Confidentialité des données statistiques : un enjeu majeur pour le service statistique public

*Patrick Redor**

Le service statistique public (SSP) est chargé de produire et de diffuser de l'information statistique à partir de données issues de fichiers administratifs ou d'enquêtes. Le SSP est ainsi dépositaire d'un large éventail de données confidentielles sur des individus, des ménages, des entreprises ou des organisations. Pour répondre à ses obligations légales et éthiques, le SSP doit garantir la confidentialité des données collectées ou produites à des fins statistiques, en appliquant le secret statistique et en respectant les obligations de protection des données personnelles formulées par la loi Informatique et libertés et le règlement général sur la protection des données (RGPD). Le SSP est dispensé de répondre aux demandes de réquisitions, et en cas de non-respect du secret statistique, des sanctions pénales sévères sont appliquées. Les obligations relatives au secret statistique découlent de la loi de 1951 sur l'obligation, la coordination et le secret en matière de statistiques et des règlements européens, tels que le règlement général sur la protection des données (RGPD) et le règlement n°223 sur les statistiques européennes.

The Official Statistical Service (SSP)[1] is responsible for producing and disseminating statistical information based on administrative data or surveys. The SSP therefore holds a wide range of confidential data on individuals, households, businesses and organisations. To fulfil its legal and ethical obligations, the SSP must ensure the confidentiality of data collected or produced for statistical purposes, by applying statistical confidentiality and observing the personal data protection obligations set out in the Data Protection Act and the General Data Protection Regulation (GDPR). The SSP is not required to answer requisitions, and in case of non-compliance with statistical confidentiality, severe criminal sanctions are applied. Statistical confidentiality obligations stem from the 1951 Act on Legal Obligation, Coordination and Confidentiality in the Field of Statistics and from European regulations, such as the General Data Protection Regulation (GDPR) and Regulation No 223 on European Statistics.

* Chef de l'Unité Affaires juridiques et contentieuses, Insee, *patrick.redor@insee.fr*

1 The Official Statistical Service (SSP) is composed of INSEE and 16 Ministerial Statistical Offices (MSOs) who carry out statistical operations in their field of competence.

Le service statistique public (SSP) est composé de l'Insee et des services statistiques ministériels (SSM). Il a pour mission de produire et diffuser de l'information statistique et la « data » ou donnée numérique est au cœur de ses métiers[2].

Les statistiques produites par le SSP mesurent des faits économiques et sociaux. Les données mobilisées portent ainsi sur les comportements et les situations de personnes (individus, ménages), d'entreprises ou d'organisations. Ces données sont le plus souvent confidentielles.

Les missions du SSP ne dépendent pas seulement de sa capacité à maîtriser les outils ou les méthodes nécessaires à la production d'une information de qualité, mais aussi de sa capacité à protéger et à garantir la confidentialité des données qui lui sont confiées. Cette protection est la condition pour continuer à disposer de ces données.

La confidentialité des données, un enjeu crucial de maîtrise des risques

> ***La maîtrise des risques de perte ou de violation de confidentialité des données dont il est dépositaire représente un enjeu crucial pour le SSP.***

La maîtrise des risques de perte ou de violation de confidentialité des données dont il est dépositaire représente un enjeu crucial pour le SSP.

L'ampleur de ces risques se mesure globalement au nombre de personnes – personnes physiques ou morales[3] – concernées par les données dont le SSP dispose à des fins statistiques, ainsi qu'au volume et à la sensibilité de ces données.

Des dispositifs légaux, au premier chef la loi de 1951[4], confèrent au SSP la possibilité de réaliser des enquêtes ou d'accéder aux fichiers détenus par l'administration. Sous certaines conditions, le SSP peut également se voir communiquer les bases de données de certains organismes de droit privé[5]. L'activité du SSP se nourrit de la collecte d'un volume important de données d'origines diverses. Plus d'une centaine d'enquêtes statistiques[6] sont réalisées chaque année, mais c'est l'accès aux fichiers des administrations qui constitue, en volume, la principale source de production du SSP. De façon générale, pour différentes raisons (coût, charge pour les enquêtés, etc.), le SSP s'est engagé depuis de nombreuses années dans la valorisation des fichiers administratifs. Plus récemment, il s'est tourné vers les bases de données privées, sous la contrainte néanmoins que la loi, sauf exception, ne permet pas d'imposer cette cession aux organismes qui les produisent et les détiennent.

2 Le SSP, outre ses missions statistiques, peut aussi être chargé de la gestion de certains répertoires ou fichiers dont les finalités sont administratives et dont les données ne relèvent pas du secret statistique. À titre d'exemple, on peut citer le répertoire national d'identification des personnes physiques (RNIPP) pour l'Insee ou le Fichier national des établissements sanitaires et sociaux (Finess) pour la Drees, service statistique des ministères sanitaires et sociaux.

3 En droit français, une personne physique est un être humain doté, en tant que tel, de la personnalité juridique. Une personne morale est un groupement doté de la personnalité juridique. Généralement une personne morale se compose d'un groupe de personnes physiques réunies pour atteindre un objectif commun.

4 Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques (voir fondements juridiques).

5 Article 3 bis de la loi de 1951 (voir fondements juridiques).

6 Voir les arrêtés de programme d'enquêtes publiés sur le site du Cnis : *https://www.cnis.fr/arretes-au-journal-officiel-du-programme-denquetes-2022/.*

Ces enquêtes et cette collecte de fichiers, cumulées au fil du temps, aboutissent à une situation où le SSP est dépositaire d'un très large éventail de données, de sensibilité parfois très forte (handicap et santé, pratiques religieuses, etc.), sur une large population d'individus, ménages, entreprises ou organisations.

> ***À la différence des sources administratives, les enquêtes ne concernent qu'une petite fraction de la population.***

L'Insee dispose, grâce aux fichiers que lui communiquent l'administration fiscale ou les organismes de protection sociale, de la déclaration de revenus de chaque Français, de ses prestations sociales, de ses périodes d'emploi, de chômage ou d'inactivité, de ses salaires et rémunérations. Les données d'origine fiscale ou douanière permettent également de connaître, pour chaque entreprise, ses comptes, ses achats, ventes, importations et exportations, la nature de ses actifs, les prix qu'elle pratique, ses effectifs, ses activités.

