Insee
Mesurer pour comprendre
Direction générale

CONCOURS PROFESSIONNEL 2022 POUR L'ACCÈS AU GRADE DE CONTRÔLEUR PRINCIPAL DE L'INSEE

Septembre 2022

ÉPREUVE DE RÉDACTION D'UNE NOTE DE SYNTHÈSE

(durée 3 heures – coefficient 2)

SUJET B :

RÉDACTION D'UNE NOTE DE SYNTHÈSE À L'AIDE DES ÉLÉMENTS D'UN DOSSIER À CARACTÈRE ADMINISTRATIF

Le sujet comporte 21 pages

Une attention particulière sera accordée à la présentation, à l'orthographe et à la syntaxe.
L'usage de la calculatrice n'est pas autorisé.

La loi pour une République numérique

Vous expliciterez les conséquences de la loi pour une République numérique sur la statistique publique. Dans un premier temps vous décrirez son impact sur la diffusion des données à l'Insee. Vous expliquerez ensuite comment cette loi contribue à renforcer l'accès aux données. Enfin, vous montrerez comment elle permet de simplifier certains traitements de données à caractère personnel.

Documents :

<u>Document 1</u> :Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique *(Publiée sur le site de legifrance.gouv.fr le 8 Octobre 2016)* : extraits commentés par l'inspection générale de l'Insee *(d'après le rapport « Conséquences pour l'Insee de la loi pour une République numérique » du 12 janvier 2017)*

[...]

Article 12

I.-Le chapitre IV du titre II du livre III du même code est ainsi modifié :
1° A la première phrase de l'article L. 324-4, les mots : « de ces redevances » sont remplacés par les mots : « des redevances mentionnées aux articles L. 324-1 et L. 324-2 » ;
2° Il est ajouté un article L. 324-6 ainsi rédigé :

« Art. L. 324-6.-La réutilisation des informations publiques produites par le service statistique public mentionné à l'article 1er de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques ne peut donner lieu au versement d'une redevance. »

II.-Le 2° du I du présent article entre en vigueur le 1er janvier 2017.

Commentaires & conséquences pour l'Insee

La liste des « informations publiques » détenues par le SSP tombant sous le coup de cette mesure devra être établie clairement. Il est déjà assuré que la base SIRENE en ferait partie. Mais d'autres bases de données pourraient être concernées (fichiers des décès issus du RNIPP, etc.).

La notion de documents administratifs et d'informations publiques (voir paragraphe 3.3.1) renvoie à des documents préexistants, achevés dans une version non provisoire, ne contenant pas d'informations protégées. Les documents résultant de travaux réalisés spécifiquement pour un client n'échappent pas à cette définition, même s'ils peuvent difficilement en première analyse faire l'objet d'une communication.

Toutefois les documents issus de ces travaux peuvent ultérieurement, dans le cadre d'une politique de diffusion et sous les réserves des protections déjà mentionnées, entrer dans le champ des documents administratifs publiables.

[…]

Article 19

La loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques est ainsi modifiée :

1° Le second alinéa de l'article 3 est supprimé ;

2° Après le même article 3, il est inséré un article 3 bis ainsi rédigé :

« Art. 3 bis. - I. - Le ministre chargé de l'économie peut décider, après avis du Conseil national de l'information statistique, que les personnes morales de droit privé sollicitées pour des enquêtes transmettent par voie électronique sécurisée au service statistique public, à des fins exclusives d'établissement de statistiques, les informations présentes dans les bases de données qu'elles

détiennent, lorsque ces informations sont recherchées pour les besoins d'enquêtes statistiques qui sont rendues obligatoires en application de l'article 1er bis.

« Cette décision est précédée d'une concertation avec les personnes de droit privé sollicitées pour ces enquêtes et d'une étude de faisabilité et d'opportunité rendue publique.

« Les données transmises par ces personnes morales ne peuvent faire l'objet d'aucune communication de la part du service dépositaire. Seules sont soumises au livre II du code du patrimoine les informations issues de ces données qui ont été agrégées et qui ne permettent pas l'identification de ces personnes morales.

« Les conditions dans lesquelles sont réalisées ces enquêtes, notamment leur faisabilité, leur opportunité, les modalités de collecte des données de même que, le cas échéant, celles de leur enregistrement temporaire et celles de leur destruction sont fixées par voie réglementaire. « II. - Par dérogation à l'article 7, en cas de refus de la personne morale sollicitée pour l'enquête de procéder à la transmission d'informations conformément à la décision prise dans les conditions mentionnées au I du présent article, le ministre chargé de l'économie met en demeure cette personne. Cette mise en demeure fixe le délai imparti à la personne sollicitée pour l'enquête pour faire valoir ses observations. Ce délai ne peut être inférieur à un mois.

« Si la personne sollicitée pour l'enquête ne se conforme pas à cette mise en demeure, le ministre saisit pour avis le Conseil national de l'information statistique, réuni en comité du contentieux des enquêtes statistiques obligatoires. La personne sollicitée pour l'enquête est entendue par le comité. « Au vu de cet avis, le ministre peut, par une décision motivée, prononcer une amende administrative. Passé un délai de deux ans à compter de la date de réception de la mise en demeure, le ministre ne peut plus infliger d'amende.

« Le montant de la première amende encourue à ce titre ne peut dépasser 25 000 €. En cas de récidive dans un délai de trois ans, le montant de l'amende peut être porté à 50 000 € au plus. « Le ministre peut rendre publiques les sanctions qu'il prononce. Il peut également ordonner leur insertion dans des publications, journaux et supports qu'il désigne, aux frais des personnes sanctionnées. »

Commentaires & conséquences pour l'Insee

L'article 19 étend l'obligation de réponse (pilier de la loi de 1951) au cas d'informations issues de certaines bases de données. Il permet à la statistique publique de se voir transmettre sous forme électronique des informations issues de ces bases dans le seul but de réaliser des enquêtes statistiques obligatoires, et ce afin de simplifier des processus manuels actuels qui sont longs et coûteux.

Cette procédure est entourée de nombreuses précautions à la hauteur des enjeux : concertation ex-ante avec les personnes concernées ; adaptation du dispositif contentieux en cas de refus de réponse ; délais de prescription de deux ans. Les conditions de la transmission sont définies par voie réglementaire.

Cet article d'inspiration SSP vise clairement le cas où des transmissions massives d'informations pourront simplifier grandement le processus de construction des données (exemple des données de caisse).

[…]

Article 34

La loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés est ainsi modifiée :

1° Après le I de l'article 22, il est inséré un I bis ainsi rédigé :

« I bis.-Par dérogation au 1° des I et II de l'article 27, font également l'objet d'une déclaration auprès de la Commission nationale de l'informatique et des libertés les traitements qui portent sur des données à caractère personnel parmi lesquelles figure le numéro d'inscription des personnes au répertoire national d'identification des personnes physiques ou qui requièrent une consultation de ce répertoire, lorsque ces traitements ont exclusivement des finalités de statistique publique, sont mis en œuvre par le service statistique public et ne comportent aucune des données mentionnées au I de l'article 8 ou à l'article 9, à la condition que le numéro d'inscription à ce répertoire ait préalablement fait l'objet d'une opération cryptographique lui substituant un code statistique non signifiant, ainsi que les traitements ayant comme finalité exclusive de réaliser cette opération cryptographique. L'utilisation du code statistique non signifiant n'est autorisée qu'au sein du service statistique public. L'opération cryptographique est renouvelée à une fréquence définie par décret en Conseil d'Etat pris après avis motivé et publié de la Commission nationale de l'informatique et des libertés. » ;

