N°9

COURRIER DES STATISTIQUES

Juin 2023

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE

Rédaction en chef
Catherine Fresson-Martinez
Contribution
Insee : Mathias André, Yves-Laurent Bénichou, Franck Cotton, Alexis Dondon, Lionel Espinasse, Jean-Marc Germain, Séverine Gilles, Olivier Haag, Pierre Lamarche, Patrick Redor, Michaël Sicsic
Cnav : Bertrand Dubrulle, Olivier Rosec, Christian Sureau
DGAFP : Gaël de Peretti
Université de Lille : Béatrice Touchelay
Directeur de la publication
Jean-Luc Tavernier
Directeur de la collection
Pascal Rivière

Rédaction
Catherine Fresson-Martinez, Pierre Glénat, Marine Le Roux, Pascal Rivière
Composition
Agence LATITUDE Nantes
5, rue Jacques Brel
« Les Reflets » Bâtiment A
44800 SAINT-HERBLAIN
0190/23
02 51 25 06 06
www.agence-latitude.fr
Photo de couverture
Adobe Stock®

Éditeur
Institut national de la statistique et des études économiques
88, avenue Verdier
92541 MONTROUGE CEDEX

www.insee.fr

Courrier des statistiques N9

SOMMAIRE

PRÉSENTATION DU NUMÉRO

Chaque numéro du Courrier des statistiques obéit à sa propre logique. Certes, le hasard des propositions d'articles ne permet pas d'aboutir spontanément à une homogénéité de thèmes, et la logique du numéro n'apparaît qu'*a posteriori*. Le numéro N9 est caractérisé par le choix d'articles portant sur des sujets parfois ardus, et inhabituels pour la revue. Nous allons donc tenter ici de faciliter leur compréhension.

Une fois n'est pas coutume, commençons par la fin, avec les trois derniers articles, qui portent sur des sujets liés, et qu'on pourrait à tort considérer comme plus adaptés à une revue d'informatique que de statistique publique. Ils sont en réalité essentiels dans un « monde de *data* », où statisticiennes et statisticiens vont de plus en plus puiser des données externes, administratives par exemple, pour leur propre usage[1] .

Ainsi l'article n°5, écrit par ***Alexis Dondon et Pierre Lamarche***, porte sur les formats de données. En première approche, on pourrait considérer qu'il s'agit d'un sujet annexe, d'une dimension purement opératoire, secondaire pour l'usage statistique. Il n'en est rien : qu'ils soient imposés à l'utilisateur ou au contraire délibérément choisis, les formats ont des propriétés, présentent des limites, des opportunités. Les auteurs expliquent qu'il n'existe pas de format idéal, mais au contraire que chaque format répond à une gamme de besoins, de contraintes. Ils présentent le format Parquet, moins connu, plus récent, adapté à de très gros volumes de données.

Ces données, on se les procure « ailleurs », dans des sources administratives. Il y aurait beaucoup à dire sur cette notion de source, mais partons de ce qui existe et voyons comment on élabore des statistiques à partir de cela. Il faut intégrer les données puis les transformer pour les rendre aptes à être digérées par un processus de production statistique. C'est cette phase méconnue de transformation que décrivent ***Franck Cotton et Olivier Haag*** dans l'article n° 6. Ils en décomposent les étapes, du recodage à la pseudonymisation en passant par le contrôle, du renommage au filtrage en passant par la caractérisation des unités statistiques. Ils insistent également sur la nécessité d'en faire un traitement automatisé et réplicable, un véritable *pipeline* piloté par les métadonnées. La gestion des formats est ici un des aspects importants du processus de transformation et de contrôle.

Vous avez dit contrôle, transformation ? ***Bertrand Dubrulle, Olivier Rosec et Christian Sureau*** (Cnav) s'y intéressent dans l'article n°7, mais dans un contexte très différent : les échanges massifs de données au sein de la protection sociale, par exemple pour l'alimentation de référentiels, comme le Répertoire de gestion de carrière unique (RGCU), ou pour les déclarations administratives. Pour maîtriser les flux de données transmis, et dans une optique de traitements automatisés, la structure attendue des données et les règles qu'elles respectent doivent être très clairement définies : c'est ce qu'on appelle une *norme d'échange*. Dans un contexte où cette structure peut évoluer fréquemment en raison des changements de réglementation, la Cnav a mis au point un outil (Saturne) qui permet de décrire formellement une norme et de générer automatiquement, sur cette base, toute la documentation associée et les outils de contrôle. Une telle démarche est particulièrement pertinente pour assurer la qualité des données, sujet essentiel pour les statisticiens.