À la différence des sources administratives, en principe exhaustives sur leur champ, les enquêtes ne concernent qu'une petite fraction de la population. Elles peuvent néanmoins collecter des informations sensibles[7] sur les revenus, le patrimoine, la situation sur le marché du travail (enquête emploi, enquête SRCV[8] sur les ressources des ménages, enquête histoire de vie et patrimoine, enquête sur le vécu du travail et du chômage durant la crise sanitaire), voire très sensibles sur les croyances religieuses (enquête Trajectoires et origines), la santé (enquête vie quotidienne et santé, enquête « Épidémiologie et Conditions de vie »), les violences physiques et sexuelles (enquête cadre de vie et sécurité, enquête nationale de climat scolaire et de victimation auprès des collégiens). Elles peuvent aussi cibler des populations fragiles ou vulnérables (enquête auprès des sans-domicile).

Le recensement constitue parmi les enquêtes un cas à part. Son questionnaire est relativement court par rapport à d'autres enquêtes spécialisées, mais il interroge une fraction élevée de la population[9] et couvre les principales caractéristiques des individus, de leur ménage et de leur logement.

▶ Les traitements statistiques imposent très souvent la collecte et la conservation de données d'identification —

Le SSP dispose de nombreuses informations sur les caractéristiques des personnes ou des entreprises, et très souvent ces informations sont accompagnées de données qui permettent d'identifier précisément ceux qu'elles concernent. Ces données particulières sont qualifiées de **« données d'identification »**.

7 Les données « sensibles » qualifient au premier chef les « catégories particulières » de données définies par l'article 9 du règlement général sur la protection des données (informations concernant les opinions politiques ou la santé, par exemple) ainsi que les données relatives aux condamnations pénales ou aux infractions. Au-delà des dispositions du RGPD, peuvent être qualifiées de « sensibles » des données hautement personnelles ou dont la violation serait susceptible d'entraîner des incidences graves pour les personnes concernées (données sur les revenus par exemple) (voir les Lignes directrices du G29 sur l'analyse d'impact relative à la protection des données : *https://www.cnil.fr/sites/default/files/atoms/files/wp248_rev.01_fr.pdf*).

8 Statistiques sur les ressources et conditions de vie.

9 Sur un cycle de cinq ans, la couverture est de 40% pour les grandes communes et de 100% pour les petites communes.

On peut disposer directement de l'identité de la personne : son état-civil (nom et prénoms) s'il s'agit d'une personne physique, sa raison sociale s'il s'agit d'une personne morale, entreprise ou autre organisation. On parle alors d'identification **directe**.

On peut aussi reconnaître la personne à travers un « pseudonyme » : numéro de téléphone, adresse mail, numéro fiscal, numéro client sur une facture, Nir[10], Siren[11], code statistique non signifiant ***(encadré 1)***. On parle alors d'identification **indirecte** : l'identité n'est pas directement révélée, mais il est possible de la retrouver de manière univoque à partir du pseudonyme.

On peut fort justement se demander pourquoi ces données d'identification, directes ou indirectes, sont présentes dans des sources statistiques, alors qu'elles ne sont pas l'objet, par elles-mêmes, de statistiques.

En fait, l'utilisation de données d'identification à des fins statistiques est indispensable pour répondre à deux types essentiels de besoins :

- Pour les différentes phases des traitements d'enquête : tirage des échantillons, localisation et contact avec les personnes à enquêter ;
- Pour les appariements, c'est-à-dire pour connecter les données de fichiers d'origines différentes. Ces appariements permettent, sur la base de données déjà disponibles et sans avoir à recourir à de nouvelles enquêtes toujours plus coûteuses, d'enrichir ou d'améliorer les traitements statistiques. Par exemple, l'appariement des revenus déclarés dans les fichiers fiscaux et des prestations versées par les organismes sociaux pour la production d'indicateurs sur les niveaux de vie et de taux de pauvreté localisés ; ou encore, l'appariement des données de Pôle emploi et celles de l'enquête Emploi à des fins de mesure de la qualité de ces sources.

▶ Encadré 1. Le cas particulier du Nir*

Parmi les « pseudonymes », le Nir, présent par exemple dans certains traitements statistiques comme l'échantillon démographique permanent ou les panels d'actifs, présente un caractère particulier. Il est l'identifiant des personnes dans le répertoire national d'identification des personnes physiques **(RNIPP)**, qui contient l'ensemble de la population française (hors collectivités d'outre-mer). Pour cette raison, il est utilisé comme identifiant unique de référence notamment dans la sphère médico-sociale et auprès des organismes de sécurité sociale.

Le pouvoir d'identification associé au Nir rend son usage particulièrement sensible. En outre, le Nir est en partie signifiant : il incorpore des informations sur le sexe, le lieu et la date de naissance, ce qui le rend partiellement reconstructible à partir d'informations connues sur une personne. La sensibilité particulière du Nir explique que l'Insee et le SSP s'orientent de plus en plus vers l'utilisation du code statistique non signifiant (CSNS)**, qui substitue au Nir un identifiant de nature aléatoire et dont l'usage est exclusivement réservé au SSP.

** Numéro d'identification au répertoire ou « numéro de sécurité sociale ».*

*** Voir l'article de Yves-Laurent Bénichou, Séverine Gilles et Lionel Espinasse sur le Code statistique non signifiant (CSNS) dans ce même numéro.*

10 Numéro d'identification au répertoire national des personnes physiques (RNIPP), plus connu comme le « numéro de sécurité sociale ».

11 Le numéro Siren (pour « système d'identification du répertoire des entreprises ») est le numéro unique d'identification de chaque entreprise.

La suppression de données d'identification, qu'elles soient directes ou indirectes ne suffit pas cependant à garantir **l'anonymat** des entités, personnes, entreprises ou organismes concernés.

Même sans nom ou pseudonyme, les autres informations relatives à une personne ou une organisation – âge, sexe, revenus, commune de résidence, activité principale, chiffre d'affaires, etc. – si elles sont combinées, peuvent suffire à réidentifier l'individu exactement.