2° Le I de l'article 25 est complété par un 9° ainsi rédigé :

« 9° Par dérogation au 1° du I et aux 1° et 2° du II de l'article 27, les traitements qui portent sur des données à caractère personnel parmi lesquelles figure le numéro d'inscription des personnes au répertoire national d'identification des personnes physiques ou qui requièrent une consultation de ce répertoire, lorsque ces traitements ont exclusivement des finalités de recherche scientifique ou historique, à la condition que le numéro d'inscription à ce répertoire ait préalablement fait l'objet d'une opération cryptographique lui substituant un code spécifique non signifiant, propre à chaque projet de recherche, ainsi que les traitements ayant comme finalité exclusive de réaliser cette opération cryptographique. L'opération cryptographique et, le cas échéant, l'interconnexion de deux fichiers par l'utilisation du code spécifique non signifiant qui en est issu ne peuvent être assurés par la même personne ni par le responsable de traitement. L'opération cryptographique est renouvelée à une fréquence définie par décret en Conseil d'Etat pris après avis motivé et publié de la Commission nationale de l'informatique et des libertés. » ;

Article 34 : pseudonymisation du NIR

L'article 34 modifie les dispositions de l'article 27 de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés en introduisant deux nouvelles procédures spécifiques d'accès à certaines données publiques à des fins statistiques ou à des fins de recherche publique.

Traitement statistique de données individuelles non sensibles : la première concerne « les traitements qui portent sur des données à caractère personnel parmi lesquelles figure le NIR ou qui requièrent une consultation du RNIPP, lorsque ces traitements ont exclusivement des *finalités de statistique publique,* sont mis en œuvre par le SSP » et ne comportent pas de données à caractère personnel (ethnie, religion, opinions, santé, condamnations, etc.).
Pour ces traitements, à l'actuel régime d'autorisation par un décret du Conseil d'État, se substitue le (simple) régime de déclaration à la CNIL prévu à l'article 22 de la loi du 6 janvier 1978.

Traitements à des fins de recherche publique : la seconde concerne les traitements qui portent sur des données à caractère personnel parmi lesquelles figure le NIR ou qui requièrent une consultation du RNIPP, lorsque ces traitements ont exclusivement des finalités de *recherche scientifique ou historique.*
Pour ces traitements, à la place de l'actuel régime d'autorisation par un décret du Conseil d'État, se substitue le régime d'autorisation par arrêté après avis de la CNIL tel que prévu à l'article 25.

La condition technique permettant ces assouplissements est l'introduction du NIR « haché » :
le NIR aura préalablement fait l'objet d'une opération cryptographique lui substituant un code statistique (ou spécifique) non signifiant.

Pour les traitements à finalité de statistique publique, l'utilisation du code non signifiant n'est autorisée qu'au sein du service statistique public.
Pour les traitements à des fins de recherche, le code spécifique non signifiant est propre à chaque projet de recherche. L'opération cryptographique et, le cas échéant, l'interconnexion de deux fichiers par l'utilisation du code spécifique non signifiant qui en est issu ne peuvent être assurés par la même personne ni par le responsable de traitement.

Dans les deux cas, les traitements ayant comme finalité exclusive de réaliser cette opération cryptographique font également l'objet d'une déclaration auprès de la CNIL.
Dans les deux cas, le législateur stipule que l'opération cryptographique est renouvelée à une fréquence définie par décret en Conseil d'État pris après avis motivé et publié de la CNIL. »

<u>Document 2</u> : La diffusion en accès libre et gratuit des données : l'exemple du répertoire Sirene

Paru sur le blog de l'Insee, le 16 septembre 2020

Dès 2003, l'Insee a fait le choix de l'open data en prenant la décision d'ouvrir l'accès sur son site internet à toutes les données qu'il produit, ainsi que ses publications. La mise à disposition de l'interface de programmation ou API Sirene permet de faire face au changement d'échelle de l'open data. Elle s'intègre par ailleurs dans une démarche nationale de simplification administrative, plus connue sous le nom « Dites-le nous une fois ». À l'occasion du salon Big Data 2020 qui s'est tenu les 14 et 15 septembre au parc des expositions de la Porte de Versailles à Paris, nos experts font le point, en trois questions, sur une offre de données par API qui ne va pas s'arrêter là.

Pourquoi une diffusion des données Sirene par API ?

Le programme « Dites-le nous une fois » a pour objectif de simplifier les démarches administratives des entreprises en leur évitant de déclarer plusieurs fois leurs données d'identité (raison sociale, adresse…). Dans ce cadre, et pour rationaliser le système d'information de l'État, les administrations peuvent interfacer leur système d'information avec les données du répertoire inter-administratif Sirene mises à jour quotidiennement via une API (Application Programming Interface, interface de programmation applicative). Cette API est exposée sur le catalogue des API de diffusion de l'Insee, ouvert en juillet 2018.

Par ailleurs, l'API-Sirene a permis de construire une offre de diffusion open data cohérente et complémentaire, via trois canaux : l'API elle-même avec les services de requêtes unitaires ou multiples sur des siren (identifiant d'entreprise) et des SIRET (identifiant d'établissement) ; le site sirene.fr, orienté plus grand public, qui appelle l'API pour les recherches et la constitution de listes ; et les fichiers stocks téléchargeables déposés chaque premier du mois sur data.gouv.fr.

L'Insee conforte ainsi son rôle d'acteur de l'open data avec cette offre API, Sirene étant l'une des neuf bases du service public de la donnée : en moyenne, 20 000 listes sont téléchargées mensuellement ; 6 000 comptes sont connectés à l'API Sirene et jusqu'à 23 millions de requêtes par mois sont effectuées en 2020.

Quelles requêtes les utilisateurs effectuent-ils à partir de l'API Sirene ?

Plus de la moitié des requêtes effectuées correspondent à des recherches unitaires sur siren ou siret : les utilisateurs vérifiant qu'un siren ou un siret existe, ou cherchant les données associées à ces identifiants. Ensuite, une part importante correspond à des requêtes multicritères : les utilisateurs cherchent à obtenir les unités légales ou les établissements qui correspondent à plusieurs critères. Pour ces recherches multicritères, il y a trois sortes d'utilisations : la mise à jour du référentiel (en sélectionnant l'ensemble des unités légales ou établissements qui ont été mis à jour depuis une certaine date), l'identification (« je ne connais pas le numéro siren ou siret mais je le recherche à partir d'éléments dont je dispose, par exemple la raison sociale ou l'adresse »), ou la liste, à des fins de recherche d'emploi, d'étude, de prospection commerciale, etc.

Quelles évolutions à venir en matière d'open data à l'Insee ?

Les évolutions à venir s'inscrivent dans une perspective historique de près de deux décennies.

L'Insee, précurseur de l'open data, a mis en place très tôt, dès 2003, une politique d'ouverture des données. Elle accompagnait l'essor d'internet, canal par excellence de diffusion directe de l'information auprès de tous les utilisateurs. Les deux bénéfices les plus fréquemment évoqués de l'open data sont la possibilité de valoriser les données, et de favoriser l'innovation par la data. Pour les institutions publiques s'y ajoute la transparence de l'action de l'État. Toutes ces valeurs rejoignent celles de la statistique publique, mise au service des citoyens, pour éclairer les débats publics.

En complément de la mise en ligne gratuite des publications et des fichiers de données, l'institut a étendu les formats de ces derniers à des standards ouverts comme csv et xml. Puis, à partir de 2018, il s'est lancé dans une politique de diffusion par API. Après Sirene et les nomenclatures associées, sont désormais accessibles depuis cette année d'une part nos séries longues d'indicateurs économiques et sociaux, et d'autre part nos données locales (aux mailles communales et au-delà).

