 sont un frein notable à l'utilisation des données de nature privée. Or celles-ci constituent le point d'achoppement actuel pour l'Insee, l'empêchant d'aller de l'avant dans l'exploitation des données privées à des fins d'enquête statistique. En effet, cet article restreint la possibilité de recourir à des données de nature privée sauf à remplir trois conditions cumulatives :

D'abord, en amont de l'utilisation des données, une concertation préalable avec les détenteurs des données doit être établie ; ensuite, une étude de faisabilité et d'opportunité, dont les critères sont définis par voie réglementaire [(1)], doit permettre de s'assurer que les informations présentes dans les bases de données répondent aux objectifs de l'enquête et améliorent la connaissance du secteur ; enfin, les données mobilisées doivent par ailleurs répondre à « des besoins d'enquêtes statistiques rendues obligatoires en application de l'article 1er *bis* ». Ainsi, les données privées mobilisées au titre de l'article 3bis doivent venir en remplacement d'enquêtes existantes.

L'inadaptation du cadre est telle que l'Insee a aujourd'hui essentiellement recours à des conventions de gré à gré avec les acteurs privés – ce qui constitue, en creux, un contournement de l'article 3 *bis*.

(1) Décret n° 2017-463 du 31 mars 2017 portant application de l'article 3 bis *de la loi du 7 juin 1951 relative à l'obligation, la coordination et le secret en matière statistique.*

b. L'encadrement européen délimite le traitement des données et limite le recours aux données privées pour les services statistiques

Le droit de l'Union européenne s'est révélé particulièrement attentif à encadrer l'utilisation des données personnelles à des fins de production statistique.

Le règlement du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation des données (dit règlement RGPD) (1) a prévu plusieurs obligations pour les services publics de statistiques :

Certaines de ces obligations sont de nature éthique. Lors de la collecte des données, un consentement libre, éclairé et univoque de la personne dont les données sont collectées doit être recueilli par les services de statistique. L'étendue des données collectées doit respecter le principe de minimisation des données, suivant lequel, ne doivent être collectées que les données qui sont directement nécessaires et adaptées à la production de statistiques.

Certaines obligations visent directement à prévenir des atteintes à la vie privée. Les données recueillies doivent faire l'objet d'une anonymisation et d'une pseudonymisation, ce qui limite les possibilités d'identification de la personne dont les données ont été collectées. Celles-ci doivent par ailleurs être protégées des divulgations, des accès non-autorisés, des altérations et des destructions, ce qui nécessite la mise en place de systèmes de sécurité des données appropriés.

La réglementation européenne a également une incidence sur l'accès à un certain type de données, à l'instar des données de télécommunications.

La directive dite vie privée et communications électroniques (*ePrivacy Regulation*) (2) restreint l'accès aux données à caractère personnel, notamment celles transitant par les communications électroniques au sein de l'Union européenne, au nom du respect de la vie privée.

Cet acte de droit dérivé limite les possibilités d'utiliser les données de signalisation pour des finalités statistiques et de recherche sans le consentement exprès de chaque détenteur de téléphone portable. Le texte fait l'objet de discussions depuis 2017, mais les difficultés se cristallisent précisément autour de la définition des conditions dans lesquelles les données de connexion peuvent être réutilisées.

Ce cadre juridique empêche aujourd'hui l'utilisation massive des données de téléphonie mobile à des fins statistiques. Celles-ci représentent pourtant un enjeu tout particulier dans la mesure où elles permettent de connaître, pratiquement en temps réel, la présence d'une population sur un territoire.

---

(1) Règlement UE 2016/679 du Parlement européen et du Conseil du 27 avril 2016.

(2) Directive 2002/58/CE du Parlement européen et du Conseil du 12 juillet 2002 concernant le traitement des données à caractère personnel et la protection de la vie privée dans le secteur des communications électroniques (directive vie privée et communications électroniques).

2. Un cadre normatif en pleine transformation, source d'incertitudes et de doutes pour l'Insee

Au niveau européen, trois textes actuellement en discussion sont susceptibles d'emporter des effets sur la statistique publique et l'Insee.

En premier lieu, le *Data Governance Act* (DGA) adopté en mai 2022 et dont l'entrée en vigueur est prévue pour septembre 2023 vise à encourager le partage volontaire de données, en favorisant la réutilisation des données du secteur public confidentielles, même si elles sont protégées par le secret statistique.

Le traitement des demandes d'accès aux données couvertes par le secret statistique

L'accès aux données couvertes par le secret statistique relève de la compétence du Comité du secret statistique. Les demandes d'accès à ces données sont déposées et gérées via un portail, le Centre d'accès sécurisé aux données (CASD). Ce portail est placé sous la responsabilité du directeur général de l'Insee, lequel est le coordonnateur du service statistique public.

Après l'examen du dossier de demande par le Comité du secret statistique et l'obtention de l'accord des services producteurs, puis la validation des Archives nationales, la mise à disposition des données confidentielles par l'intermédiaire d'une plateforme sécurisée est assurée par le CASD.

L'Insee se montre favorable à une telle démarche pour plusieurs raisons. D'abord, cela permettrait de pallier la faible diffusion des données entre administrations publiques identifiées précédemment, en raison d'une meilleure visibilité sur l'ensemble des données administratives disponibles au niveau national, ainsi que d'améliorer les relations de travail entre les services de statistiques et les autres administrations. Ensuite, cela permettrait de capitaliser sur les investissements déjà réalisés en matière de développement de plateformes permettant un partage sécurisé de la donnée, telles que le CASD par exemple.

L'Insee aspire également à tirer profit des dispositions relatives à « *l'altruisme des données* » prévu par le DGA. Cette mesure consiste en une mise à disposition volontaire de données par des particuliers ou des entreprises pour des objectifs d'intérêt général. Cependant, il est pour l'heure difficile de prédire l'ampleur des données qui seront concernées par ces échanges, car la diffusion altruiste est purement tributaire des arbitrages individuels des particuliers et des entreprises.

En second lieu, le *Data Act*, adopté par la Commission européenne en février 2022 et validé par le COREPER en mars 2023 présentait, à l'origine, une ambition forte en matière d'accès et de partage des données.

Cependant, ce texte constitue, pour les services de l'Insee auditionnés, essentiellement une déception sinon un vecteur de craintes. L'ambition du texte était de faciliter le recours à des données privées pour les acteurs publics et notamment

les services de statistiques. Or le périmètre retenu pour l'utilisation des données privées apparaît étriqué, car il ne prévoit un recours à celles-ci qu'en cas « d'urgence publique » [(1)] c'est-à-dire pendant une période strictement délimitée et pour des finalités bien précises. De telles conditions ne sont pas pertinentes pour l'établissement de statistiques qui suppose un accès aux données pérenne. Le *Data Act* n'ayant pas permis d'avancer suffisamment sur l'accès aux données privées, les espérances se concentrent maintenant sur la révision en préparation de la loi statistique européenne (Règlement 223/2009). Ce texte pourrait être l'occasion de dépasser les limites imposées par l'article 3 *bis* de la loi de 1951 à la faveur d'un cadre renouvelé et clarifié sur l'accès aux données privées. Néanmoins, tout cela demeure au stade de l'expectative, une proposition de révision du Règlement pouvant être adoptée par la Commission au milieu de l'année 2023.

*

Le cadre normatif applicable à la production de statistiques publiques est à la fois ancien et peu adapté à l'évolution digitale, laquelle encourage une croissance des données disponibles. Tout spécialement, ce cadre apparaît dépassé sur le plan de l'accès aux données privées. Or, les évolutions attendues au niveau européen ont été décevantes ou sont encore sujettes à des incertitudes. Les acteurs privés quant à eux ont accès à de nombreuses données, faisant de ces derniers des potentiels concurrents de l'Insee ou, à tout le moins, des détenteurs de données attrayantes pour les services statistiques.

B. L'INSEE EST CONCURRENCÉ, TANT DU POINT DE VUE DE L'EXPLOITATION QUE DE L'ACCÈS AUX DONNÉES, PAR LES ACTEURS PRIVÉS

1. Les grands acteurs de la donnée bénéficient d'un accès plus étendu aux données que l'Insee

a. Les Big Tech ont un accès exponentiel aux données

Le secteur public ne peut plus aujourd'hui s'affranchir des acteurs privés dans le domaine des données. Ces derniers disposent d'une masse de données plus importante et ont également des capacités de traitement parfois supérieures à celles des acteurs publics en raison de politiques d'investissement dans l'exploitation des données.