Il suffit de relativement peu de données : selon une étude parue dans le magazine *Nature Communications (Rocher, Hendrickx et de Montjoye, 2019)*, 15 variables ou caractéristiques suffisent pour réidentifier une personne.

Le degré d'exposition d'une base de données à un risque de pertes de confidentialité dépend des moyens qui doivent être mis en œuvre pour réidentifier une personne ou une organisation. Ces moyens vont du plus simple, si la base contient des données d'identification directe, au plus compliqué, si elle ne contient aucune donnée d'identification directe ou indirecte et s'il faut combiner les informations contenues dans la base.

La suppression de tout risque de rupture de confidentialité pour une base de données est possible moyennant son **anonymisation**. L'anonymisation implique non seulement la suppression de toute donnée d'identification, mais également le traitement des données pour que toute réidentification par combinaison soit impossible. Dans la plupart des cas, l'anonymisation des bases de données nuirait gravement aux missions du SSP, car il a besoin de conserver des données d'identification, ou *a minima* de conserver une information détaillée, sous une forme qui facilite la réidentification.

Des mesures de protection indispensables, à la mesure de la gravité de l'impact pour les personnes concernées

La nature et le volume des données dont dispose le SSP, couplés au fait que ces données sont généralement identifiantes font peser sur lui une forte responsabilité. En cas de rupture de confidentialité, l'impact pour les personnes ou organisations concernées peut varier en fonction de la nature des données ; il peut se limiter à de simples désagréments mais peut aussi avoir de très graves conséquences.

> ***En cas de rupture de confidentialité, l'impact pour les personnes ou organisations concernées peut […] se limiter à de simples désagréments mais peut aussi avoir de très graves conséquences.***

Si les données divulguées sont relatives à la vie privée, l'impact peut être mineur et ne causer que peu ou pas de préjudice tangible à la personne concernée. Cette divulgation peut néanmoins, de manière subjective, affecter la personne, qui considère que des informations ou des données personnelles ont été divulguées sans son consentement ou sans qu'elle ait été informée de manière adéquate.

Dans d'autres cas, une violation de la vie privée peut avoir des conséquences graves pour un individu, sur le plan personnel ou professionnel. Par exemple, la divulgation de données sensibles comme des informations sur ses revenus ou sa santé peut entraîner pour une personne des préjudices financiers ou juridiques importants. Néanmoins, la simple divulgation d'une adresse, si elle permet de retrouver une personne, peut avoir parfois des conséquences dramatiques si celle-ci est l'objet de menaces.

Si les données concernent des activités entrepreneuriales, les conséquences sont susceptibles de se traduire en termes de pertes d'avantages concurrentiels, de préjudices d'image auprès du public et des consommateurs.

> ***Il est important de prendre en compte la violation de confidentialité afin de mettre en place des mesures de protection efficaces.***

Quelle que soit sa gravité, une atteinte à la vie privée peut affecter la confiance des individus dans les institutions. Il est donc important de prendre en compte la violation de confidentialité afin de mettre en place des mesures de protection efficaces pour minimiser les risques pour les individus et organisations concernés.

Ces mesures peuvent inclure des dispositifs techniques (protection et surveillance des accès et des réseaux, mots de passe, antivirus, etc.) et organisationnels (règles et procédures de désignations des agents habilités à accéder aux données confidentielles, définition d'une politique de sécurité, formation et sensibilisation des agents aux enjeux de la confidentialité, etc.). Par exemple, les agents de l'Insee ou des SSM n'ont pas accès à l'ensemble des données confidentielles dont dispose l'institut, mais uniquement à celles nécessaires au titre de leurs fonctions. Ce principe de compartimentation est l'une des règles essentielles pour limiter les risques de divulgation de données protégées.

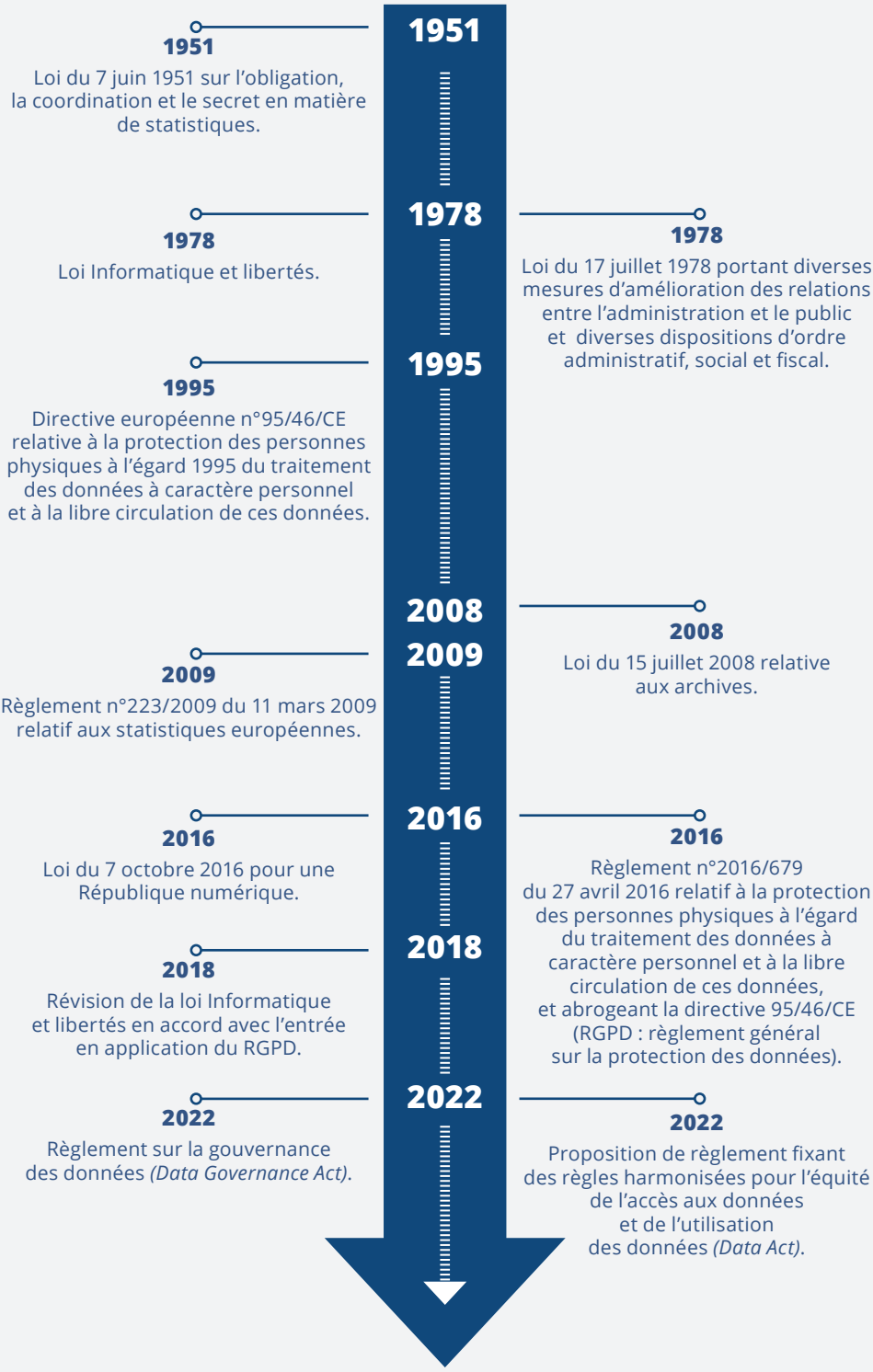
Toutes ces mesures, pour développées et sophistiquées qu'elles soient, sont insuffisantes sans la maîtrise du facteur humain, qui passe par les moyens que le SSP consacre à la formation et à la sensibilisation de ses agents aux enjeux de la confidentialité.