Une API sur les métadonnées sera prochainement proposée au public : non seulement les nomenclatures, mais aussi les concepts, sources et définitions seront accessibles aux utilisateurs par API, en complément des consultations actuelles sur le site insee.fr. Les systèmes d'information des utilisateurs pourront ainsi directement et automatiquement se servir des bons concepts et définitions de variables lorsqu'ils utilisent nos sources : un gage d'efficacité et de cohérence de l'information.

À moyen terme, toutes nos bases de données ont vocation à être mises à disposition par API, et leurs variables mises en cohérence avec notre catalogue central de métadonnées (définitions et concepts). L'objectif de l'Insee, à travers ce projet ambitieux compte tenu de l'étendue de son offre, 5 000 nouveaux fichiers de données chaque année, 150 000 séries chronologiques, est de créer de la valeur pour les utilisateurs en mettant à disposition des outils performants et en favorisant l'interopérabilité entre les données, qu'elles viennent de l'Insee ou d'autres sources.

Enfin, à plus long terme, nous menons un projet de recherche en partenariat avec d'autres instituts de statistique européens, pour préparer la prochaine révolution dans le domaine de la data : rendre les données interopérables au niveau le plus fin. Cela permettra à l'utilisateur de formuler des requêtes à distance sur nos statistiques sans passer par un fichier ou une base de données. Cette prochaine révolution porte le nom de « statistiques ouvertes liées » (linked open statistics), comme composante des données ouvertes liées. Nous parlerons alors de « lacs de données » (datalakes), de requêtes « SPARQL », etc. Mais ceci est une autre histoire.

<u>Document 3</u> : Extrait du Courrier des Statistiques N°3

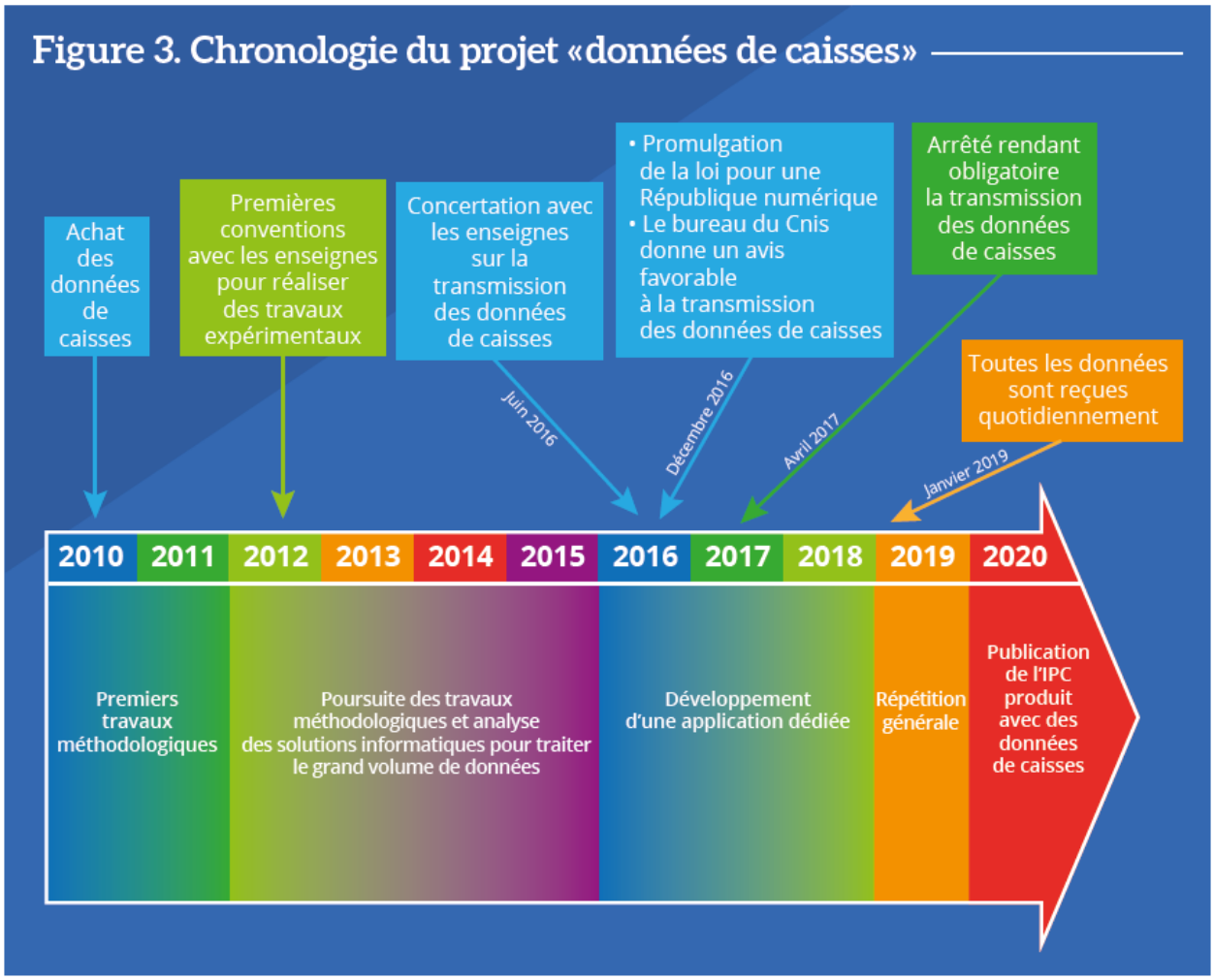
Paru en décembre 2019

LE LONG CHEMIN POUR ACCÉDER AUX DONNÉES PRIVÉES

Si l'apport des données de caisses pour la production de statistiques de prix est incontestable, leur utilisation par l'Insee n'est pas sans poser un certain nombre de difficultés.

La première, dans l'ordre chronologique, est tout simplement de pouvoir y accéder : les données de caisses sont des actifs incorporels des entreprises qui les produisent, et celles-ci n'avaient donc pas d'obligation à en donner un accès à l'Insee, même à des fins d'intérêt général pour la production de statistiques publiques. Dans un premier temps, des contacts ont été pris avec certaines enseignes afin de les convaincre de les transmettre à l'Insee. Depuis 2012, quatre enseignes (40 % environ du marché de la grande distribution) fournissaient ainsi les données à titre expérimental et dans le cadre de conventions ***(figure 3)***. Pour d'une part obtenir les données sur l'ensemble du champ des super et hypermarchés, et d'autre part pérenniser la mise à disposition, la loi de 1951 sur l'obligation, la coordination et le secret en matière statistique a été amendée. La loi[7] prévoit désormais la possibilité de rendre obligatoire la transmission de certaines données privées, après concertation des acteurs et uniquement pour remplacer des enquêtes statistiques obligatoires. Cet amendement, outre l'accès aux données de caisses, pourrait permettre de faciliter, à terme, l'accès à d'autres données privées.

Figure 3. Chronologie du projet «données de caisses»



degré de finesse nécessaire ; le champ des données de caisses peut être aisément complété par des données d'enquêtes (dans les autres formes de vente, pour les produits frais vendus en super et hypermarché, par exemple) pour couvrir l'ensemble de la consommation des ménages ; il couvre même dans certains cas mieux la consommation des ménages, en intégrant les données sur le *drive* par exemple, non couvert par l'IPC jusqu'à présent.