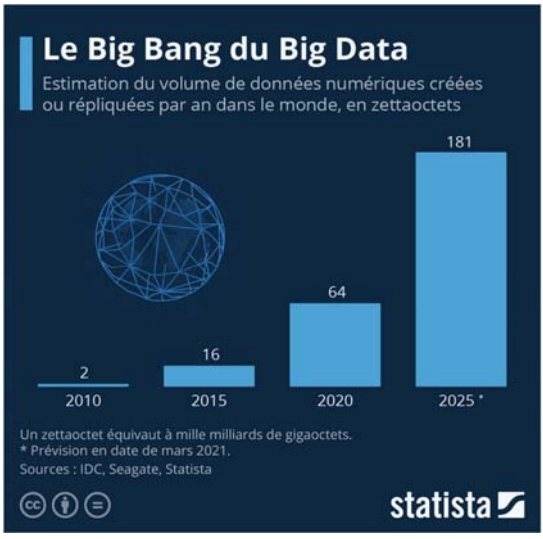
Les investissements réalisés par les grandes entreprises du numérique reposent à la fois sur **les capacités de stockage,** via la création de grands centres de

(1) Le texte définit l'« urgence publique » comme une situation exceptionnelle, telle que les urgences de santé publique, les urgences résultant de catastrophes naturelles, ainsi que les catastrophes majeures d'origine humaine, comme les incidents majeurs de cybersécurité, affectant négativement la population de l'Union, d'un État membre ou d'une partie de celui-ci, avec un risque de répercussions graves et durables sur les conditions de vie ou la stabilité économique, ou de dégradation substantielle des actifs économiques dans l'Union ou le ou les États membres concernés, et qui est déterminée et déclarée officiellement selon les procédures respectives prévues par le droit de l'Union ou le droit national.

données permettant de recueillir un volume massif de données, sur des infrastructures permettant la réalisation de calculs complexes, ainsi que sur **des investissements dans les technologies de réseaux et de connectivité de haut débit** pour assurer la rapidité et la fiabilité de l'accès aux données.

Les services statistiques des États accusent souvent un retard technologique par rapport à certaines entreprises – les *Big tech* (1) – et se retrouvent ainsi contraints de sous-traiter un certain nombre de services auprès de ces entités (2). Ces dernières ont recours à des données issues de leur propre activité commerciale. Ainsi, Google a annoncé traiter plus de 3,5 milliards de requêtes de recherche par jour, ce qui constitue une quantité de données particulièrement importante.

Le marché de la donnée est actuellement en plein essor. Selon des études de l'*International Data Corporation* (IDC), le volume mondial de données devrait augmenter selon un taux de croissance annuel moyen d'environ 26 % d'ici à 2025. Cela signifie que la quantité de données devrait doubler tous les deux ans. La Commission a estimé quant à elle que le volume mondial des données devrait augmenter de 530 % à horizon 2025 (3). Même si les écarts d'estimation sont considérables, il en ressort que l'accroissement du volume de données disponibles est en expansion.



Pour prendre la mesure d'une telle évolution, « *il faudrait se procurer environ 640 millions des plus gros disques SDD actuellement commercialisés (disposant d'une capacité de 100 To de stockage) pour être à même de sauvegarder les 64 zettaoctetts de données générées en 2020* » selon une étude réalisée par Statista au cours de 2021 (4).

Source : Statista, présentant une estimation du volume annuel de données numériques générées dans le monde entre 2010 et 2025.

---

(1) Les grands groupes sont connus : Alphabet, Apple, Meta, Amazon, Microsoft, ect.

(2) Revue Pouvoirs 2023/2 (n° 185), Asma Mhalla, « Les Big Tech, de nouveaux États parallèles ».

L'auteur remarque à cet égard que : « En France par exemple, depuis 2015, la Direction générale de la sécurité intérieure coopère avec Palantir, qui lui fournit des logiciels de traitement de données ainsi que des logiciels de cybersécurité, et ce, en lien avec des fonctions régaliennes particulièrement sensibles en matière de sûreté national ».

(3) Commission européenne, Open data maturity report, 2019.

(4) Statista, « Le Big Bang du Big Data », article écrit par Tristan Gaudiaut le 19 octobre 2021.

Le **rapporteur spécial** appelle à être vigilant sur les conséquences de l'accroissement des données en circulation, qui pose la question de la protection de la vie privée.

b. La donnée représente une manne financière pour les acteurs privés

La donnée acquiert une valeur économique sur l'ensemble de sa chaîne de valeur qui peut être décomposée en trois temps :

La **donnée est d'abord une matière première** qui peut être vendue par les entreprises via des brokers (1) ; La **donnée peut ensuite être mobilisée comme levier et gain de productivité** permettant de perfectionner les systèmes d'information, notamment via le *deep learning* ; la donnée peut enfin revêtir la dimension d'un actif stratégique sur un marché, les études de Jean Tirole et Jean Charles Rochet ont montré que la donnée était un élément central des « marchés bifaces » (2) sur lesquels reposent les plateformes numériques (Facebook, Google, Twitter, etc.).

Pour les détenteurs de données, leur exploitation représente un enjeu financier particulièrement significatif. L'enjeu est d'autant plus fort que le marché mondial de la donnée tend à s'accroître de manière exponentielle. En 2020, le marché du *Big Data* mondial, représentait environ 200 milliards d'euros de chiffre d'affaires et, pour la France, près de 4 milliards d'euros (3).

Dans de telles conditions, l'Association française des entreprises privées (Afep) fait valoir que la mise à disposition gratuite de données privées est susceptible de déstabiliser les équilibres d'investissement et commerciaux existants et d'avoir un effet désincitatif. Ainsi l'Afep estime que les données transférées ne doivent pas devenir publiques et qu'un droit de refus du propriétaire des données à destination des organismes du secteur public devrait être établi dans les textes européens en discussion (4). Le droit de propriété intellectuelle devrait demeurer protecteur des données détenues par les personnes privées.

Dans ces conditions, **le rapporteur spécial** estime qu'une plus grande ouverture des données privées au public suppose d'engager au préalable une concertation entre les acteurs concernés. Cette concertation devrait notamment évoquer : L'étendue et la finesse des données privées ainsi rendues accessibles ; les modalités d'accès à ces données ; les éventuelles compensations à prévoir.

---

(1) Selon une définition retenue dans un Rapport conjoint de l'Inspection générale des finances (IGF), du Conseil d'État et du Conseil général de l'économie, relatif aux données d'intérêt général (septembre 2015) : « les data brokers sont des sociétés (comme Acxiom ou Bluekai aux Etats-Unis) spécialisées dans la collecte de données, généralement personnelles, récupérées sur les sites publics internets ou achetées auprès d'autres entreprises (sites de commerce en ligne, entreprise de grande distribution, etc.). Ces données sont ensuite agrégées, triées et vendues à des entreprises souhaitant par exemple mieux connaître leurs clients ou cibler leurs publicités ».

(2) Jean-Charles Rochet and Jean Tirole, « Platform Competition in Two-Sided Markets » Journal of the European Economic Association, June 2003.

(3) Rapport CESE, « Économie et gouvernance de la donnée », février 2021.

(4) Position de l'Afep sur la proposition de règlement européen sur les données (Data Act) de juillet 2022.

2. Les données des entreprises privées sont couvertes par un ensemble normatif protecteur, limitant les possibilités de réutilisation par le secteur public en l'absence de cadre approprié

a. Les données privées sont protégées

Les personnes privées sont titulaires de certains droits sur les données dont elles disposent. Les pouvoirs publics doivent en tenir compte et ne sauraient agir d'une manière telle que ces droits seraient méconnus.

Tout d'abord, l'utilisation des données par une personne privée – à des fins d'exploitation commerciale ou non – relève de la liberté d'entreprendre, principe dont la valeur constitutionnelle a été reconnue dès 1982 par le Conseil constitutionnel (1).

Par ailleurs, ces données, qui révèlent souvent des informations sur des personnes physiques, posent également la question du droit au respect de la vie privée, lequel est une composante de la liberté personnelle également garantie par la Constitution. C'est la raison pour laquelle le Conseil constitutionnel juge que « *la collecte, l'enregistrement, la conservation, la consultation et la communication de données à caractère personnel doivent être justifiés par un motif d'intérêt général et mis en œuvre de manière adéquate et proportionnée à cet objectif* » (2).

Ensuite, le droit de l'Union européenne protège la détention des données dont disposent les acteurs privés. L'article 7 de la directive 96/9/CE (3) concernant la protection *sui generis* des bases de données prévoit que le producteur d'une base de données dispose du droit « ***d'interdire l'extraction ou la réutilisation de la totalité ou d'une partie substantielle […] du contenu de celle-ci****, lorsque l'obtention, la vérification ou la présentation de ce contenu attestent un investissement substantiel du point de vue qualitatif ou quantitatif* ».