Des obligations strictement définies et encadrées par la loi, pour les données individuelles comme pour les résultats statistiques

La définition et la mise en œuvre de mesures de sécurité répondent à un engagement éthique et déontologique, ainsi qu'à des obligations posées par la loi. Chacun des agents du SSP est soumis en particulier à deux lois : celle de 1951, ou loi sur l'obligation, la coordination et le secret en matière de statistiques[12], qui définit le secret statistique, et celle de 1978 ou loi Informatique et libertés[13], qui définit la protection des données personnelles ***(frise chronologique)***. Le secret statistique s'applique aux données confidentielles obtenues ou exploitées à des fins de production de résultats statistiques. Pour les agents publics qui y sont soumis, c'est un secret professionnel.

12 Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.
13 Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (voir fondements juridiques).

Frise chronologique



Le secret statistique a un double rôle juridique. D'une part, à l'égard des agents du SSP, le non-respect de la confidentialité les expose à des sanctions pénales sévères, allant jusqu'à un an d'emprisonnement et 15 000 euros d'amende. D'autre part, le secret statistique agit comme une **protection**. Il dispense le SSP de répondre aux demandes de réquisitions émises par des autorités administratives ou judiciaires concernant des données statistiques. Le secret statistique est opposable à ces réquisitions, ce qui n'est pas le cas du secret professionnel en général. Cependant, le secret statistique autorise la communication de données statistiques confidentielles auprès de chercheurs ou d'autres administrations, **en réponse à des demandes d'accès formulées pour des motifs statistiques ou de recherche scientifique ou historique** et sous la condition que les demandeurs s'engagent eux-mêmes à respecter la confidentialité des données, sous peine des sanctions prévues par la loi.

En parallèle du droit français avec la loi de 1951, le règlement n°223[14] définit, dans le cadre européen les obligations qui s'attachent à l'usage de données confidentielles, **lorsque celles-ci sont exclusivement obtenues pour la production de statistiques**[15]. Comme la loi de 1951, le règlement 223 interdit dans ce cas toute communication, sauf à des fins de statistiques ou de recherche. Il impose aux États membres de prévoir des sanctions en cas de violation du secret statistique.

> ***La construction d'une statistique par agrégation n'est cependant pas une condition suffisante pour garantir le secret.***

Les résultats — agrégats, indicateurs, tableaux, graphiques et autres — issus de l'exploitation de données protégées par le secret statistique sont eux-mêmes soumis au secret statistique. Leur diffusion est possible s'ils ne permettent aucune réidentification des personnes ou organismes. La construction d'une statistique par agrégation n'est cependant pas une condition suffisante pour garantir le secret, pour peu que l'effectif que représente cet agrégat soit faible ou en fonction de la distribution de la valeur d'une caractéristique au sein de cet effectif (si par exemple une même caractéristique est partagée par tous les individus, ou si un individu prévaut excessivement sur les autres) ***(encadré 2)***.

14 Règlement (CE) N° 223/2009 du Parlement européen et du Conseil du 11 mars 2009 relatif aux statistiques européennes et abrogeant le règlement (CE, Euratom) n° 1101/2008 relatif à la transmission à l'Office statistique des Communautés européennes d'informations statistiques couvertes par le secret, le règlement (CE) n° 322/97 du Conseil relatif à la statistique communautaire et la décision 89/382/CEE, Euratom du Conseil instituant un comité du programme statistique des Communautés européennes (voir fondements juridiques).

15 Article 20 du règlement 223. Cet article ne s'applique cependant qu'autant que des données confidentielles sont nécessaires à la production et au développement de statistiques prévues par le programme statistique européen. (voir fondements juridiques).

La diffusion sous la contrainte des « règles du secret statistique »

Pour savoir quelles statistiques peuvent être librement diffusées, le SSP se réfère à des règles définies en fonction de critères simples.

La loi de 1951 ne spécifie pas directement d'obligation pour la diffusion de résultats statistiques ***(encadré 2)***. Le code des relations entre le public et l'administration indique que les données à caractère personnel ne peuvent être rendues publiques qu'après avoir été traitées de manière à empêcher l'identification des personnes concernées (article L312-1-2). Cependant, il ne précise ni les règles ni les méthodes à suivre pour atteindre cet objectif.