Par ailleurs, les données de caisses se distinguent au sein du *big data* parce qu'il s'agit en réalité de données très structurées, qui n'ont du *big data* que les deux premiers « V » du volume et de la vélocité.

Au total, alors que d'autres pistes d'utilisation du *big data* visent plutôt à produire de nouvelles statistiques, complémentaires mais non substituts de statistiques publiques existantes, les données de caisses ont cela de spécifique qu'elles peuvent réellement se substituer à des données d'enquête sans avoir à modifier les concepts ou le cadre méthodologique de ce que l'on veut mesurer.

- Ce dictionnaire de codes-barres permet également de traiter correctement le cas des « **relances commerciales** ». Ces relances consistent en une modification marginale du *packaging* du produit avec souvent un prix en hausse, ou un prix stable mais avec un volume de produit vendu plus faible. Ces relances commerciales masquent en général une hausse de prix, il est donc fondamental de pouvoir les repérer. Dans une collecte sur le terrain, c'est l'enquêteur qui identifie la relance ; avec les données de caisses, le code-barres change en général avec le *packaging* et il faut être capable, de manière automatique, de relier le produit initial à sa relance. En Suède, Tongur (2019) estime à 0,1 point le biais qui pourrait s'ensuivre si la relance n'était pas identifiée dans les données de caisses. Dans le cas français, l'existence d'un dictionnaire de codes-barres permet de faire le lien nécessaire entre un produit et sa relance commerciale et de bien enregistrer la hausse de prix associée.

1,7 MILLIARD D'ENREGISTREMENTS : NOUVELLE ARCHITECTURE INFORMATIQUE…

… ET NÉCESSAIRE AUTOMATISATION DES PROCESSUS STATISTIQUES

D'un point de vue statistique, le volume des données ne permet plus les traitements manuels qui pouvaient être effectués, en général par les enquêteurs, et qui doivent donc être automatisés. Trois exemples peuvent être donnés : être capable de classer un produit dans une nomenclature détaillée, identifier les relances commerciales et remplacer un produit lorsqu'il disparaît.

- Dans les données de caisses, **les produits sont identifiés par leur code-barres** ***(figure 1)*** et celui-ci ne donne pas d'information sur la nature du produit suivi. Compte tenu du nombre de codes-barres présents dans les données de caisses (près de 9 millions), il est impensable de rechercher manuellement à quel produit correspond chacun d'eux. Pour résoudre ce problème, l'Insee achète à un panéliste un dictionnaire de codes-barres, décrivant très précisément les caractéristiques du produit associé à chaque code-barres. **Classer les produits** revient alors à construire une simple table de passage entre ce dictionnaire et la nomenclature Coicop.

- Le **remplacement des produits** appartenant au panier de l'IPC et disparaissant en cours d'année est également une opération impliquant fortement les enquêteurs. Ils choisissent le produit remplaçant de manière à être le plus proche possible du produit disparu et décident s'il est nécessaire ou non d'effectuer un ajustement qualité[10]. Cette opération stratégique pour l'IPC peut elle aussi être automatisée (Léonard *et alii,* 2017). Le produit remplaçant est tiré aléatoirement parmi les produits de la même variété et un ajustement qualité est réalisé systématiquement, en comparant le prix du produit remplacé et du produit remplaçant au cours d'une même période : la différence de qualité est estimée égale à la différence de prix. Cette méthode d'ajustement de la qualité est couramment utilisée dans l'IPC, mais les prix comparés sont presque toujours imputés car lorsqu'on recourt à des relevés de prix, il n'est en général pas possible de comparer au cours d'une même période le prix du produit disparu et remplaçant. Par définition, on n'a pas anticipé que le produit allait disparaître et on n'a pas relevé le prix du produit remplaçant avant même de savoir que le produit remplacé allait disparaître. Puisque les données de caisses sont exhaustives, on peut y rechercher *a posteriori* le prix passé d'un produit.

<u>Document 4 :</u> Evolution des services Sirene, extrait du rapport de l'inspection générale de l'Insee « Conséquences pour l'Insee de la loi pour une République numérique » du 12 janvier 2017

Publié le 12 janvier 2017

[…]

2.1.3 De la nécessité de revoir l'offre de diffusion de Sirene

Les principaux éléments de contexte relatifs à Sirene peuvent être déclinés comme suit :

1- L'Insee va perdre les recettes liées à la diffusion de listes SIRENE et aux redevances qui y sont attachées versées par sa soixantaine de rediffuseurs, ses 400 abonnés et le millier d'utilisateurs ponctuels (chaque année) de listes d'établissements. La direction du Budget s'est engagée à compenser les pertes de recettes estimées à 11 millions d'euros en 2017 et les années suivantes.

2- Dans le même temps la demande ponctuelle adressée à Sirene-diffusion stagne, voire baisse alors que les rediffuseurs étoffent leur offre en réalisant des enrichissements des données achetées à l'Insee avec d'autres sources d'information.

3- L'offre de l'Insee s'appuie sur une partie de la division « grands comptes » qui assure également la maîtrise d'ouvrage déléguée du pôle « diffusion de l'information sirene » (DIS) placé à Nantes. Ce pôle a déjà fait l'objet de conjecture sur son maintien à Nantes, sur une intégration dans la division « Grands comptes », voire sur son déplacement dans une autre région16. L'intérêt spécifique porté à ce pôle tient à la situation de la DR de Nantes qui se caractérise par une démographie plus vieillissante que la moyenne. Par ailleurs, des motifs d'efficacité peuvent conduire à vouloir regrouper des activités proches, quitte à les placer dans d'autres établissements de l'Insee que la DG.

4- Dans le cadre de DLN1X (« Dites-le nous une fois »), l'Insee projette de créer une API-Sirène qui permettra à tout utilisateur ayant obtenu une autorisation de faire une interrogation sur la base Sirène pour en tirer tous les établissements qui l'intéressent avec l'ensemble des variables.

5- Dans le cadre de cette diffusion de données publiques, la mission Etalab rattachée au SGMAP, a vocation à assurer cette mise à disposition gratuite aux différents utilisateurs intéressés. Elle dispose d'un site data.gouv.fr qui fonctionne pour l'instant surtout comme un portail signalant l'offre publique d'information. Etalab envisage d'étoffer l'offre en accueillant de grosses bases de données qu'il se chargerait de mettre à disposition. Le sujet continue d'être discuté dans les cabinets des ministres concernés.

6- Du fait de la centralisation à Metz des moyens informatiques, les migrations en cours continuent de mobiliser les équipes du Secrétariat général informatique, rendant difficile toute mobilisation de moyens pour modifier sensiblement la configuration actuelle. Tout choix d'adaptation devra donc être économe en moyens informatiques.

2.2 L'offre loi pour la République numérique Sirene est à définir

La loi indique que la base Sirène doit être mise à disposition de tout utilisateur (public ou privé) de façon accessible avec des moyens informatiques. Le respect de la loi a minima consisterait à mettre sur un support informatique le fichier complet de Sirene.
Il convient de noter que c'est ce que permet aujourd'hui la chaîne de production des fichiers Syracuse, mais de manière payante. On ne peut envisager de simplement modifier la dimension payante en offrant simplement le service gratuitement, sans changer la nature de ce service.