1 Voir le dossier « Le statisticien et les sources administratives » du numéro N1 de décembre 2018.

Remontons d'un cran dans le sommaire, avec deux articles (les numéros 3 et 4) portant sur la question de la confidentialité des données et la façon de gérer efficacement cette confidentialité.

L'article n° 3 de ***Patrick Redor*** fournit le cadre d'analyse, en posant la confidentialité des données comme un enjeu majeur de la statistique publique, en raison des risques encourus en cas de violation de cette confidentialité. Mais pour l'activité statistique, les éléments d'identification des personnes sont le plus souvent indispensables, et on ne peut se borner à les enlever. Il faut donc des mesures de protection et un cadre juridique, qui s'est enrichi dans le temps : loi de 1951, loi Informatique et Libertés, loi pour une République numérique, règlement général sur la protection des données (RGPD), *Data Act*. Les règles du secret statistique, si elles ne figurent pas dans la loi, se révèlent subtiles dans leur application, avec notamment le « secret secondaire ». Tout ceci s'inscrit dans un contexte évolutif, où la demande de données ne cesse de croître, et l'on peut parfois se demander si confidentialité et *open data* ne sont pas deux injonctions contradictoires. Promouvoir un large accès aux données peut avoir pour conséquence paradoxale de… restreindre l'accès à certaines statistiques.

Pour assurer la protection des données confidentielles tout en profitant de la richesse des données provenant de sources différentes, il existe une possibilité : s'appuyer sur un « code statistique non signifiant » (CSNS). Ainsi, dans chaque fichier à apparier, on enlèvera les éléments d'identification, en ne conservant que ce code. Celui-ci servira de pivot pour l'appariement, tout en ne permettant pas de remonter à l'individu. Comme l'expliquent ***Yves-Laurent Bénichou, Lionel Espinasse et Séverine Gilles*** dans l'article n°4, le CSNS, plus qu'un code, est un véritable « service » rendu par l'Insee à l'ensemble du service statistique public. Il peut s'appliquer à des NIR (numéro de sécurité sociale), auquel cas l'opération est un pur chiffrement, ou à des traits d'identité (nom, prénom, date et lieu de naissance). Dans le second cas, un algorithme préalable d'identification, i.e. la détermination du NIR à partir des traits, est nécessaire. L'article explicite les différentes étapes de cet algorithme et la mesure de la qualité de l'identification. Celle-ci est indispensable, car la procédure CSNS est entièrement automatisée : disposant des niveaux de qualité, l'utilisateur décidera des seuils à appliquer.

Notre cheminement dans le sommaire nous conduit à l'article n°2, portant sur un sujet sophistiqué, multiforme et en même temps novateur : les comptes nationaux redistribués. ***Mathias André, Jean-Marc Germain et Michaël Sicsic*** en expliquent de façon synthétique les tenants et les aboutissants, ce qui constitue un véritable défi pédagogique. Il s'agit, dans un premier temps, de se replacer dans le cadre « classique » de la comptabilité nationale (certes dans une vision simplifiée)… pour aussitôt faire un pas de côté en posant des questions nouvelles et en inventant pour cela un nouveau cadre, centré sur les ménages. De façon directe ou (très) indirecte, les ménages sont destinataires finaux des revenus et des transferts des autres secteurs institutionnels, par exemple à travers les services rendus par les administrations publiques (santé, éducation, etc.). Ce cadre permet ainsi de construire un revenu « avant transferts » et « après transferts ». On s'intéresse alors aux mécanismes

de redistribution (élargie, complémentaire à l'approche usuelle monétaire), selon le niveau de vie, par cohorte d'âge, par catégorie socio-professionnelle, réconciliant ainsi comptabilité nationale et statistiques sociales. Les auteurs explicitent la genèse de cette démarche, les sources utilisées, la méthode de calcul et en détaillent les principales hypothèses. L'article propose quelques enseignements à tirer et trace des perspectives opérationnelles pour cette méthode internationalement reconnue et promise à un bel avenir.

Ce n'est pas d'avenir, mais de passé dont il est question dans l'article n° 1. ***Gaël de Peretti et Béatrice Touchelay*** nous racontent une histoire, celle de la statistique publique dans les 40 années qui ont suivi la création de l'Insee, sous l'angle de son insertion dans le débat social et politique. Dans une première période, dite de construction, on va poser des bases qui vont structurer le fonctionnement de l'institut : les enquêtes « ménages », avec le souci d'étudier les conditions de vie des ménages, le cadre de la comptabilité nationale, la loi de 1951, la coordination statistique, dans un contexte où l'intérêt de la statistique publique est loin d'être acquis. Apparaissent très tôt des controverses, par exemple sur l'indice des prix. Mais c'est encore un auditoire limité, une institution au service d'un petit nombre de décideurs. Dans la deuxième période, dite de consolidation, de nouveaux publics, de nouvelles ouvertures apparaissent dès les années 60 : la création des observatoires économiques régionaux, la création du conseil national de la statistique qui deviendra plus tard le conseil national de l'information statistique (Cnis). C'est aussi l'ouverture vers le grand public, avec la création d'un département de la diffusion, et une reconnaissance croissante *via* plusieurs succès éditoriaux. Entre-temps, on est passé de la mécanographie à l'informatique.