Dans tous les cas, la seule obligation résultant de la loi est de s'assurer que les résultats statistiques rendus publics ne permettent pas l'identification des personnes ni de leurs caractéristiques. Face à cette injonction, le SSP a dû mettre en place des méthodes

Encadré 2. Le secret statistique s'applique aussi aux résultats statistiques en complément des données individuelles

Le secret statistique a essentiellement pour fonction de garantir la confidentialité d'informations individuelles, relatives à des personnes physiques ou morales. L'article 6 de la loi de 1951 interdit ainsi la communication des « renseignements individuels figurant dans les questionnaires » des enquêtes statistiques*. Si l'on suit le texte à la lettre, on pourrait penser que le secret statistique n'est alors opposable qu'en cas de communication ou diffusion de données individuelles.

Cette interprétation méconnaît le fait que les résultats statistiques peuvent indirectement, par recoupement avec d'autres informations connues par ailleurs du public, révéler les caractéristiques de personnes.

Une statistique est une valeur numérique ou une mesure qui est utilisée pour résumer, analyser ou interpréter des données dans une population. Une statistique peut se présenter sous de très nombreuses formes : somme, moyenne, pourcentage, taux de croissance, etc. Une statistique occulte les informations individuelles dont elle est issue. Dès lors que l'on a calculé le revenu moyen des habitants pour un territoire donné, il n'est plus possible de retrouver, sur la seule base de cette information, le revenu de chacun des habitants de ce territoire. Une statistique est anonyme.

Cependant, cet anonymat peut être levé si l'on utilise d'autres informations disponibles ou accessibles grâce au lien entre des personnes et une statistique. À titre d'exemple, considérons la statistique constituée par le chiffre d'affaires total réalisé par des entreprises d'un secteur donné dans un territoire donné. Grâce au répertoire des entreprises et des établissements (Sirene), qui est public, on peut savoir quelles sont les entreprises concernées. Si deux entreprises seulement composent la population décrite par cette statistique, chaque entreprise, à partir de la connaissance de son propre chiffre d'affaires, est capable de déduire celui de sa concurrente.

À travers cet exemple, la réidentification de données individuelles est d'autant plus forte que la population concernée est peu nombreuse**, raison pour laquelle, et de longue date, le service statistique public a défini des « règles de secret statistique » selon des critères de taille d'effectifs.

** Sauf décision de l'administration des archives, prise après avis du comité du secret statistique.*

*** Voir à ce sujet les conclusions du rapporteur public pour la décision n° 186073 du 7 octobre 1998 du Conseil d'État.*

et des moyens facilement opérationnels pour répondre à cette obligation sans nuire aux besoins d'information du public. Traiter les résultats statistiques pour prévenir les risques de réidentification a pour conséquence dans la plupart des cas d'appauvrir le contenu de ces résultats. De manière intuitive, plus les résultats sont détaillés, plus les risques de réidentification sont grands, et donc, inversement, des résultats moins détaillés permettront de réduire ces risques.

Certaines de ces méthodes, relativement simples à mettre en œuvre (à l'exception du problème du secret secondaire, ***encadré 3***), consistent à définir des seuils de diffusion, également appelés «règles de secret statistique». Ces règles sont élaborées de manière à minimiser les risques de réidentification sans compromettre la pertinence ou l'intérêt des données diffusées. Parmi les règles de secret les plus connues au sein du SSP, on peut citer celles qui limitent la diffusion de valeurs pour les statistiques d'entreprises (une valeur ne doit pas s'appliquer à moins de trois unités[16] ou une unité ne doit pas représenter plus de 85% du total de la valeur[17]), ou encore celle qui s'applique aux statistiques issues de sources fiscales (une valeur ne doit pas représenter moins de 11 unités[18]).

Une fois établies, ces règles doivent également être appliquées par les utilisateurs ultérieurs de données individuelles protégées par le secret statistique, tels que les chercheurs y ayant accès *via* le Comité du secret statistique ***(infra et encadré 4).***

▶ Le secret statistique fondé dans le droit français et dans le droit européen

> ***Le secret statistique s'applique à toutes données de nature confidentielle collectées ou produites à des finalités statistiques.***

Le secret statistique s'applique à toutes données de nature confidentielle collectées ou produites à des finalités statistiques, qu'elles concernent des personnes physiques ou des personnes morales (entreprises, collectivités territoriales, associations, etc.).

Le secret statistique implique de la part des statisticiens l'engagement que les données confidentielles obtenues pour la production de statistiques, ne servent à aucune autre fin que l'établissement de statistiques ou à des travaux de recherche. Cet engagement exclut notamment tout usage à des fins de contrôle ou pour toute mesure ou décision à l'égard d'une personne en particulier.

La loi de 1951 établit trois régimes distincts par lesquels le service statistique public peut obtenir l'accès à des données individuelles confidentielles. Le bénéfice de ces régimes est lié à une obligation de secret. L'article 3 bis définit le régime d'accès aux bases de données détenues par des personnes morales de droit privé, les articles 1 bis, 2 et 6 bis définissent ensemble le régime applicable aux enquêtes statistiques obligatoires, et l'article 7 bis définit le régime d'accès aux données détenues par les administrations au sens large (y compris les personnes morales de droit privé exerçant une mission de service public).

16 Décision du directeur général de l'Insee du 13 juin 1980.
17 Règle de diffusion définie le 7 juillet 1960 par le Comité de coordination des enquêtes statistiques, prédécesseur du Cnis, Conseil national de l'information statistique.
18 Définie suite à un avis du 27 mai 1997 rendu par la Cnil et publiée au bulletin officiel des finances publiques (voir fondements juridiques).

▶ Encadré 3. Secrets primaire et secondaire

L'application de règles de secret statistique définies en fonction de seuils de diffusion conduit à « blanchir » ou masquer les cases de tableaux qui ne respectent pas ces seuils à des fins de publication. Il faut alors distinguer le *secret primaire* du *secret secondaire*.