Il y a quelques conséquences et limites à cette stratégie :

- La gratuité aura pour effet d'accroître fortement la demande faite à l'Insee. En 2015, au pôle DIS, il y a eu 2223 demandes ponctuelles de listes sur Internet dont 1371 ont fait l'objet d'un devis. Seules 208 d'entre elles se sont concrétisées par un paiement et par la fourniture de listes. Le taux apparent est alors de 15%. Si on s'intéresse aux commandes internet qui ont impliqué en 2015 une assistance d'un chargé de clientèle (CDC) : sur 107 demandes, 32 ont abouti à un paiement. Si on veut avoir un ordre d'idée du ratio entre l'intérêt pour les produits de sirene et l'abandon au moment du paiement, on trouve un ratio de 1 pour 6 dans un cas et de 1 pour 3 dans l'autre. L'ordre de grandeur de l'abandon des demandes du fait de la non-gratuité pourrait être de l'ordre de 3 demandes abandonnées pour 4. On imagine que dans ce cas, la gratuité devrait multiplier par un facteur 4 les demandes qui nous seraient faites avec une conséquence potentielle sur la charge de travail pour DIS et pour les capacités informatiques de diffusion de ces données.
- La base de diffusion Syracuse ne fournit pas toutes les données qui pourraient potentiellement être publiées. En particulier les données historiques sur chaque établissement qui sont actuellement demandées par la DSS pourraient être prises en compte dans le cadre de DLN1X.
- La loi Macron prévoit que les établissements fournissent une adresse e-mail qui pourrait servir de lien pour les administrations. Cette adresse pourrait s'ajouter à l'adresse physique dans le fichier Syracuse.

Le principe de gratuité impose à terme de revoir le contour de l'activité de diffusion : elle mobilise des moyens qui ne sont plus justifiés pour l'Insee par l'obtention de ressources financières additionnelles (même s'ils peuvent continuer de trouver une justification dans le cadre de la politique de diffusion).
Mais ce redimensionnement de l'offre, qui se traduira probablement par un désengagement par rapport à la situation actuelle, doit se faire progressivement. Sa vitesse dépendra, pour partie, de la localisation de la base publique que nous qualifions d'« offre loi pour la République numérique ». Il est en effet envisagé que ce soit Etalab qui puisse devenir responsable de la diffusion de cette offre loi pour la République numérique. Dans la suite, nous proposons de différencier l'offre loi pour la République numérique qui concerne la base sirene complète et la réponse à des demandes limitées de listes ou d'autres prestations de service.

Dans nos publics, nous distinguons les rediffuseurs, les abonnés à usage final, les clients à demandes ponctuelles, mais également les administrations. L'analyse précise des échanges avec les administrations n'a pas encore été conduite par la mission.
S'agissant de l'offre loi numérique, les administrations profiteront des principes d'ouverture et de gratuité pour tous qui les placent au même niveau que les rediffuseurs ou les abonnés à usage final. Elles auront accès aux mêmes informations dans les mêmes conditions. Pour le reste, il est suggéré de maintenir avec la sphère publique des relations étroites dans un premier temps (même si cela présente quelques coûts, ces relations pouvant concerner plusieurs unités : DDAR, DSE, projets).

* * *

Il convient tout d'abord de définir précisément l'offre loi pour la République numérique SIRENE universelle, gratuite et simplement accessible. L'offre actuelle sirene.fr (adossée à Syracuse) ne répond pas complètement à ces exigences. Il est suggéré de proposer en ligne une combinaison des fichiers stocks (mensuel) et des fichiers mouvements (quotidien et hebdomadaire) dans les formats longs actuels de Syracuse.

Recommandation S1 : Définir l'offre loi numérique au 1er janvier 2017. Ce devrait être l'offre maximale permise par la chaîne de production des fichiers Syracuse aujourd'hui, avec l'accès sur un site dédié (Etalab, Insee ou autre). Il s'agit pour l'essentiel des fichiers fournis aux rediffuseurs aujourd'hui : cette offre loi numérique exclut toute spécialisation (mise à jour sur profil, créations, sirénage). Les services web actuels d'aide à la sélection n'en font pas partie.

* * *

Pour réduire l'augmentation de la demande induite par la gratuité, il convient d'augmenter la taille des listes qui peuvent être obtenues de manière automatique et sans intervention d'un chargé de clientèle. Pour les demandes de fichiers de plus grosse taille, il convient d'adresser le demandeur à l'accès au fichier complet défini par l'offre loi pour la République numérique qu'il pourra se procurer sur le site où elle est placée : Insee ou Etalab. En 2015, sur 836 demandes de listes ponctuelles, il y en a eu 687 (82%) de moins de 10 000 lignes. Il y en a eu 126 entre 10 000 et 100 000 lignes (15%). Il n'y en a eu que 23 de plus de 100 000 lignes (dont 17 de plus de 1 000 000).

Recommandation S2 : Modifier Syracuse pour permettre aux usagers d'obtenir des fichiers d'une taille allant jusqu'à 100 000 lignes, en self service sans intervention des chargés de clientèle. Pour les demandes plus larges, le système les redirigera sur l'offre loi numérique.

* * *

Pour les rediffuseurs et les abonnés à usage final, il semble difficile de les mettre dès le 1er janvier 2017 en situation de devoir modifier toute leur organisation pour qu'ils se branchent sur l'offre loi pour la République numérique qui commencera à se mettre en place. Afin de les aider dans cette période de transition, le service actuel pourrait être maintenu (au seul prix de mise à disposition avec gratuité des données20), mais il faudrait par contre définir une date butoir du service actuel fourni pour qu'ils adaptent leurs systèmes d'information qui s'alimenteront sur l'offre loi pour la République numérique là où elle sera positionnée. Cette date butoir pourrait être le 30 août 2017, ou plus certainement, en termes d'opportunité, la date de basculement vers la solution API-Sirene. Les conditions qui pourront être faites dans ce cadre doivent leur être communiquées au plus vite ; idéalement en même temps que la dénonciation ou normalisation des conventions rendues nécessaires au 31 décembre 2016

Recommandation S3 : maintenir gratuitement le service actuellement rendu aux rediffuseurs et aux abonnés et les informer formellement dès à présent de son extinction d'ici la fin de l'année 2017.

* * *

2.3 Le positionnement de l'offre loi pour la République numérique : la place d'Etalab

La question de la place qu'aura Etalab dans la fonction de mise à disposition de bases de données est en discussion. Les arguments favorables à un transfert complet sont contrebalancés par ceux que l'Insee pourrait leur opposer. L'annexe 4 les explicite. Pour l'instant, avec Etalab, le SGMAP fournit un portail qui permet de valoriser les données publiques en permettant d'en avoir rapidement connaissance. Cette fonction peut devenir plus proactive en visant à rassembler les grands fichiers administratifs et en dotant Etalab de moyens d'interrogation performants s'appuyant sur des technologies développées notamment par des start-ups. S'il est décidé, dans le cadre de la politique d'Open Data, que la diffusion de l'offre loi pour la République numérique de Sirene doit être placée auprès d'Etalab qui aura été doté des moyens de gérer et diffuser la base, l'Insee devra mettre à disposition d'Etalab l'offre maximale qu'il fournit aujourd'hui aux rediffuseurs et ceci à un rythme journalier. Cette fourniture deviendra l'offre loi pour la République numérique. Devenu rediffuseur public de Sirene, Etalab serait alors alimenté quotidiennement comme le sont actuellement certains de nos rediffuseurs (échanges automatisés). La charge pour l'Insee serait donc d'un coût marginal, ce

qui va dans le sens d'un allègement des contraintes informatiques dans le contexte de forte mobilisation de moyens pour la migration des applications à Metz.

Encadré 1 : La mission « ETALAB »

Le décret n° 2011-194 du 21 février 2011 crée la mission Etalab, placée sous l'autorité du Premier ministre et d'abord rattachée au secrétaire général du Gouvernement.