À cet égard, le droit français a largement transposé les dispositions européennes de protection des bases de données détenues par des entités privées. Le code de la propriété intellectuelle (CPI) reconnaît des droits spécifiques aux producteurs de bases de données. Les articles L. 342-1 et suivants de ce code prévoient les limites à l'extraction et à la réutilisation des données dès lors qu'un investissement substantiel a été entrepris pour pouvoir réaliser ces bases de données.

Il n'existe aujourd'hui en droit aucun dispositif permettant de contraindre à l'ouverture inconditionnelle et automatisée des données privées pour le secteur public à des fins de statistiques publiques.

---

(1) Décision n° 81-132 DC du 16 janvier 1982.

(2) Voir en particulier la décision n° 2012-652 DC du 22 mars 2012. Cette exigence vaut pour les fichiers privés (décision n° 2004-499 DC du 20 juillet 2004) mais également pour les fichiers publics (décision n° 2003-467 DC du 12 mars 2003).

(3) Directive 96/9/CE du Parlement européen et du Conseil du 11 mars 1996 concernant la protection juridique des bases de données.

b. Le recours aux données à des fins d'intérêt général, une notion ambitieuse, mais définie de manière trop réductrice

Dès 2015, le rapport relatif aux données d'intérêt général dirigé par l'Inspection générale des finances, le Conseil d'État et le Conseil général de l'économie de l'industrie, de l'énergie et des technologies avait mis en avant la nécessité de fonder une notion de « donnée d'intérêt général ».

À l'époque, de nombreuses réticences avaient été émises, notamment sur les risques d'inconstitutionnalité et d'inconventionalité d'une loi laissant une trop grande latitude d'interprétation aux autorités administratives chargées de définir, au cas par cas, la notion de donnée d'intérêt général.

En introduisant dans la loi pour une République numérique la notion de donnée d'intérêt général, le législateur a créé une notion devant favoriser l'utilisation de données privées.

Toutefois, la notion de donnée d'intérêt général ne cible que certains acteurs : d'une part, les acteurs privés concessionnaires d'un service public (article 17), d'autre part, les personnes chargées d'une mission de service public industriel et commercial (article 18). Le lien de sujétion à la puissance publique apparaît ainsi comme un critère nécessaire à la qualification de « donnée d'intérêt général ». Pour autant, ces données ne sont pas assimilables à des données privées, mais davantage à des documents administratifs, dont la communication est encadrée par les articles L. 311-5 et L. 311-6 du code des relations entre le public et l'administration (CRPA).

En 2020, le rapport du député Éric Bothorel pour une politique publique de la donnée, révélait l'échec de la loi pour une République numérique à consacrer une notion juridique claire qui révélerait l'existence de données d'intérêt général « par nature ».

Enfin, en 2022, les propositions de textes déposées au niveau européen, qu'il s'agisse du *Data Governance Act* ou du *Data Act*, n'ont pas permis d'entrevoir l'émergence d'une donnée d'intérêt général.

Pourtant, une définition moins limitative de cette notion de donnée d'intérêt général pourrait être utile aux services statistiques.

Le rapporteur spécial estime ainsi nécessaire d'apporter une définition législative large de la notion de « donnée d'intérêt général » afin de faciliter une utilisation des données privées à des fins de statistiques publiques. Il insiste sur le fait que la notion de donnée d'intérêt général doit permettre une mobilisation permanente des données pour que celles-ci soient utiles aux services statistiques comme l'Insee.

Proposition n° 1. Établir une définition législative large de la notion de donnée d'intérêt général et de ses modalités d'utilisation à des fins de production statistique, afin de permettre un accès facilité à ces données pour l'Insee et les services statistiques ministériels.

Cette définition doit permettre d'identifier des données qui revêtent, par nature, des caractéristiques d'intérêt général. La liste des catégories de données est établie par décret en Conseil d'État.

Le rapporteur spécial a estimé par ailleurs que cette notion pouvait être, dans un premier temps, expérimentée dans le domaine environnemental. Plusieurs études ont démontré l'intérêt des statistiques publiques pour agir sur le changement climatique.

L'intérêt de la consécration de la notion de donnée d'intérêt général : l'exemple de la donnée environnementale.

Dans un avis du Conseil national du numérique de juillet 2020 intitulé « *faire des données environnementales des données d'intérêt général* », le Conseil a mis en avant l'intérêt, pour le secteur environnemental, de faire naître une telle notion.

Retenant une conception large de la notion, le Conseil national du numérique estime que toutes les données produites dans un cadre privé, associatif ou encore par les citoyens pourraient être utilement mises à profit par les collectivités publiques.

Le domaine environnemental serait particulièrement propice à l'émergence d'une telle notion, car la Charte de l'environnement de 2004 dont la valeur constitutionnelle a été reconnue (1) appréhende l'environnement comme « un bien commun des êtres humains ». Selon eux, il est possible « *de relier la notion de données environnementales d'intérêt général à celle de patrimoine commun* » (2).

Récemment, dans une étude de France Stratégie réalisée par Jean Pisani-Ferry parue le 22 mai 2023 intitulé « *Les incidences économiques de l'action pour le climat* », l'économiste consacre une partie de son argumentaire aux « indicateurs et données » (3) qui pourraient être utilement mis à profit de l'action publique pour agir sur le climat.

Proposition n° 2. Permettre, pour les données en lien avec les politiques publiques environnementales, une expérimentation d'une définition législative large de la notion de donnée d'intérêt général.

L'évaluation de cette expérimentation permettrait d'apprécier dans un second temps l'intérêt de la généralisation du dispositif. Au-delà de l'intérêt de la

(1) Décision n° 2008-564 DC du 19 juin 2008.

(2) Rapport du Conseil national du numérique, « Faire des données environnementales des données d'intérêt général », juillet 2020.

(3) Rapport thématique France Stratégie, Nicolas Carnot et Nicolas Riedinger, « indicateurs et données », paru au sein du rapport « les incidences économiques de l'action pour le climat » le 22 mai 2023.

consécration d'une notion de donnée d'intérêt général à des fins environnementales, le rapporteur spécial estime que plusieurs secteurs de politique publique pourraient également en tirer profit. Par exemple, pour le secteur de la santé, alors que le Conseil constitutionnel a eu l'occasion de consacrer la protection de la santé publique comme principe à valeur constitutionnelle [(1)], il serait possible de considérer que l'utilisation de données privées, à ces fins spécifiques, poursuit des finalités d'intérêt général.

(1) Décision n° 90-283 DC du 8 janvier 1991.

III. AFIN DE CONFORTER LA QUALITÉ DE L'INFORMATION STATISTIQUE, L'INSEE ENTREPREND UNE ÉVOLUTION DE SES PRATIQUES

Si l'Insee s'est longtemps limité à l'utilisation des ressources tirées des enquêtes et des données administratives, la crise sanitaire a eu pour effet de faire évoluer ses pratiques.

Désormais, l'Institut entend entreprendre des démarches partenariales avec les acteurs de la société civile pour tirer profit de la diversité des données qu'offre un monde numérisé (A).

Mais pour permettre une véritable utilisation par l'Insee et les services statistiques des données issues du secteur privé, le cadre juridique aujourd'hui applicable doit évoluer vers une plus grande clarté en vue de faciliter les partenariats (B).

A. L'INSEE S'EST ENGAGÉ DANS UNE DÉMARCHE PARTENARIALE AVEC LES ACTEURS PRIVÉS, ÉTENDANT L'ACCÈS À DE NOUVELLES DONNÉES

1. Le développement des approches partenariales : apports et limites

a. Le partenariat avec La Banque Postale : l'exemple d'une réussite en matière de mise à disposition des données privées

La crise sanitaire a révélé la pertinence et l'intérêt pour les acteurs publics d'avoir accès aux données de nature privée. Coupé de ses possibilités d'enquêtes par collecte directe, l'Insee s'est orienté vers des données alternatives, jusqu'alors peu explorées dans le cadre des enquêtes statistiques. Les seules données de nature privée qui avaient été mobilisées jusqu'alors concernaient les données de caisse de la grande distribution. Ce temps de crise a agi comme un accélérateur.

L'Insee a ainsi fait évoluer sa doctrine en matière d'exploitation des données.