La suite de cette aventure à lire dans un prochain numéro de la revue, probablement en 2024 !

Pascal Rivière
Directeur de la collection, Insee

Statistiques publiques et débat démocratique : de la création à la consolidation (1946-1987)

Gaël de Peretti et Béatrice Touchelay* **

Le développement et la diffusion de l'information statistique est au cœur des préoccupations du Conseil national de la Résistance pour instaurer une démocratie économique et sociale. Au travers d'une esquisse historique des quarante premières années de fonctionnement de l'Insee, deux périodes se dessinent. Dans la première, qualifiée de construction, l'appareil statistique s'étoffe et assoit sa position. L'institution, reconnue et écoutée, s'affirme mais l'essentiel de ses regards sont portés vers les décideurs politiques et économiques. Il s'agit avant tout de répondre aux besoins de la reconstruction. Dans un deuxième temps, nommé "consolidation par l'ouverture", il s'agit, tout en s'appuyant sur les travaux déjà réalisés, de consolider ces acquis mais surtout de s'ouvrir aux autres utilisateurs, que ce soit au niveau local en lien avec les différentes vagues de décentralisation, au niveau national avec la mise en place d'instances de concertation, et plus généralement à destination du grand public à travers une politique offensive de diffusion de l'information statistique.

The development and dissemination of statistical information is a major concern of the "Conseil national de la Résistance" to establish economic and social democracy. Through a historical sketch of the first forty years of INSEE, two periods emerge. In the first period, described as construction, the official statistic system expands and establishes its position. It is a recognized and listened institution that asserts itself but whose main eyes are focused on political and economic decision-makers. This is primarily to meet rebuilding needs. In a second step called consolidation through openness, it is a question, while relying on the work already carried out, of consolidating these achievements but above all of opening up to other users, whether at the local level in connection with the different waves of decentralization, at national level with the establishment of consultation bodies, and more generally aimed at the general public through a volontarist policy of disseminating statistical information.

* *Sous-directeur, sous-direction des études, des statistiques et des systèmes d'information, DGAFP*
gael.de-peretti@finances.gouv.fr

** *Professeure d'histoire contemporaine, Université de Lille*
beatrice.touchelay@univ-lille.fr

Le programme du Conseil national de la Résistance (CNR) du 15 mars 1944 définit les réformes à mener dès la libération du territoire pour reconstruire sur des bases nouvelles et instaurer une démocratie économique et sociale[1]. Il fait du développement de l'information et de sa diffusion l'un des piliers de la reconstruction et il confie cette mission à l'État.

La création de l'Institut national des statistiques et des études économiques (Insee) pour la France et l'Outre-mer par la loi de finance d'avril 1946 doit répondre à cet impératif. L'Institut est doté d'une double fonction, technique et d'étude, qui fait son originalité parmi ses homologues européens focalisés sur la seule production statistique *(Desrosières, 1989)*. Doté de sa propre école d'application, il est chargé de produire des données pour éclairer les décisions et de fournir des études économiques. Son nom est discuté, il est question un temps d'ajouter *« et de la documentation »* à un intitulé déjà bien long pour confirmer que l'Institut est au service du public[2].

En suivant l'évolution de l'Insee voire du service statistique public, depuis sa création jusqu'au milieu des années 1980 ***(figure)***, il s'agit de préciser dans quelle mesure cet appareil statistique répond à cette mission de production et de diffusion de l'information statistique et nourrit le débat social et politique. Ces quarante premières années sont découpées en deux périodes au cours desquelles l'auditoire des travaux de l'Insee et sa contribution aux débats démocratiques s'étoffent. Les frontières sont évidemment floues et il n'existe pas dans cette histoire de rupture aussi abrupte. Comme Amossé *et al.* l'écrivaient dans leur analyse historique sur les hommes et les femmes en ménage statistique : *« Aussi, notre article se présente davantage comme l'esquisse d'une lecture historique proposée à la discussion qu'un travail véritable d'historien des statistiques. »*

▶ Période I : construction

La première période correspond aux mandats de Francis-Louis Closon (1946-1961) et de Claude Gruson (1962-1967). Ils fondent une institution qui s'éloigne progressivement du projet du CNR en étant d'abord au service de l'État, en contradiction sans doute avec les aspirations des jeunes recrues de l'institution issues de l'école d'application de l'Insee *(Brunaud et al., 2020)*, mais en concordance avec ses moyens humains et budgétaires, avec les capacités moyennes d'appréhension des statistiques et avec les conceptions du débat démocratique de la majorité des gouvernants *(Porter, 1995)*.