Etalab a pour mission de créer le portail unique interministériel « data.gouv.fr » destiné à rassembler et à mettre à disposition librement l'ensemble des informations publiques de l'État, de ses établissements publics administratifs et, si elles le souhaitent, des collectivités territoriales et des personnes de droit public ou de droit privé chargées d'une mission de service public.

Elle coordonne en outre les actions des administrations de l'État et apporte son appui aux établissements publics administratifs afin de faciliter les réutilisations de leurs informations publiques.

Le portail data.gouv.fr poursuit les trois objectifs suivants :
– permettre la réutilisation des informations publiques la plus facile et la plus large possible ;
– encourager l'innovation par toute la communauté des développeurs et des entrepreneurs pour soutenir le
développement de l'économie numérique ;
– contribuer à renforcer la transparence de l'action de l'État, mettre en valeur le travail des administrations et éclairer le débat public.

L'arrêté du 31 octobre 2012 place la mission Etalab au sein du SGMAP créé par le décret n° 2012-1198 du 30 octobre 2012. La situation actuelle est définie par l'arrêté du 21 septembre 2015 portant organisation du SGMAP avec création de la direction interministérielle du numérique et du système d'information et de communication de l'État (DINSIC) à laquelle est rattachée la mission Etalab, à côté d'un service « performance des services numériques », d'une mission « incubateur de services numériques » et du service à compétence nationale « Réseau interministériel de l'État » (RIE). Sa mission est reprécisée dans l'article 3 :
La mission « Etalab » coordonne les actions des administrations de l'État et leur apporte son appui pour faciliter la diffusion et la réutilisation de leurs informations publiques. Elle contribue à leur conception et coordonne leur mise en œuvre interministérielle. Elle développe et anime le portail unique interministériel destiné à rassembler et à mettre à disposition librement l'ensemble des informations publiques de l'État, de ses établissements publics et, si elles le souhaitent, des collectivités territoriales et des personnes de droit public ou de droit privé chargées d'une mission de service public. Elle contribue, avec les administrations de l'État, à l'ouverture des données publiques et à la promotion des sciences des données.

Depuis 2014, le directeur de la Dinsic est aussi administrateur général des données. Il coordonne à ce titre l'action des administrations en matière d'inventaire, de gouvernance, de production, de circulation et d'exploitation des données par les administrations et organise, dans le respect de la protection des données personnelles et des secrets protégés par la loi, la meilleure exploitation de ces données et leur plus large circulation, notamment aux fins d'évaluation des politiques publiques, d'amélioration et de transparence de l'action publique et de stimulation de la recherche et de l'innovation.

[...]

<u>Document 5</u> : Rapport annuel de l'Autorité de la Statistique Publique (ASP) 2020

Paru en 2021

•Une intensification de l'exploitation de données privées et administratives pour éclairer la crise et au-delà

Outre les enquêtes nouvelles, le service statistique public a cherché à mobiliser autant que possible les sources privées ou administratives dont l'exploitation permettait d'enrichir la connaissance de l'impact de la crise et plus généralement de produire des statistiques infra-annuelles économiques ou sociales. La crise sanitaire a joué un rôle de catalyseur pour l'ouverture et l'exploitation par la statistique publique de données détenues par des sociétés privées d'énergie, de données de caisse, et de cartes bancaires. Les sources de données et les méthodes utilisées revêtaient un caractère expérimental, mais les capacités d'innovation développées au fil des ans à l'Insee ont permis d'en tirer rapidement le parti. Les points de conjoncture se sont en partie appuyés sur des exploitations nouvelles de fichiers détenus par des sociétés privées. Certains contacts entre ces sociétés et l'Insee étaient engagés depuis plusieurs mois avant la crise. Des partenariats inédits ont été rapidement noués, notamment avec le groupement d'intérêt économique «Cartes bancaires». Certaines de ces sources privées n'ont été mises à disposition du SSP que sur une durée limitée.

➢*Exploitation de données privées*

<u>Données d'électricité</u>

Au début de la crise du Covid, les données de consommation d'électricité, parce que disponibles dans un délai très court, ont permis d'approcher rapidement les baisses d'activité économique. Des contacts engagés dès mai 2020 entre l'Insee avec les opérateurs RTE et Enedis ont permis l'obtention de données plus adaptées au besoin que celles en open data mais avec un délai de transmission plus long.La comparaison entre consommation d'électricité et production (via l'indice de la production industrielle) ou chiffre d'affaires (via les données de TVA) devrait permettre d'identifier les branches pour lesquelles les variations de consommation d'électricité sont les plus liées aux variations d'activité économique.

<u>Données de caisse des enseignes de la grande distribution</u>

La loi pour une République numérique avait autorisé l'exploitation des données de caisse dans le cadre du calcul de l'indice des prix à la consommation. Avec l'accord de certaines enseignes, les données agrégées issues des données de caisse ont pu être utilisées pour l'exercice de « nowcasting », pour estimer l'évolution des volumes.
Des conventions ont été conclues avec 4 enseignes en vue de l'utilisation des données de caisse pour le calcul d'indices d'activité dans le commerce.

<u>Cartes bancaires</u>

Dès le 16 mars, le GIE Carte bleue a accepté de donner des données quotidiennes agrégées par département et secteur de magasins. Ces données se sont révélées assez fiables pour mesurer la consommation et l'activité dans les secteurs concernés. Le fait de disposer de ces indicateurs a permis de publier le chiffre de -35 % de baisse du PIB et aussi de consommation.L'analyse du potentiel des données de paiement par cartes bancaires pour le calcul d'indices précoces d'activité se poursuit mais le fait que l'Insee n'ait pas accès aux données individuelles de transactions freine l'exercice. Par ailleurs, l'Insee a fait part au GIE de son souhait de disposer de données (agrégées) par croisement de code d'activité et de département, ceci notamment à des fins de mobilisation de ces données pour le diagnostic conjoncturel régional.

Données de téléphonie mobile

Depuis le début de la crise sanitaire, les opérateurs de téléphonie mobile ont été fortement sollicités par l'action publique pour fournir des indicateurs utiles à la gestion de la crise. À ce titre et au-delà du projet de recherche MobiTic en cours avec Orange et financé par l'agence nationale de la recherche, l'Insee a contacté des opérateurs de téléphonie mobile pour leur demander des comptages anonymes afin de pouvoir renseigner sur la population présente et sa répartition sur le territoire et pour renseigner sur la chute de l'activité, telle que mesurée par les déplacements sur le territoire. Les trois opérateurs Orange, Bouygues Télécom et SFR ont répondu favorablement sous la condition que ces informations ne soient mobilisées que dans le cadre du suivi et de la gestion de la crise. L'Insee a ainsi pu proposer en avril 2020 des premiers résultats statistiques expérimentaux sur les déplacements générés par l'annonce de confinement et sur la nouvelle répartition de la population sur le territoire au niveau départemental, issus d'une collaboration avec Orange. L'analyse des mouvements de population à l'occasion du déconfinement a également été menée à partir des données des trois opérateurs de téléphonie mobile. Les discussions sont actuellement dans l'impasse avec les opérateurs, la cession d'indicateurs issus des offres commerciales ne convenant pas à l'Insee. Par ailleurs, la possibilité d'exploiter des données détaillées des opérateurs pour des finalités statistiques s'avère incertaine, notamment du fait des négociations européennes en cours sur la directive e-privacy.