L'Institut s'est tourné vers plusieurs établissements bancaires, à savoir La Banque Postale (LBP) et le Crédit Mutuel Alliance Fédérale, afin d'estimer, les effets de la crise sanitaire sur les ménages français. L'Institut n'a pas été la seule institution publique à avoir un tel réflexe : le Conseil d'analyse économique a également sollicité les services de LBP en vue d'obtenir des données avec des fréquences élevées.

Le rapporteur spécial a souhaité, dans un premier temps, s'intéresser spécifiquement au partenariat avec La Banque Postale, ce dernier constituant un modèle d'accord de gré à gré vertueux.

Contactée dès le début de la crise sanitaire, au moment du premier confinement, LBP a répondu positivement à la demande de l'Insee. Le processus de

discussion permettant de déterminer les données mises à disposition (l'échantillon) ainsi que les moyens sécurisés assurant le respect de la vie privée et évitant les risques cybercriminels a duré plus de six mois.

Concrètement, à l'issue de la convention signée entre l'Insee et LBP, cet établissement bancaire a fourni des données anonymisées relatives aux comptes bancaires d'environ 300 000 clients sélectionnés aléatoirement. Seul un échantillon de « clients engagés » – soit des clients connaissant des mouvements bancaires mensuels d'un montant minimum de 150 euros et des opérations de débit – a été concerné. Les données contiennent les soldes des comptes en fin de mois (comptes courants individuels ou joints, livrets, assurances vies et comptes titres), toutes les transactions effectuées (montants et dates des opérations par carte bancaire, chèques, virements, prélèvements, retraits et dépôts) et des données sociodémographiques (âge, sexe, département, statut marital, type d'habitat, catégorie socioprofessionnelle). Ces données constituent une source d'information en temps quasi réel sur la consommation, le revenu et l'épargne des ménages.

Pour mettre à disposition ces données de manière confidentielle, LBP a décidé de passer par le Centre d'accès sécurisé aux données (CASD), perçu comme le moyen le plus rationnel et le plus sécurisé par les services de LBP. Afin d'assurer la non-identification de ses clients, l'établissement bancaire a employé la méthode de « pseudonymisation » consistant à remplacer l'identifiant initial d'une personne par un autre identifiant attribué aléatoirement.

L'Insee a utilisé ces données précieuses à des fins de production statistique. Les données concernées ont permis d'établir :

Deux publications déterminant les effets de la crise sur les ménages [(1)] et une étude interprétant les effets des tensions inflationnistes sur les ménages dans les situations financières les plus fragiles.

Ce partenariat ne s'est pas estompé à l'issue de la crise sanitaire : l'Institut et La Banque Postale prolongent leur collaboration.

D'abord, LBP met régulièrement à jour ses bases de données afin de permettre à l'Insee d'obtenir des informations qui soient toujours corrélées à la conjoncture.

Ensuite, il a été décidé de créer un « indicateur de la fragilité ». La méthodologie est fournie par l'Insee qui, sur ce point, bénéficie d'une expérience largement éprouvée et les données sont fournies par l'établissement bancaire.

Toutefois, trois limites méthodologiques concernant ce partenariat sont à mentionner :

---

(1) Bonnet, Loisel, Olivia, « Impact de la crise sanitaire sur un panel anonymisé de clients de La Banque Postale », 3 novembre 2021.

La première limite tient au vieillissement de l'échantillon fourni. Malgré une actualisation des données disponibles en téléchargement, la base de données n'inclut pas de nouveaux clients par rapport à ceux sélectionnés en 2020.

La seconde limite concerne la représentativité même de la clientèle de LBP. Le profil de clientèle de cet établissement est particulier. En effet, sur 4,1 millions de clients fragiles (1) en France, 40 % d'entre eux sont clients de La Banque Postale, soit près de 1,6 million (2). En outre, certains clients, par exemple les individus non bancarisés, sont par principe absents des jeux de données.

La troisième limite correspond à la complétude des données. Les agrégats constitués ne sont pas directement comparables avec ceux issus des concepts définis par la statistique publique.

Le rapporteur spécial souhaite souligner le caractère vertueux de ce type de partenariat et appelle à ce que d'autres partenariats similaires se développent.

b. Les partenariats avec les acteurs de la téléphonie, prometteurs mais frappés de dissonances

Les partenariats entre l'Insee et des groupes de téléphonie mobile (Orange Business Services France, Bouygues Telecom et SFR) illustrent de manière paroxystique les limites actuelles des partenariats public-privé en matière de données.

Ce partenariat est intervenu juste avant le premier confinement de la crise sanitaire ordonné par le Premier ministre. D'importants mouvements de la population ont eu lieu, conduisant à une nouvelle distribution de la population sur le territoire. Or l'Insee ne disposait d'aucune donnée permettant de mesurer et documenter ce phénomène.

L'Institut a mis en place une collaboration limitée à la période du confinement, sans flux financier et spécifique au suivi de la crise sanitaire.

Les données mises à disposition par les opérateurs téléphoniques ont été des indicateurs anonymisés et agrégés respectant les limites de la directive *ePrivacy* précédemment évoquée. Ces dernières étaient issues des offres commerciales des différents opérateurs.

Les informations fournies ont révélé le potentiel fort des données de téléphonie mobile. Elles permettent d'obtenir des informations précises, en temps réel, sur la présence de population sur un territoire à un instant donné et d'apprécier

(1) Un client est considéré comme « fragile » financièrement suivant le décret n° 2020-889 du 20 juillet 2020 modifiant les conditions d'appréciation par les établissements de crédit de la situation de la fragilité financière de leurs clients titulaires de compte, dès lors qu'il répond aux critères définis par l'article R. 312-4-3 du code monétaire et financier.

(2) Observatoire de l'inclusion bancaire (OiB), rapport d'activité 2021.

la mobilité des personnes. L'Insee a souligné l'intérêt de telles informations, même hors période de crise, afin d'aider à la prise de décision publique.

Toutefois, les données mises à disposition comportaient un certain nombre de biais limitant l'utilisation à des fins statistiques. Les principales limites étaient les suivantes :

La méthodologie de calcul des indicateurs n'a jamais été rendue pleinement transparente par les opérateurs, malgré des demandes répétées de l'Insee. Les explications n'ont été données qu'à la marge. **Les opérateurs ont en effet estimé que la divulgation détaillée de leur méthodologie risquait de révéler leurs propres innovations**.

Les données des opérateurs peuvent révéler leurs parts de marché respectives à des échelles géographiques infra-nationales. Or, les opérateurs craignent que la divulgation des données aux services statistiques entraîne une fuite des données à leurs concurrents.

Le risque de mauvaise presse et de coût en matière d'image de marque peut désinciter les opérateurs à fournir les données. L'acceptabilité sociale de l'utilisation des données de téléphonie n'est en effet pas évidente.

Malgré ce triple frein, l'Insee insiste sur l'intérêt et la nécessité de renforcer le partenariat avec les acteurs de la téléphonie. Ainsi, en 2023, l'Insee collabore avec Orange et l'Université Gustave Eiffel dans le cadre d'un partenariat de recherche jusqu'en 2025.

Toutefois, le partenariat mis en œuvre durant la crise sanitaire n'a pas été renouvelé. Pour l'Insee, l'accès à ces données constitue le principal enjeu des années à venir en termes d'accès à des données privées.

Néanmoins un point d'achoppement majeur est le fait que les opérateurs se refusent à donner leurs données à titre gratuit. Ils considèrent que cela constituerait une perte sèche de revenus. Par exemple, Orange dispose d'une offre commerciale « Flux-Vision », se déclinant en quatre sous-offres (Tourisme, Transport, Cœur de ville et Retail), fournissant des informations quantitatives et fiables sur un secteur.

Les collectivités territoriales sont particulièrement consommatrices de tels services. L'Insee, de son côté, refuse catégoriquement d'avoir recours à l'achat direct de donnée.

c. Le partenariat avec le groupement des cartes bancaires CB : des données atypiques

Lors de la crise sanitaire, l'INSEE a également conclu une convention avec le groupement des cartes bancaires, qui est un groupement d'intérêt économique (GIE-CB), afin de pouvoir accéder aux données relatives aux transactions par carte bancaire en France.

Ces données représentent un intérêt pour la statistique publique puisqu'elles permettent de gommer certains biais identifiés dans l'utilisation des données bancaires. Les données de carte bancaire transmises permettent d'apprécier de manière exhaustive et quasiment en temps réel les pratiques de consommation des ménages, et cela d'autant plus que les paiements par carte deviennent majoritaires. Ces données viennent ainsi compléter les données de caisse des enseignes de grande distribution.