Exemple d'un tableau de revenu déclaré mensualisé où un seuil de diffusion de 11 ménages est à appliquer

Nombre de ménages	< 1 000 €	Entre 1 000 et 2 000 €	> 2 000 €	Ensemble
Zone A	3	12	75	90
Zone B	15	35	30	80
Ensemble	18	47	105	170

Légende : **Secret primaire** **Secret secondaire**

Seule la case comprenant moins de 3 unités est sous le seuil des 11 personnes. Néanmoins, si seule celle-ci est masquée – *secret primaire* – il reste la possibilité d'en retrouver la valeur par différence entre le total en ligne ou en colonne et les valeurs des autres cases. Il faut donc masquer deux cases supplémentaires, l'une sur la même ligne, l'autre sur la même colonne, plus une autre case, pour empêcher le calcul de ces deux nouvelles cases masquées. Ce masquage de nouvelles cases pour éviter le calcul de valeurs directement soumises au secret statistique constitue ce qu'on appelle le *secret secondaire*.

Le tableau publié après secret statistique primaire et secondaire se présente alors de la manière suivante :

Nombre de ménages	< 1 000 €	Entre 1 000 et 2 000 €	> 2 000 €	Ensemble
Zone A	ss	ss	75	90
Zone B	ss	ss	30	80
Ensemble	18	47	105	170

où « ss » signifie « secret statistique ».

L'exemple ci-dessus correspond à la définition du secret secondaire dans sa version la plus simple. La gestion du secret secondaire se complique sensiblement s'il faut tenir compte du même tableau construit sur des périodes antérieures, ou de tableaux de mêmes caractéristiques produits à partir de sources de nature similaire (par exemple des effectifs salariés en nombre de personnes et en équivalent temps plein).

Cependant, il serait erroné de limiter l'application du secret statistique aux seules données obtenues par l'un ou l'autre de ces trois régimes.

En effet, ne considérer que les données confidentielles obtenues par les régimes d'accès définis par la loi de 1951 exclut les nombreux autres cas où le service statistique public détient des données confidentielles. Par exemple, les données obtenues par accord de gré à gré ainsi que celles dont le service statistique public est directement destinataire grâce à des dispositions législatives ou réglementaires spécifiques, comme c'est le cas pour la déclaration sociale nominative[19]. De plus, il convient de prendre en compte les données individuelles construites ou déduites par le service statistique public à partir des informations auxquelles il a accès. Ces données, qui sont issues de méthodes ou de calculs statistiques, ou qui peuvent être obtenues par appariement[20] sont des résultats statistiques individualisés.

▶ Encadré 4. Le Comité du secret statistique

Créé en 1984 et défini par la loi de 1951, le Comité du secret statistique est une commission administrative à caractère consultatif. Il relève des dispositions de l'article L311-8 du code des relations entre le public et l'administration, du chapitre II du 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, ainsi que de l'article 116 du décret n° 2019-536 du 29 mai 2019 pris pour l'application de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

Ce Comité, selon le premier alinéa de l'article 6 bis de la loi de 1951, « est appelé à se prononcer sur toute question relative au secret en matière de statistiques. Il donne son avis sur les demandes de communication de données individuelles collectées en application de la présente loi. »

La compétence du Comité a longtemps été restreinte aux données sur les entreprises. La loi de juillet 2008 sur les archives l'a étendue à celles recueillies sur les ménages. Le Comité exerce désormais sa compétence sur toutes les demandes d'accès à des données individuelles collectées par le service statistique public à des fins statistiques en application de la loi de 1951, aux données fiscales (article L135D du livre des procédures fiscales), et, depuis l'article L311-8 introduit par la loi pour une République numérique, à toute autre base administrative dès lors que l'administration concernée en saisit le Comité.

Placé sous la présidence d'un conseiller d'État, le Comité du secret statistique est composé de représentants des services producteurs de données, de chercheurs, de l'administration des Archives, de corps institutionnels (Assemblée nationale, Sénat), des entreprises et de leurs salariés, de la Cnil. Ses membres actifs en sont les services producteurs – pour l'essentiel le service statistique public –, les chercheurs et les Archives.

Son secrétariat est assuré par l'unité des Affaires juridiques et contentieuses de l'Insee. Le Comité entretient des liens étroits avec le CASD*, groupement d'intérêt public qui organise et met à disposition des services des accès sécurisés pour les données confidentielles à des fins non lucratives de recherche, d'étude ou d'évaluation. Il est également en relation avec Quetelet PROGEDO Diffusion, portail d'accès dont la mission est de mettre à la disposition de la communauté scientifique les bases de données et enquêtes en sciences humaines et sociales produites par la statistique publique (Insee, services statistiques des ministères, autres institutions gouvernementales et collectivités territoriales) et par le monde académique (organismes de recherche et universités).

** Centre d'accès sécurisé aux données.*

19 Décret n° 2013-266 du 28 mars 2013 relatif à la déclaration sociale nominative (voir fondements juridiques).
20 Un appariement consiste en l'interconnexion d'au moins deux fichiers sur la base d'un identifiant ou de données d'identité communes. Un appariement permet d'établir une relation nouvelle entre des données, absentes des

Leur utilisation à d'autres fins que statistiques peut ne pas être neutre pour les personnes concernées.

Au-delà du droit français, les obligations relatives au secret statistique découlent aussi des obligations que définissent les règlements européens, règlements sectoriels propres à un domaine particulier, ou règlements transversaux comme le règlement général sur la protection des données (RGPD, voir plus bas) ou le règlement 223 sur les statistiques européennes.

> ***Au-delà du droit français, les obligations relatives au secret statistique découlent aussi des obligations que définissent les règlements européens.***

Le principe de finalité, au regard du considérant 162[21] du RGPD, implique ainsi que des données relatives à des personnes physiques et traitées pour des finalités statistiques ne peuvent être réutilisées qu'à des fins statistiques ou de recherche, ce qui en pratique revient à imposer le secret statistique à la diffusion ou la communication de ces données. Le règlement 223 interdit toute communication de données confidentielles obtenues pour la production de statistiques européennes, sauf motif statistique ou de recherche. Relevant de règles définies par une norme européenne, de rang supérieur, ces données échappent donc à toute règle de communication ou de diffusion applicable strictement en droit français.

La protection des données personnelles renforce les obligations de confidentialité

À partir de 1978, la loi Informatique et libertés vient définir les obligations qui s'attachent à la protection des données des personnes physiques, lorsque ces données font l'objet de traitements automatisés, y compris les traitements statistiques. Elle renforce, pour le SSP déjà soumis au secret statistique, les obligations relatives à la protection de la vie privée.