➢***Exploitation de données administratives***

Mesures de soutien aux entreprises

A la fin de l'été 2020, l'Insee a engagé des démarches pour accéder à différentes données administratives permettant de documenter les mesures de soutien aux entreprises pendant la crise : notamment auprès de l'Acoss pour les reports de cotisations sociales, de la Dares pour l'activité partielle, de la DGFiP pour le fonds de solidarité. L'objectif est d'effectuer des études visant à examiner les effets de ces mesures de soutien, en particulier une étude conjointe Insee-Banque de France sur l'effet de ces aides sur les chroniques de trésorerie des entreprises. Ce sujet a vocation à alimenter les travaux du comité de suivi et d'évaluation des mesures économiques d'urgence présidé par M. Coeuré et placé auprès du Premier ministre. Concernant les aides régionales, l'absence de remontées d'information administrative centralisée empêche leur exploitation nationale.

1.3 Les principales avancées du SSP sans lien direct avec la crise

Comme il a été noté dans le chapitre précédent, l'investissement du SSP pour éclairer les conséquences de la crise sanitaire a été très soutenu, mais ceci n'a pas empêché le SSP de conduire en parallèle des travaux novateurs, constituant de réelles avancées.

•*À l'Insee*

Une première publication de l'indice des prix à la consommation (IPC) prenant en compte les données de caisse

Depuis la publication de l'IPC de janvier 2020, les données de caisses de la grande distribution sont utilisées pour le calcul de l'indice sur le champ des produits alimentaires industriels, des produits d'entretien et d'hygiène beauté en France métropolitaine. Près de 80 millions de produits sont suivis chaque mois sur ce champ contre 30 000 environ auparavant par des enquêteurs de l'Insee. Cela représente 1,7 milliard d'enregistrements par mois, transmis à l'Insee à un rythme quotidien, et exploités grâce aux technologies des « big data » (données massives)Le recours aux données de caisses a été précédé d'une phase expérimentale longue afin de s'assurer de l'obtention effective et sécurisée des données, de l'architecture informatique nécessaire pour traiter un tel volume de données et de la méthodologie pour exploiter ces données tout en restant à concept de l'IPC constant. En particulier, un double calcul a été réalisé tout au long de l'année 2019 afin de mesurer précisément l'impact du recours aux données de caisses et de s'assurer de la robustesse du processus de production.

<u>Document 6</u> : Extrait du compte rendu de la réunion du bureau du Conseil national de l'information statistique (Cnis)

Paru le 9 décembre 2020

APPARIEMENT DE SOURCES ADMINISTRATIVES EN STATISTIQUE PUBLIQUE : RAPPORT DE L'IG ET NOTE EXPLORATOIRE

Présentation du rapport de l'IG

Pascal RIVIERE, Institut National de la statistique et des études économiques (INSEE) - Inspection générale, indique que le rapport de l'inspection générale joue un rôle important dans une démarche d'ensemble. Il sera donc présenté en chapeau des autres présentations. Ce rapport de l'inspection générale, préparé par Renan Duthion et Michel Isnard, est né des difficultés rencontrées par les statisticiens pour apparier des fichiers en combinant, par exemple, des enquêtes et des sources administratives. Ces difficultés sont de deux ordres : ce sont des difficultés liées à l'offre de services et des difficultés résultant d'un manque de visibilité sur le cadre juridique.

A la suite d'échanges avec différents services et avec les SSM, le constat est fait de l'existence d'une forte demande d'appariement des SSM qui n'est aujourd'hui pas pleinement satisfaite. D'une part, pour mener à bien ces appariements, le cadre juridique n'est pas clair : un préalable est nécessaire pour définir une stratégie d'utilisation des identifiants. Pour être efficace, il est important d'utiliser des identifiants communs et de trouver un accord collectif sur une stratégie. D'autre part, l'organisation du service mérite d'être revue dans son ensemble car le travail sur les appariements s'effectue aujourd'hui au cas par cas avec des travaux menés en îlots. En outre, les demandeurs des appariements ne sont pas forcément assez explicites pour préciser leurs demandes afin que les appariements répondent aux attentes. Enfin, l'environnement méthodologique autour de l'appariement n'est pas assez structuré. En effet, il n'existe pas aujourd'hui une méthodologie commune ni un appui centralisé en direction des équipes qui réalisent l'appariement, alors que cette méthodologie commune existe dans d'autres domaines comme le tirage d'échantillons. En termes de méthode et d'outils, la présence de la France à l'international est très faible alors que les autres instituts nationaux de statistique (INS) sont extrêmement actifs. La littérature comme les sites Internet des autres INS montrent qu'il existe un très fort dynamisme sur cette matière en termes d'innovation méthodologique mais aussi en termes d'organisation d'ensemble de l'appariement.
En définitive, l'organisation actuelle consiste à rendre un service alors qu'il conviendrait de tendre vers une offre de services.

Après avoir posé ce constat, Pascal RIVIERE propose de soumettre quelques recommandations. Tout d'abord, le décret-cadre sur le NIR peut jouer un rôle central dans le préalable juridique. Il faut aussi préciser la stratégie d'utilisation des identifiants à mettre en place en indiquant quels identifiants utiliser, sur quelle durée, comment les conserver, etc. Sans stratégie identifiée, il faudra traiter les demandes au cas par cas. Enfin, il convient de mettre en place une offre de services aussi automatisée que possible. Cependant, un paramétrage ne permettra pas de proposer une offre standard, valable dans tous les cas, dès lors qu'il existera des spécificités selon les fichiers. Dans une logique d'appariement standard, la gratuité du service sera possible. Dans le cadre d'une offre de service automatisée, certaines fonctions pourront être proposées dans une logique de libre-service. Ces fonctions sont décrites en détail dans le rapport. Ce sont les fonctions d'identification, de hachage, de couplage, etc. Un rééquilibrage des rôles entre le demandeur et le détenteur doit être envisagé puisque le demandeur dispose d'une expertise sur les caractéristiques des fichiers qu'il apporte mais aussi sur son domaine métier. Ces connaissances doivent être prises en compte pour être plus efficace dans la démarche d'appariement.

Pascal RIVIERE ajoute qu'il apparaît nécessaire de mettre en place progressivement une culture d'appariement en créant un réseau d'experts en méthodologie et par type de domaines métiers. Cette culture pourra essaimer via un jeu de questions réponses et un travail sur le contenu des fichiers. Cette approche nécessite des développements méthodologiques afin de fixer les modalités de paramétrage et de déterminer des attendus sur la qualité des résultats.

Il est cependant impossible d'attendre la fin du projet avant de proposer une première offre de services. Il convient donc de mettre en place des mesures transitoires. Ces mesures transitoires peuvent passer par des programmes expérimentaux. Il est aussi possible de diffuser la culture d'appariement lors d'événements comme des séminaires. L'identification statistique doit aussi jouer son rôle. Aujourd'hui, il existe un algorithme d'identification au RNIPP qui permet de repérer dans le référentiel, à partir d'un nom, d'un prénom et d'une date de naissance, la personne qui correspond à la demande pour obtenir le NIR. Pour réduire tout risque, l'algorithme n'aboutit à une identification que s'il existe une très grande proximité entre les données d'entrée et les données du référentiel. Dans une logique d'identification à des fins statistiques, il est toutefois possible d'accepter de légers écarts. Avec un algorithme moins exigeant, il sera possible de parvenir à une efficacité plus importante de l'appariement.