Pour l'Insee cette source figure parmi les plus prometteuses « *car elles tirent parti de la dématérialisation de l'économie tout en retraçant au plus près les achats de biens et services qui constituent directement une partie de la consommation des ménages telle qu'elle sera ensuite mesurée par les comptes nationaux* » (1).

Ce partenariat continue aujourd'hui d'être enrichi. La convention établie entre le GIE-CB et l'Insee prévoit des réunions régulières afin d'échanger sur la qualité et la couverture de ces données.

En vue d'enrichir les travaux entrepris, l'Insee a intégré, fin 2022, la Chaire Finance digitale de l'établissement Télécom Paris à laquelle est partie le groupement des cartes bancaires.

*

Le rapporteur spécial insiste sur la nécessité de surmonter les difficultés qui ont pu parfois être rencontrées à l'occasion du développement de ces nouveaux usages statistiques de données privées, en redéfinissant les partenariats avec un niveau de finesse et de sécurité de la donnée idoine.

2. Vers un *open data* du privé ?

Aujourd'hui, une attention croissante est portée à l'ouverture des données du secteur privé à des fins d'amélioration des services publics et d'orientation dans la prise de décision publique.

Poursuivre l'objectif d'ouverture des données privées à destination du secteur public n'est pas une ambition vaine.

La Commission européenne a élaboré une Stratégie européenne pour les données, présentée le 19 février 2020 et dont l'ambition est de « Bâtir l'avenir numérique de l'Europe ». La Commission plaide pour le partage et l'ouverture des données du secteur privé pour deux raisons principales.

D'abord, les données détenues par les entreprises peuvent aider à orienter les décisions politiques et à améliorer les services publics. La Commission évoque à cet égard plusieurs exemples : Les données privées peuvent permettre d'apporter une réponse plus ciblée aux épidémies ; elles peuvent favoriser une meilleure

(1) Blog de l'Insee, « Nouvelles données pour suivre la conjoncture économique pendant la crise sanitaire : quelles avancées ? quelles suites ? », 28 juillet 2020.

planification urbaine ; les données permettent également d'envisager une meilleure protection de l'environnement.

Aussi, les données privées seraient un outil vertueux et particulièrement bénéfique à l'action publique.

La Commission remarque même que, lors de l'élaboration des statistiques officielles, l'analyse des données détenues par les entreprises privées est souvent plus rentable et peut produire des résultats plus rapides que l'utilisation des données publiques sur certains points, par exemple pour connaître les mouvements de population, les prix, mesurer l'inflation ou l'état de l'économie de l'internet.

Sans évoquer l'hypothèse d'une ouverture totale des données privées qui reviendrait à instituer une véritable *open data* privée, les détenteurs de ces données pourraient soutenir la fourniture de données dans des conditions préférentielles de réutilisation.

Un groupe d'experts de haut niveau s'est réuni sur ce point et a présenté les conclusions de ses travaux [(1)]. Il insiste sur la nécessité de définir un cadre juridique favorable à l'investissement dans le développement et le partage des données de nature privée.

Ce dernier envisage, afin de faciliter les flux de données Business-to-Government (B2G), de mettre en place une fonction reconnue de « gestionnaires de données » dans les organismes privés et publics en vue de faciliter les collaborations. Ensuite, des incitations à destination des entreprises afin que celles-ci développent leurs structures de partage de la donnée à destination des acteurs publics pourraient favorablement être envisagées selon les auteurs. Pour ce faire, le rapport recommande la réalisation d'études permettant de présenter les avantages, pour le secteur privé, de procéder à de tels partages.

Le rapporteur spécial tient à souligner qu'actuellement, les freins à l'intensification des flux entre les acteurs privés et publics sont encore trop nombreux.

Le transfert de données nécessite la mise en place d'infrastructures numériques sécurisées qui peuvent être coûteuses. De surcroît, il peut également exister une crainte d'un effet boomerang consistant à se servir des données privées exploitées pour renforcer l'encadrement d'un secteur d'activité. Enfin, l'absence d'un cadre juridique clair, spécifiquement sur les enjeux de propriété intellectuelle et de droit de la concurrence, constitue un frein puissant aux développements des échanges.

Ensuite, le partage de données entre acteurs du secteur privé pourrait être encouragé, selon la Commission européenne. Sur ce point, les limites peuvent apparaître comme d'autant plus exacerbées.

---

(1) Final report prepared by the High-Level Group on Business-to-Government Data Sharing, « Towards a European Strategy on Business-to-Government data sharing for the public interest », 2020.

En effet, les entreprises ont peu intérêt, à première vue, à échanger leurs données avec une entreprise qui est potentiellement concurrente. Le risque est simple, il peut consister en une perte sèche de revenu pour l'entreprise concernée.

La Commission s'intéresse aux développements d'accords entre les entreprises, mais ceux-ci, pour être rendus possibles, doivent bénéficier d'un renouvellement du cadre du droit de la concurrence. Le risque est en effet celui de la constitution d'oligopoles en situation de position dominante et d'ententes entre les entreprises, celles-ci étant prohibées par les articles 101 et 102 du Traité sur le fonctionnement de l'Union européenne (TFUE).

B. DES ÉVOLUTIONS DU CADRE JURIDIQUE ET DES PRATIQUES ENTOURANT L'ACCÈS AUX DONNÉES PRIVÉES POURRAIENT ÊTRE MISES À PROFIT PAR L'INSEE POUR SES MISSIONS QUOTIDIENNES

1. Fluidifier les échanges de données entre les acteurs publics et les acteurs privés

a. Permettre une utilisation plus extensive des données privées

Le rapporteur spécial a souhaité rendre possible une utilisation plus extensive des données privées à des fins de production statistique.

À cette fin, il apparaît nécessaire de faire évoluer la rédaction du texte régissant l'activité statistique en France.

Proposition n° 3. Modifier la rédaction actuelle de l'article 3 *bis* de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistique.

La nouvelle rédaction de l'article devrait permettre d'avoir un accès élargi aux données privées à des fins de statistiques publiques. Elle pourrait être similaire à celle de l'article 7 *bis* de la même loi relatif aux données administratives.

Néanmoins, une telle mesure, pour être acceptée par les acteurs privés, doit s'accompagner d'un certain nombre de garde-fous. Ceux-ci sont nécessaires pour fluidifier les échanges de données entre le secteur public et les acteurs privés.

b. Prévoir des mesures d'accompagnement et de délimitation de l'usage des données privées transmises

Tout d'abord, pour être acceptée, cette évolution du cadre juridique doit faire l'objet d'une évaluation des besoins en matière de données susceptibles d'être utilisées pour les études statistiques publiques.

Proposition n° 4. Recenser, au cas par cas, avec les acteurs concernés, dans le cadre de partenariats de gré à gré, les données que les acteurs privés pourraient communiquer à l'Insee et plus largement aux acteurs de la statistique publique. Cette évaluation pourrait faire l'objet d'une révision périodique pour s'assurer de la nécessité et de la proportionnalité des informations fournies au secteur public.

Cette concertation permettra de déterminer le volume, la finesse et la méthodologie retenue lors du traitement des données concernées.

Ensuite, il apparaît nécessaire au rapporteur spécial, pour assurer l'acceptabilité d'une telle ouverture des données, d'assortir celle-ci de garanties relatives à leur utilisation.

Proposition n° 5. Proposer des garanties renforcées aux acteurs privés concernés par les échanges de données.

Le rapporteur spécial préconise de prendre pour modèle ce qui a été fait dans le cadre du partenariat entre l'Insee et La Banque Postale.

Le Centre d'accès sécurisé aux données (CASD) pourrait être utilisé comme une plateforme unique et de référence pour tous les échanges entre les services statistiques et les acteurs privés.

Les données individuelles doivent faire l'objet d'une anonymisation et d'une agrégation afin d'empêcher la possibilité d'identification directe de la source.

Enfin, malgré toutes ces précautions, il est probable que les acteurs privés éprouvent des réticences à divulguer leurs données. En effet, certaines d'entre elles exploitent ces données à des fins commerciales. Pourtant, ces données pourraient être nécessaires voire indispensables aux services statistiques à l'avenir.

c. Contraindre à l'ouverture des données.

Le rapporteur spécial formule **deux propositions** visant à obtenir les données en question.

Proposition n° 6. Prévoir un dispositif de réquisition administrative des données privées sous certaines conditions. Le ministre de tutelle concerné pourrait demander au Conseil d'État, statuant en premier et dernier ressort, d'ordonner la communication des données nécessaires, dès lors que cette communication est adaptée et proportionnée à l'objectif poursuivi.