Tout comme le secret statistique, la protection des données personnelles que définit la loi Informatique et libertés s'applique aussi longtemps que des données rendent possibles l'identification de personnes ou de leurs caractéristiques.

La loi de 1978, en commun avec la loi de 1951, vise pour les personnes physiques la protection de la confidentialité de leur vie privée. La loi de 1978 fait néanmoins entrer cette protection dans une nouvelle dimension, en tenant compte des risques spécifiques inhérents à l'automatisation des traitements de données individuelles.

données disjointes d'origine. On peut apparier par exemple des données d'origine fiscale avec des données issues des régimes d'assurance sociale sur les prestations versées, afin d'établir le revenu disponible des ménages et calculer des taux de pauvreté.

21 Par « fins statistiques », on entend toute opération de collecte et de traitement de données à caractère personnel nécessaires pour des enquêtes statistiques ou la production de résultats statistiques. Ces résultats statistiques peuvent en outre être utilisés à différentes fins, notamment de recherche scientifique. Les fins statistiques impliquent que le résultat du traitement opéré ne constitue pas des données à caractère personnel mais des données agrégées, et que ce résultat ou ces données à caractère personnel ne sont pas utilisés à l'appui de mesures ou de décisions concernant une personne physique en particulier.

> ***La loi de 1978, en commun avec la loi de 1951, vise pour les personnes physiques la protection de la confidentialité de leur vie privée.***

À partir des années 60 en effet, le développement de l'informatique a ouvert la possibilité de collecter et traiter de grands volumes de données et de faciliter leur concentration et leur recoupement. Par les obligations qu'elle impose aux responsables de traitement, la loi Informatique et libertés cherche à répondre aux conséquences pour la protection de la vie privée de l'informatisation croissante de nos sociétés.

La loi Informatique et libertés a instauré de nouvelles obligations qui requièrent de la part des responsables de traitement de données de justifier la légitimité des finalités poursuivies, de garantir que les données traitées sont adaptées à ces finalités et de limiter leur conservation à la durée strictement nécessaire pour atteindre ces finalités.

Le renforcement des objectifs de protection va de pair avec des sanctions fortes en cas de manquement, plus fortes qu'en cas de violation du secret professionnel ou statistique. Les sanctions pénales actuellement prévues par la loi Informatique et libertés peuvent ainsi atteindre 300 000 euros d'amende et 5 ans d'emprisonnement.

La loi Informatique et libertés fait par ailleurs intervenir deux nouveaux acteurs dans le champ de la protection des données personnelles :

- La personne concernée par les données, absente de la loi de 1951, qui bénéficie non seulement d'un droit d'information sur les traitements qui la concerne, mais aussi le droit d'accéder à ses données, de les faire éventuellement corriger, et dans certains cas de s'opposer à leur traitement ou à leur conservation ;
- La Commission nationale Informatique et libertés (Cnil), en tant qu'autorité de contrôle qui agit aussi dans le cadre de missions de conseil auprès des responsables de traitement et de sensibilisation du public. Jusqu'en 2018, la Cnil agit par contrôle *a priori* systématique : tout traitement de données personnelles est soumis à une formalité de déclaration obligatoire auprès de la Cnil, qui pour les traitements les plus sensibles par leur impact sur les personnes concernées peut impliquer une autorisation ou un avis de la Cnil.

Depuis 2018, la loi de 1978 s'inscrit dans le cadre juridique européen harmonisé défini par le règlement général sur la protection des données (RGPD). Le RGPD renforce les droits des personnes concernées, notamment le droit à l'information sur les traitements, d'où une obligation de transparence plus grande encore que par le passé.

L'action de la Cnil évolue, d'un contrôle *a priori* vers un contrôle *a posteriori* : la mise en œuvre d'un traitement ne dépend plus systématiquement d'une formalité auprès de la Cnil. Néanmoins, la conformité du traitement peut être vérifiée à tout moment et il incombe au responsable de traitement de s'assurer par lui-même de cette conformité. Pour cela, la loi impose la tenue d'un registre des activités de traitement et la réalisation d'études d'impact sur la vie privée pour les traitements les plus sensibles, en se faisant assister par un délégué à la protection des données, qui agit auprès du responsable de traitement pour des missions d'assistance et de conseil et comme relais auprès de la Cnil.

Confidentialité et *Open Data*, des injonctions contradictoires ?

Protection des données personnelles, secret statistique ne sont pas les seules obligations légales liées à la collecte et la production de données statistiques auxquelles le SSP est tenu de se conformer.

Comme toute administration, l'Insee et les services statistiques ministériels sont tenus de répondre aux demandes de communication relatives aux documents qu'ils produisent ou collectent, sur le fondement soit du Code des relations entre le public et l'administration, soit du Code du patrimoine. Ce dernier organise spécialement l'accès aux archives publiques, sachant que tout document administratif constitue dès sa production une archive publique.

Chacun a le droit de demander et d'obtenir l'accès aux documents et archives produits par les administrations, sous réserve que cette communication ne porte pas « une atteinte excessive aux intérêts que la loi a entendu protéger » autrement dit aux secrets définis par la loi. Les bases de données statistiques sont des documents ou des archives publiques au sens où le définissent ces deux Codes et sont donc soumises au droit d'accès.

Le Code des relations entre le public et l'administration organise par ailleurs les conditions de diffusion des documents – au sens large – produits par les administrations. Ces documents ne peuvent être diffusés sur un site Internet par exemple, que dans le respect des secrets définis par la loi et sous réserve que les informations relatives à des personnes physiques soient préalablement traitées pour ne permettre aucune réidentification de ces personnes, autrement dit anonymisées.

> ***Avec la loi pour une République numérique ou loi Lemaire, le paradigme change.***

Jusqu'en 2016, la diffusion était une possibilité offerte aux administrations. Il leur était possible de publier ou pas. Avec la loi pour une République numérique[22] ou loi Lemaire, le paradigme change. La diffusion devient la règle ; tout document, dès lors qu'il est communiqué, doit être diffusé, et cette diffusion doit se faire par publication sur un site Internet.