Le rapport de l'Inspection générale aboutit à la conclusion qu'il existe un déficit en termes d'offre de services. Pour combler ce manque, il est proposé de mettre en place une offre de services automatisée et un soutien plus important sur le plan méthodologique. Avant de parvenir à la cible, des mesures transitoires pourront être envisagées notamment pour les fonctions d'identification statistique, de hachage, de couplage, d'extraction/transmission et d'expertise. C'est tout l'objet de la présentation à suivre de Christel COLIN portant sur un projet visant à éviter un effet tunnel en attendant que le projet dans son entièreté soit livré.

Présentation du CSNS

Christel COLIN, Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS), souligne que la mission de l'inspection générale sur les appariements recommande de développer une offre de services, laquelle a vocation à s'appuyer notamment sur un projet en cours à l'Insee et qui porte sur le code statistique non signifiant (CSNS).

Pour rappel, le projet CSNS trouve son origine dans des évolutions juridiques intervenues depuis 2016. Son point de départ est la loi n°2016-1321 du 7 octobre pour une république numérique qui introduit la notion de code statistique non signifiant dans son article 34. Ce code statistique non signifiant est obtenu par une opération cryptographique appliquée au NIR. Il s'agit d'un code statistique pour le service statistique public qui vise à simplifier les démarches auprès de la Cnil. En effet, cette disposition vise à ce qu'il n'y ait plus besoin de décret en Conseil d'Etat pour les traitements ayant uniquement des finalités de statistique publique et ne comportant pas de données sensibles, lorsque le NIR est préalablement crypté et lorsqu'il lui est substitué un code statistique non signifiant.

Cet objectif de simplification des démarches s'est concrétisé dans le décret n°2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche. Ce décret constitue le cadre des dispositions de mise en oeuvre du CSNS. Il précise notamment que cette opération cryptographique est mise en oeuvre par un service de l'Insee.

Ces dispositions sont complétées par le décret-cadre NIR de 2019 et par l'arrêté du 28 septembre 2020 pris en application du décret n°2016-1930 du 28 décembre 2016, notamment pour préciser les modalités de conservation du CSNS.

Le projet CSNS en cours à l'Insee vise à mettre en oeuvre ces dispositions juridiques et à développer une offre de services automatisée pour faciliter les appariements de données individuelles au sein du service statistique public dans l'objectif de répondre à une demande sociale croissante, notamment sur les questions d'évaluation des politiques publiques ou pour les analyses longitudinales, et ceci en limitant la charge d'enquête directe par l'intermédiaire d'un recours accru aux données administratives. C'est un projet qui fait l'objet d'attentes fortes, notamment de la part des services statistiques ministériels, et tout particulièrement des grands SSM sociaux que sont la Dares et la Drees qui sont destinataires de nombreuses sources administratives et qui expriment des besoins importants en matière d'appariement.

Christel COLIN ajoute qu'il est également envisagé d'insérer progressivement le code statistique non signifiant dans des fichiers dits « pivot », c'est-à-dire dans des fichiers de référence qui contiennent des variables potentiellement utiles à de multiples appariements. Le CSNS pourrait par exemple être intégré dans

Fideli, la DSN ou encore le recensement de la population, ce qui permettrait d'assurer ensuite un fort potentiel d'appariements.

Le projet CSNS comprend aussi un volet ayant trait à l'organisation et à l'échange de données puisqu'il a aussi pour objectif de proposer une organisation visant à faciliter la diffusion du CSNS au sein du service statistique public de façon sécurisée, ce qui est essentiel lorsque l'on parle d'identifiants.

Le projet vise à développer un service pour l'ensemble du service statistique public en délivrant un code statistique non signifiant. Ce code sera délivré par un service de l'Insee à un responsable de traitement (service statistique ministériel ou autre unité de l'Insee). Le responsable de traitement aura en amont fourni soit un NIR, soit des éléments d'état civil permettant de déterminer le NIR et d'opérer le passage au code statistique non signifiant. L'objectif poursuivi est de fournir un service pérenne. Le renouvellement du CSNS devra avoir lieu a minima tous les dix ans, d'après les textes, mais pourra avoir lieu plus fréquemment, en particulier s'il est identifié une alerte ou des risques sur l'intégrité des données ou leur sécurité.

Concrètement, le CSNS est obtenu par un hachage du NIR. Puis, est appliquée une clé secrète qui permet de transformer le NIR en un identifiant non porteur d'informations, et donc non signifiant. Si le responsable de traitement ne transmet pas le NIR, parce qu'il n'en dispose pas, mais qu'il transmet des traits d'identité, c'est-à-dire des éléments d'état civil, une étape préalable d'identification au Répertoire national d'identification des personnes physiques (RNIPP) pour retrouver le NIR sera nécessaire, avant d'exécuter l'opération de hachage et de chiffrement.

Le projet a démarré en 2019. Une première version sera disponible fin 2020 et permettra d'obtenir un code statistique non signifiant à partir du NIR. Une deuxième version sera disponible fin 2021 et permettra de réaliser les étapes à franchir lorsque le NIR n'est pas disponible et qu'il faut passer par une étape d'identification statistique préalable. Enfin, la version finale du projet est prévue fin 2022 avec des compléments sur le moteur d'identification et une interface d'utilisation pour les responsables de traitement.

Christel COLIN ajoute que ce projet doit aussi conduire à définir une gouvernance afin que les traitements qui mobilisent les appariements se réalisent dans la transparence, afin de garantir l'information des personnes concernées et la justification de la finalité des traitements. Ce préalable vaut en particulier pour les appariements mobilisant le code statistique non signifiant. A l'instar de ce qui est fait pour les enquêtes, il est proposé que le Cnis soit conduit à formuler un avis préalable sur les demandes de traitements d'appariements formulées par le service statistique public. En particulier, les traitements impliquant le CSNS à des fins de croisement de données feraient l'objet d'un avis préalable du Cnis conditionnant l'accès au service rendu par l'Insee. Dans la même logique de transparence, le Cnis pourrait également être consulté sur les traitements relatifs à l'insertion du CSNS dans des fichiers pivots.

Document 7: Glossaire

Acoss : Agence Centrale des Organismes de Sécurité Sociale

API : les API (Application Programming Interface – ou Interface de programmation d'applications en français.) se définissent comme un ensemble de fonctions informatiques par lesquelles deux logiciels vont interagir sans intermédiation humaine. C'est ainsi un mode de diffusion adapté aux utilisateurs avancés souhaitant exploiter les données de manière automatique

Cnil : Commission nationale de l'informatique et des libertés

Cnis : Conseil national de l'information statistique

Coicop : Classification des fonctions de consommation des ménages (Classification of Individual Consumption by Purpose - COICOP)

CSNS : Code Statistique Non Signifiant

Dinsic : Direction interministérielle du numérique et du système d'information et de communication de l'État (Dinsic) – devenue depuis octobre 2019 La direction interministérielle du numérique (Dinum) est en charge de la transformation numérique de l'État

Directive e-privacy : la directive « e-privacy » établit les règles relatives à la protection des libertés et droits fondamentaux des personnes physiques et morales en ce qui concerne la fourniture et l'utilisation de services de communications électroniques, pour les utilisateurs finaux qui se trouvent dans l'Union européenne.

DSN : Déclaration Sociale Nominative

EPA : Etablissement Public à caractère Administratif

GIE : Groupement d'Intérêt Economique

IPC : Indice des Prix à la Consommation

NIR : Numéro d'Identification au Répertoire

RNIPP : Répertoire National d'Identification des Personnes Physiques

RTE : Réseau de Transport d'Électricité

SGMAP : Secrétariat général pour la modernisation de l'action publique

SSP : Service Statistique Public

Syracuse : Système pour la rediffusion, l'accès et la commercialisation vers les utilisateurs de Sirene