Cette proposition se veut volontairement plus étendue que le cas de réquisition pour « urgence publique » que prévoit le *Data Act*.

En complément, le législateur pourrait prévoir un mécanisme de sanction en cas de refus systématique et injustifié d'accès aux données privées.

Proposition n° 7. Sanctionner les refus d'accès systématiques et injustifiés d'accès aux données privées.

Le rapporteur préconise de modifier l'article 7 de la loi de 1951 en prévoyant des sanctions en cas de non-réponse ; les montants desdites sanctions doivent être suffisamment significatifs pour être dissuasifs.

Une telle mesure pourrait toutefois entrer en contradiction avec le principe de « consentement éclairé » qui guide l'activité statistique. Il apparaît donc nécessaire de limiter les sanctions à des refus témoignant d'un comportement dilatoire.

Le rapporteur spécial souhaite également que le cadre juridique évolue de manière à prendre en compte les évolutions vers une société du numérique.

2. Garantir un cadre protecteur de l'utilisation statistique des données privées : renforcer l'existant

a. Renforcer la sécurité juridique entourant le secret statistique

Le rapporteur spécial estime nécessaire de conforter le cadre juridique applicable à la statistique publique, en rationalisant l'existant et en renforçant les protections entourant les données de nature privée dont pourrait avoir à connaître le secteur public.

En premier lieu, le rapporteur spécial souhaite que la définition de la notion de « secret statistique » soit modifiée afin d'apporter davantage de sécurité juridique tant en termes de protection des informations personnelles qu'en termes d'utilisation des informations par les services statistiques.

La triple définition que connaît la notion dans le texte de 1951 est de nature à porter à confusion et à avoir un effet désincitatif sur l'utilisation des données par les services statistiques.

Proposition n° 8. Uniformiser la notion de secret statistique afin de renforcer la sécurité juridique attachée à cette notion.

Le rapporteur spécial estime que la notion telle qu'elle est définie à l'article 6 de la loi de 1951 est la plus protectrice des droits et libertés. À cet égard, il pense que cette formulation unique pourrait être étendue aux données administratives (article 7 *bis*) et aux données privées (article 3 *bis*) afin de clarifier la rédaction du secret statistique tout en conservant un haut niveau de protection des informations personnelles.

En second lieu, le rapporteur estime que pour renforcer la confiance dans la statistique publique, il serait possible de développer davantage une démarche, lancée depuis 2021 par l'Autorité de la statistique publique (ASP), consistant à « labelliser » des données produites par d'autres personnes que l'Insee et les services statistiques ministériels (SSM) et qui ont vocation à être diffusées.

Proposition n° 9. Développer la démarche de labellisation de séries statistiques.

Le rapporteur estime pertinent d'ouvrir la labellisation d'informations statistiques produites par d'autres personnes que l'Insee et les services statistiques ministériels.

b. Anticiper les besoins en matière de protection des données personnelles dans leur utilisation statistique, par la concertation

Enfin, le rapporteur spécial propose une démarche de concertation permettant de réfléchir aux besoins futurs en matière de protection des données personnelles dans le cadre de l'intensification des flux de partage des données inter-acteurs.

Proposition n° 10. Favoriser des concertations afin d'anticiper les besoins en matière de protection des données personnelles.

Ces dernières pourraient être réalisées entre les services statistiques ministériels, l'Insee, les acteurs de la société civile dans leur diversité, que ce soit des entreprises privées, des associations voire des citoyens.

L'objectif serait d'assurer continuellement un haut niveau d'acceptabilité dans l'utilisation des données personnelles à des fins statistiques.

TRAVAUX DE LA COMMISSION

Lors de sa réunion de 8 heures 30, le 1er juin 2023, la commission des finances, réunie en commission d'évaluation des politiques publiques, a entendu M. Michel Sala, rapporteur spécial des crédits du programme Économie et statistiques.

M. Michel Sala, rapporteur spécial. Avant de vous présenter les résultats de mes travaux d'évaluation, je tiens à remercier les différentes personnes qui ont pris de leur temps dans le cadre des auditions afin d'éclairer ce travail. Je remercie tout spécialement les services de l'Insee et madame Sylvie Lagarde, qui m'a permis d'orienter ma réflexion au sein de cette politique particulièrement large qui est celle de l'*open data* vers l'accès aux données de nature privée. Je remercie également la direction interministérielle du numérique, les syndicats de l'Insee, qui ont su m'apporter des éléments d'information concrets sur la production de statistiques publiques en France, ainsi que les acteurs de la société civile, représentants d'établissements bancaires ou d'associations que j'ai eu l'occasion d'entendre.

La crise sanitaire a sonné le glas pour nos concitoyens et concitoyennes, mais également pour nos services administratifs, d'une activité que l'on pourrait qualifier de normale. La réactivité et l'adaptabilité de nos structures à l'urgence ont largement été mises à l'épreuve. À cet égard, l'Institut national de la statistique et des études économiques, l'Insee, a plus que jamais su assurer son rôle d'aiguillon de l'action publique dans cette période troublée par les incertitudes. Frappé d'une attrition de ces ressources, il s'est tourné vers une donnée de nature nouvelle, la donnée privée.

Pour réaliser ce rapport, j'ai pu m'appuyer sur l'important travail réalisé deux ans plus tôt par notre collègue Éric Bothorel, qui avait produit un rapport, à la demande du Premier ministre, relatif à la politique publique de la donnée. Ce travail conséquent m'a permis d'appréhender la politique de la donnée en France, ses réussites, ses atouts, mais aussi ses faiblesses. Pour ce rapport d'information, le thème retenu, d'abord destiné à s'intéresser à la politique d'*open data* et à la manière dont l'Insee traite et accède aux données, s'est métamorphosé, comme une évidence, vers un domaine plus circonscrit de l'accès aux données privées pour l'Insee à des fins de production statistique. Il me revient dès lors de vous expliquer les raisons de ce choix de thématique qui peut sembler un peu dissonant pour la commission des finances. J'ai été marqué, lors de l'examen du projet de loi de finances et des auditions, de voir à quel point le monde de la donnée, qui pourtant tend aujourd'hui à régir une importante partie de notre vie, m'était inconnu. En tant que responsable de l'évaluation du programme 220 de la mission *Économie* finançant les activités de l'Insee, j'ai décidé de m'intéresser à cette politique publique d'ampleur au travers des travaux de statistiques que produit l'Institut. Si aujourd'hui, tout le monde connaît l'Insee pour la qualité de ses travaux, de ses indicateurs et de ses prévisions macroéconomiques, peu sont ceux à être familiers avec la production statistique et surtout avec la manière dont celle-ci est réalisée.

Lorsque l'on évoque l'Insee, on pense bien sûr à certaines grandes études et grands indicateurs sans en connaître parfois les sous-jacents. Nous nous appuyons sur des indicateurs comme celui de l'indice des prix à la consommation (IPC), mais la réalisation de ce dernier, pour être fiable et objective, doit s'appuyer sur une étendue particulièrement forte de données. Par exemple, près de 500 000 relevés de prix sur internet, la collecte de données des caisses des enseignes de la grande distribution et près de 150 000 relevés de prix dans des points de vente sont réalisés chaque mois pour façonner cet indicateur. Ces chiffres permettent de

prendre la mesure de ce que représente aujourd'hui la masse de données disponibles. Pour pouvoir exploiter ces données à des fins de production statistique, l'Insee s'appuie sur un texte promulgué en 1951, qui continue de faire figure d'autorité en définissant la notion de secret statistique. Cela peut étonner, à l'heure où les sociétés s'accélèrent et deviennent hyper modernes, de voir que l'encadrement juridique semble en inadéquation avec son époque. La notion de secret statistique n'en demeure pas moins la pierre angulaire de l'activité statistique. Or trois remarques peuvent être émises à son égard. Premièrement, la définition du secret statistique n'est pas uniforme dans la lettre du texte de loi. Elle connaît en effet une triple définition, qui dépend de la nature de la donnée exploitée par les services statistiques. Ce faisant, elle est lourde et peu lisible pour les services statistiques et l'Insee, ce qui peut brider son activité statistique. Deuxièmement, la définition de ce secret statistique est particulièrement restrictive. Elle limite le recours aux données de nature privée à des cas précisément identifiés, ce qui ne permet pas aujourd'hui de tirer l'avantage de la variété de celles-ci. Troisièmement, comme le relevait l'association Ouvre-boîte que j'ai pu auditionner, le secret statistique serait également employé comme un paravent administratif, une vague de procédures qui me pousse à formuler des critiques, comme d'autres avant moi, sur cette culture du secret persistante au sein des administrations.