La France franchit un nouveau cap dans l'ouverture des données publiques ou *Open Data*, conçu comme un vecteur de transparence et d'amélioration de l'action publique ainsi qu'un puissant levier pour l'innovation économique. L'évolution de la législation française s'inscrit dans un mouvement porté au sein de l'Union européenne par la « stratégie européenne pour les données », dont le RGPD, le règlement sur la gouvernance de la donnée, adopté le 30 mai 2022, et le projet de règlement sur la donnée constituent les principaux vecteurs juridiques.

Néanmoins, l'obligation de diffusion portée par la loi Lemaire, en contraignant le SSP à réévaluer les risques de réidentification de certaines de ses bases de données, a eu des conséquences inattendues, voire paradoxales.

22 Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (voir fondements juridiques).

▶ Le paradoxe des fichiers de production et de recherche

L'Insee et les services statistiques ministériels ont de longue date une politique d'ouverture relativement généreuse de leurs bases de données individuelles statistiques, dans le respect des règles de confidentialité et du secret statistique, à destination en particulier du monde de la recherche.

Cette politique s'organise sous un régime de communication et d'accès restreints, essentiellement sous deux formes :

- Des fichiers complets, à l'exception des données identifiantes pour les personnes, accessibles aux chercheurs ainsi qu'aux services publics en charge de missions d'établissement de statistique ou assimilables à des travaux de recherche (l'évaluation des politiques publiques par exemple), après avis du Comité du secret statistique, qui s'assure en particulier de la compatibilité de la demande avec des finalités statistiques ; les demandeurs dans la grande majorité des cas n'accèdent aux données que *via* le centre d'accès sécurisé aux données ; l'ensemble de la procédure, par les contraintes et les obligations qu'elle impose aux demandeurs, est définie de façon à garantir les conditions optimales de protection de la confidentialité des données ***(encadré 4)*** ;
- Des fichiers de données individuelles anonymisées, dits fichiers de production et de recherche (FPR). Jusqu'en 2017, il suffisait pour y accéder de signer une licence d'usage final, ce qui permettait aux chercheurs, aux administrations et aux organismes privés à des fins commerciales (tels que des bureaux de conseil ou des agences de marketing) d'en obtenir la communication.

> ***Le risque de réidentification est toujours plus élevé pour des données mises en ligne et rendues publiques.***

S'il était admis que les FPR étaient anonymisés dans un contexte de **communication**, ce n'était plus le cas s'ils devaient être **publiés**. La mise en ligne d'un fichier de données individuelles l'expose en effet à des risques de tentatives de réidentification menées à l'aide d'autres informations, connues ou rendues publiques, sur les personnes concernées. Des travaux de recherche (*Narayanan et Shmatikov*, 2008 ainsi que *Sweeney*, 1997) ont montré la vulnérabilité de données mises en ligne, bien qu'elles aient fait l'objet d'une anonymisation poussée. En d'autres termes, le risque de réidentification est toujours plus élevé pour des données mises en ligne et rendues publiques.

Dans le contexte de la loi Lemaire, le SSP s'est trouvé contraint de publier ces FPR, et ainsi de courir le risque de rompre la confidentialité des données, sauf à décider d'en arrêter la communication. En définitive, ces fichiers sont restés accessibles aux chercheurs et aux services publics pour des finalités statistiques, moyennant une procédure qui soumet maintenant leur communication à un avis du Comité du secret statistique. En revanche ces fichiers, désormais couverts par le secret statistique, ne sont plus accessibles pour des motifs non statistiques, notamment pour des utilisations commerciales.

Ainsi, paradoxalement la loi voulue pour promouvoir l'accès le plus large possible aux données publiques, a eu pour conséquence de restreindre l'accès à certaines données statistiques. Ce que l'on peut considérer comme une manifestation de l'attachement viscéral du SSP à la protection du secret statistique et de la confidentialité des données qu'il détient.

▶ Bibliographie

- NARAYANAN, Arvind et SHMATIKOV Vitaly, 2008. *How To Break Anonymity of the Netflix Prize Dataset*. [Consulté le 26 mai 2023]. Disponible à l'adresse : *https://philpapers.org/rec/SWEWTA-2*.

- ROCHER, Luc, HENDRICKX, Julien M. et de MONTJOYE, Yves-Alexandre, 2019. *Estimating the success of re-identifications in incomplete datasets using generative models*. In : *Nature Communications*. [en ligne]. 23 juillet 2019. [Consulté le 26 avril 2023]. Disponible à l'adresse : *https://www.nature.com/articles/s41467-019-10933-3*.

- SWEENEY Latanya, 1997. *Weaving Technology and Policy Together to Maintain Confidentiality*. Volume 25, Issue 2-3. [Consulté le 26 mai 2023]. Disponible à l'adresse : *https://philpapers.org/rec/SWEWTA-2*.

▶ Fondements juridiques

- Règlement (CE) n°223/2009 du Parlement européen et du Conseil du 11 mars 2009 relatif aux statistiques européennes et abrogeant le règlement (CE, Euratom) n°1101/2008 relatif à la transmission à l'Office statistique des Communautés européennes d'informations statistiques couvertes par le secret, le règlement (CE) n°322/97 du Conseil relatif à la statistique communautaire et la décision 89/382/CEE, Euratom du Conseil instituant un comité du programme statistique des Communautés européennes. In : *Journal officiel de l'Union européenne*. [en ligne]. Mise à jour le 08 juin 2015. [Consulté le 17 mai 2023]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX%3A32009R0223.*

- Article L312-1-2 du Code des relations entre le public et l'administration. In : *site de Légifrance*. [en ligne]. [Consulté le 17 mai 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000033205514.*

- Code du patrimoine. Version en vigueur au 25 mai 2023. [Consulté le 25 mai 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006074236*.

- Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance.* [en ligne]. Mise à jour le 25 mars 2019. [Consulté le 17 mai 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. In : *site de Légifrance.* [en ligne]. Mise à jour le 26 janvier 2022. [Consulté le 17 mai 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460*.

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance.* [en ligne]. Mise à jour le 11 mars 2023. [Consulté le 17 mai 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

- Bulletin officiel des finances publiques : *https://bofip.impots.gouv.fr/bofip/7248-PGP.html/identifiant%3DBOI-DJC-CADA-20-20220126*.