En sus de cette activité de production d'études et d'indicateurs, l'Insee est désormais sommé d'ouvrir ses bases de données. La loi pour une République numérique d'octobre 2016 a transposé en droit français la logique d'*open data*, qui a eu pour effet une ouverture large et gratuite des données produites par l'Institut. Sur un plan budgétaire, la mise en œuvre de cette politique s'est matérialisée par la perte de la redevance perçue précédemment par la revente des données, soit une perte de onze millions d'euros annuels pour l'Institut. Sur un plan pratique désormais, l'*open data* induit la mise à disposition d'une information gratuite, facilement réutilisable. Or un tel exercice est parfois périlleux puisqu'il pourrait conduire, en cas de divulgation d'informations trop fines et précises, à connaître l'origine de la donnée.

La variété des données de nature privée est presque infinie. On peut donc aisément imaginer la quantité d'informations utiles pour l'action publique que celles-ci peuvent contenir. Près de 100 000 milliards de giga-octets de données ont été collectés en 2022. Google, par exemple, traite près de 3,5 milliards de requêtes d'utilisation chaque jour. Outre la richesse pour les services statistiques que peut représenter cette masse, j'attire l'attention sur les risques inhérents à la protection de la vie privée.

Enfin, j'ai fait le choix de m'intéresser à cette thématique pour sa dimension prospective et pour l'actualité forte qui entoure cette question. Tout d'abord, les partenariats tissés dans le cadre de la crise sanitaire avec des acteurs privés ont révélé une plus-value forte à l'usage de cette matière inexplorée. Le partenariat établi entre l'Insee et La Banque Postale, mis en place durant la crise sanitaire afin de suivre les effets économiques du confinement sur la population, s'est révélé prometteur. Cette convention a été déterminée d'un commun accord entre les deux acteurs et a reposé sur un échantillon déterminé de données composé d'environ 300 000 clients sélectionnés aléatoirement. Je tiens ici à souligner le caractère vertueux de ce partenariat, qui a permis aux services de l'Insee de réaliser des études novatrices. Mon regard est plus nuancé en revanche sur le partenariat établi entre l'Insee et le service de téléphonie mobile, notamment Orange. Ce dernier a consisté, dans le temps circonscrit de la crise sanitaire, à fournir aux services de l'Insee des données de localisation mobile, pour être en mesure de suivre les mouvements physiques de la population sur le territoire. On peut toutefois déplorer le manque de transparence dans la méthodologie employée par les opérateurs pour mettre à disposition des données agrégées et la réticence de ceux-ci à partager des informations, a fortiori lorsque les opérateurs disposent d'une offre commerciale.

Ensuite, l'actualité entourant cette question de l'utilisation des données privées a particulièrement attiré mon attention. Au niveau européen, plusieurs propositions des textes ont été présentées et portent spécifiquement sur cette thématique de la donnée, avec une incidence plus ou moins importante sur la production statistique. Dans le cadre de la stratégie européenne pour les données présentées le 19 février 2020, les textes *Data Gouvernance Act* et le *Data Act* ont constitué un point d'attention pour les services de l'Insee. Pour autant, ces deux textes n'apportent que des modifications marginales à l'exploitation des données privées, en prévoyant par exemple une utilisation des données privées en cas d'urgence publique pour le *Data Act*. Le règlement numéro 223, dit aussi loi statistique européenne de 2009, fait l'objet d'une proposition de révision, mais qui n'est encore qu'au stade de la discussion. J'attire l'attention sur la nécessité de demeurer attentifs à cette révision réglementaire, qui pourrait avoir une incidence notable sur les services statistiques. Le rapport Jean Pisani-Ferry paru la semaine dernière a consacré un rapport thématique à la donnée et aux indicateurs à des fins environnementales. Il relevait notamment que des ressources de nature privée, comme des relevés de compteurs de gaz ou d'électricité, pouvaient être utilement exploitées à des fins de production statistique pour orienter le décideur public.

Fort de ces constats, j'ai formulé un certain nombre de propositions qui visent à intensifier les flux de données entre les acteurs privés et les acteurs publics, mais également à rendre plus lisible un cadre normatif que je trouve ancien et peu propice à ce nouveau monde de la donnée. Je vous remercie de votre attention.

M. Roland Lescure, ministre délégué chargé de l'industrie. Merci monsieur le rapporteur et bravo pour votre intérêt pour cette belle institution qui fait à la fois la fierté des Françaises et des Français, mais aussi la fierté de la France à l'international. Votre rapport spécifique permet de mettre la lumière sur cette institution, dont j'ai été membre quelques années et qui a été particulièrement visible par nos concitoyens dans le cadre de la crise sanitaire. Nous avons pu découvrir que nous étions capables, avec de la créativité, de mesurer en France mieux qu'ailleurs l'activité économique quasiment au quotidien, grâce aux données de cartes bancaires auxquelles l'Insee a eu accès et, vous l'avez dit aussi, aux données de téléphonie mobile. Merci aux employés de l'Insee qui ont accompli un travail particulièrement créatif et utile dans le cadre de la crise sanitaire, mais qui, au jour le jour et depuis des années, nous aident à mieux comprendre le monde dans lequel nous vivons.

L'Insee, comme tous les instituts de statistiques, fait face à un défi majeur : la numérisation croissante de l'économie et de la société. C'est à la fois une opportunité pour l'Institut et un défi. Cette opportunité est de compléter les exploitations habituelles qui sont aujourd'hui fondées sur des enquêtes statistiques, téléphoniques, internet, etc., des fichiers administratifs détenus par les administrations et, depuis la loi pour une République numérique de 2016, la communication par des personnes morales de droit privé de données. Nous avons ainsi ajouté un article 3 *bis* dans la loi de 1951 à laquelle vous avez fait référence. Nous avons également accès à des données privées dans le cadre de conventions de gré à gré, ce qui permet à l'Insee d'avoir accès à des transactions bancaires, mais aussi à des données ouvertes, notamment les images satellitaires. Toutes ces exploitations sont complétées par l'Insee, par des méthodes développées ces dernières années, en matière d'intelligence artificielle, de *data science*, de *nowcasting* qui permet d'analyser la situation plutôt que de la prévoir, des extractions d'informations issues duweb, etc. Ces nouvelles méthodes permettent, en complément des anciennes, de produire des indicateurs à des niveaux géographiques extrêmement fins.

Le défi est clair : de nouveaux organismes et des opérateurs privés se mettent à produire des informations qui se veulent statistiques qui, pour certaines, sont d'une rigueur et

d'une qualité exemplaires et qui, pour d'autres, laissent à désirer. Tout cela est disponible en temps réel, difficile à évaluer et à utiliser. Les chiffres peuvent être repris par des médias et par le grand public et sont parfois relayés de manière amplifiée et détournée sur les réseaux sociaux, là où l'Insee, par sa rigueur, a parfois besoin d'un peu de temps pour produire de la donnée. Nous devons conserver, vous l'avez dit, la robustesse de l'encadrement de la statistique publique, issue de la loi de statistiques de 1951. Nous pouvons pour autant élargir l'accès par l'Insee et les services statistiques à des données privées. Au-delà des conventions dont j'ai parlé, nous pourrions envisager un partenariat plus structurel avec les opérateurs de téléphonie. Ce partenariat avait été très utile pendant la Covid. Il a été interrompu. Je sais que des discussions sont en cours entre l'Insee et les opérateurs téléphoniques. Il se pose un sujet de prix, mais aussi de qualité de la donnée. L'Insee a besoin, pour réaliser un travail exemplaire, de la donnée brute. Or les opérateurs proposent aujourd'hui des données d'ores et déjà traitées.

Nous pourrions envisager de faire évoluer la loi statistique de 1951 pour élargir l'accès aux données privées, uniquement afin d'avoir accès à l'élaboration de statistiques. Aujourd'hui, ce qui est permis est le remplacement d'enquêtes statistiques d'ores et déjà obligatoires. Nous devons pouvoir identifier de nouveaux champs dans lesquels l'Insee pourrait avoir accès à des données privées, à condition évidemment d'assurer quelques garanties et d'expliquer pourquoi ces données sont nécessaires et d'avoir une finalité bien spécifique. Ces données n'ont pas vocation à nourrir les autorités de régulation ou la recherche, mais uniquement l'appareil statistique. Elles doivent être, de ce fait, inscrites dans un programme statistique et proportionnées au strict besoin. Afin de maintenir la confiance du public, il s'agit également de poursuivre les concertations sous l'égide du conseil national de l'information statistique, un organisme important qui permet d'assurer la concertation entre les utilisateurs et les producteurs de statistiques et de communiquer en toute transparence sur les données utilisées pour la production statistique.

Merci pour ce rapport. Nous sommes évidemment prêts à travailler sur les suites avec vous et avec d'autres députés. La statistique n'appartient à personne. Elle appartient à tout le monde. C'est l'estimation de la vérité qui mobilise et motive l'ensemble des agents de l'Insee et si nous pouvons améliorer leur travail, nous devrions le faire.

M. le président Éric Coquerel. Merci, monsieur le ministre. Je remercie Michel Sala pour ce rapport qui, d'une part, nous rappelle les qualités de l'Insee et de ses agents. D'autre part, ce rapport a révélé une question que je trouve effectivement importante dans les temps qui courent. Je passe la parole aux orateurs de groupe.

M. Dominique Da Silva (RE). Monsieur le rapporteur, merci pour ce rapport d'information très technique et très riche. Merci aussi d'avoir salué le travail de notre collègue Éric Bothorel, qui avait conduit une mission sur la politique publique de la donnée.

Nous savons que l'Insee est l'acteur de référence dans la production statistique en France, mais qu'il est concurrencé par des acteurs privés disposant de toujours plus de données, au détriment parfois de la qualité de la statistique publique. Aujourd'hui, l'*open data* est entrée dans le quotidien des entreprises. Grandes consommatrices de ces données publiques ouvertes, les entreprises privées ont en revanche longtemps été récalcitrantes à l'idée de partager leur propre *data*, de peur de révéler des informations sensibles à leurs activités. Je note d'ailleurs que le cadre applicable à la statistique publique est affecté de lourdeurs, ce qui conduit à brider l'utilisation de certaines *data*. Comment le gouvernement entend-il aider les entreprises et l'Insee à jouer pleinement leur rôle pour démocratiser la donnée au service de notre compétitivité économique ? Je vous remercie.

M. Pascal Lecamp (Dem). En préambule, je souhaiterais exprimer toute ma reconnaissance à l'Insee et à ses agents qui, inlassablement, nous tiennent informés, chiffres à l'appui. Dans un monde où la notion de vérité est remise en cause, où les faits objectifs voient leurs contours de plus en plus floutés, le travail de statistique que réalise l'Institut est tout à fait essentiel. Nous en bénéficions tous chaque jour dans nos missions professionnelles ou dans nos collectivités territoriales.

L'Insee se heurte parfois à des refus dans l'accès aux données. D'où proviennent ces résistances, quand elles ne sont évidemment pas justifiées par des freins juridiques liés à la protection des données personnelles ? Quel nouveau partenariat à l'Insee pourrait-il envisager, privé ou public, pour collecter des données, en particulier si certaines des recommandations du rapporteur devaient être mises en œuvre ?

M. Mickaël Bouloux (SOC). Merci pour votre travail, monsieur le rapporteur, qui permet de mettre la lumière sur un sujet important et méconnu. J'aurais deux questions concernant l'accès et le traitement des données recueillies par l'Insee dans le cadre de la politique de l'*open data*. D'une part, si l'Insee est un outil formidable de production et d'analyse des statistiques, vous soulignez qu'il est aussi régi par une réglementation obsolète qui date de 1951. Vous prônez l'évolution de l'article 3 *bis* de cette loi de 1951, mais cet article a déjà été modifié le 7 octobre 2016 pour la loi pour une République numérique. Il permet notamment au ministre de l'économie, pour les refus des personnes morales qui ne permettraient pas d'accéder à des statistiques, de les mettre en demeure, puisqu'elles sont passibles d'une amende administrative. Quelle forme l'élargissement pour les données privées prendrait-il ?

D'autre part, la réglementation de l'Insee agit comme un carcan, notamment par rapport aux *Big Tech*, qui produisent des données de façon exponentielle. Quels outils pourrait-on envisager pour s'assurer que les données recueillies par les acteurs privés soient mieux encadrées ? Comment définir ensemble cette notion d'intérêt général de ces données privées ?

M. Roland Lescure, ministre délégué. Monsieur le député Da Silva, l'Insee fait déjà beaucoup pour démocratiser la statistique. Je vous engage à suivre les blogs de l'Insee qui sont extrêmement didactiques. Nous pouvons faire mieux et plus, notamment en collaborant encore davantage avec l'éducation nationale, en allant dans les écoles, en conduisant nos jeunes à développer un regard critique sur les informations diffusées largement sur les réseaux sociaux. Cela répondra en partie aux remarques du député Lecamp sur le fait de développer la démarche institutionnelle d'identification, voire de labellisation des statistiques à visée d'informations générales. Par définition, l'Insee, quand il résiste à des demandes d'information, le fait parce que ses agents considèrent que la loi ne leur permet pas de transmettre les données en question. Peut-être sont-elles insuffisamment sécurisées d'un point de vue juridique et nous devons renforcer ces aspects, pour nous assurer que la transparence soit absolue, compte tenu du droit absolu de préserver l'anonymat et la sécurité des données et des personnes.

Une démarche a été lancée par l'autorité de la statistique publique, qui vise à identifier des organismes qui respectent plusieurs critères de qualité pour élargir la labellisation des statistiques à visée d'information générale. La communication sur cette démarche sera nécessaire pour en accroître la visibilité et la démocratisation. Nous devons également, dans le cadre de l'élargissement potentiel de la loi de 1951, bien établir les besoins et apporter des garanties aux acteurs privés qui fonderont nos données. Nous pourrions aussi envisager d'uniformiser la notion de secret statistique. La loi de 1951 définit les secrets statistiques en fonction des modes de collecte, selon une sédimentation, je le reconnais, quelque peu historique, alors que le règlement européen 223 de 2009 propose une définition unique qui

pourrait renforcer la sécurité juridique de l'Insee et faciliter la vie de ces agents dans les réponses aux questions qui leur sont régulièrement posées.

M. Michel Sala, rapporteur spécial. Pour l'instant, la loi pour une République numérique ne permet l'accès à des données des personnes morales de droit privé que lorsqu'elles sont délégataires d'un service public, mais pas à l'ensemble des données.

Sur le sujet des partenariats, avec le Centre d'accès sécurisé aux données (CASD), comment l'Insee protège-t-il et s'ouvre-t-il à la donnée privée ? Le CASD doit avoir des moyens pour développer ce qu'il a réussi à faire, notamment avec la Banque Postale. Des études sont conduites au niveau européen pour reproduire l'action menée par le CASD en France.

Enfin, pour conclure, notre action sur le climat a et aura des incidences économiques. Pour les appréhender au mieux, la statistique publique a un rôle à jouer dans l'analyse et le suivi des enjeux de la transition climatique. Nous devons disposer de statistiques adaptées et précises pour éclairer l'action et le débat public dans toutes ses dimensions. Nous devrons estimer les émissions de gaz à effet de serre, au regard de l'activité économique, pour aider à définir et suivre les politiques publiques touchant les entreprises ou les ménages, pour permettre l'évaluation des investissements liés à la transition, répondre à la question du financeur, mais aussi évaluer l'écart entre l'ambition et les moyens mis en œuvre, ainsi que suivre l'observation des dommages et estimer l'adaptation nécessaire. Il s'agit d'un chantier indispensable et urgent qu'ont commencé à aborder l'Insee et France Stratégie, notamment dans un rapport paru en mai 2023 intitulé « Incidence économique de l'action pour le climat ».

La commission autorise, en application de l'article 146, alinéa 3, du Règlement de l'Assemblée nationale, la publication du rapport d'information de M. Michel Sala, rapporteur spécial.

LISTE DES PERSONNES AUDITIONNÉES

Institut national de la statistique et des études économiques (Insee) :

– Mme Sylvie Lagarde, directrice générale de la méthodologie de la coordination statistique et internationale.

Représentants syndicaux de l'Insee :

– M. Yohan Baillieul, co-secrétaire national de la CGT Insee-Genes ;

– Mme Zouza Hontangs, co-secrétaire nationale de la CGT Insee-Genes ;

– M. Christian Monteil, membre du bureau national Sud-Insee ;

– M. Axel Gilbert, co-secrétaire national de Sud-Insee.

Direction interministérielle du numérique :

– Mme Stéphanie Schaer, directrice interministérielle du numérique ;

– M. Pierre Sucevic, chef du pôle juridique de la Dinum.

La Banque Postale :

– M. Philippe Aurain, directeur des études économiques de La Banque Postale.

Association Ouvre-boîte :

– M. Michel Blancard.

*
* *