1 Sur la période de la reconstruction et des transformations de l'État, quelques références : Chapman H (2021), La longue reconstruction de la France. À la recherche de la république moderne. Presses de Sciences Po, « Académique » ; Andrieu C, Le Van L, Prost A (1987), Les Nationalisations de la Libération. De l'utopie au compromis, Presses de Sciences Po, 1re éd. ; Margairaz M (2017), L'État, les finances et l'économie : histoire d'une conversion, 1932-1952, vol. 1 et 2, Paris, Comité pour l'histoire économique et financière de la France *(lire en ligne [archive])*.

2 Service des archives économiques et financières (SAEF) Savigny-le-Temple, H 1573, n°299/C, 10 avril 1946, lettre de Closon à Braconnot, directeur régional Alger : *« La loi qui doit réorganiser les services du MIN et créer, notamment, un Institut National des Statistiques, des Études Économiques et de la Documentation, prévoit (…) qu'une coordination très étroite soit réalisée (…) entre la Métropole et les TOM »*.

▶ Figure - Fresque chronologique : l'Insee de 1946 à 1987

1946

- 27 avril : Création de l'Insee. Exploitation du premier recensement démographique
- 14 juin : Décret fixant les attributions de l'Insee
- 28 août : Attribution de la gestion du fichier électoral

François-Louis Closon

1951 — **7 juin :** Loi sur le secret et l'obligation statistiques, création du Comité d'organisation et de coordination des enquêtes COCOES

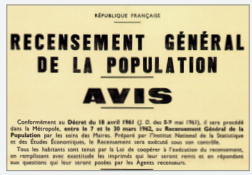
1952 — **15 septembre : décret d'application de la loi du 7 juin 1951**

Logo de l'Insee en vigueur jusqu'en 1969

1961 — L'Insee est chargé des comptes nationaux

Claude Gruson

1962 — **Du 7 au 30 mars,** recensement général de la population



1965 — **28 octobre :** Décret permettant l'utilisation des données fiscales

1967 — **Juin :** Inauguration du premier Observatoire économique régional (OER)

Jean Ripert

1969 — **Mai :** Réformes des publications nationales de l'Insee avec la création d'*Économie et statistique*, *Tendance de la conjoncture, Annales, Informations rapides*

1972 — Création du Conseil national de la statistique (CNS) qui prend le nom du Conseil national de l'information statistique (Cnis) en 1984

1973 — Publication du premier numéro de *Données sociales*
Projet de Système automatisé pour les fichiers administratifs et répertoires des individus

1974

Edmond Malinvaud

1975 — Déménagement de la Direction générale Porte de Vanves à Malakoff

1978 — **6 Janvier :** Loi Informatique et libertés

1984 — **17 juillet :** Conseil national de l'information statistique (Cnis)

1987 — Création d'un vidéotex national grand public : 3615 Insee

▶ Une institution au service d'une minorité de décideurs

Le titre premier du décret de juin 1946 définit ses attributions. L'Insee est chargé *« d'établir, de rassembler et de mettre à jour les statistiques relatives à l'État et au mouvement des personnes et des biens en utilisant, le cas échéant, les éléments qui lui sont fournis par les diverses administrations »* (alinéa 1), de *« donner et tenir à jour l'inventaire permanent de l'économie »* (alinéa 3). Il *« procède, pour le compte des administrations et organismes visés au 2° du présent article, à l'exécution des recensements approximatifs »* (alinéa 9), *« coordonne les méthodes, les moyens et les travaux statistiques des administrations publiques et des organismes privés subventionnés ou contrôlés par l'État, [...] centralise leur documentation statistique et économique et [...] réalise l'unification des nomenclatures et des codes statistiques »*. Cette fonction coordonnatrice et centralisatrice n'étant définie par aucun règlement, l'Insee tarde à s'imposer. Au service des gouvernements, de l'ensemble des administrations, et de *« toutes les personnes physiques ou morales de droit privé »* qui en feraient la demande, le périmètre de ses actions est très large. Il recouvre tous les domaines *« des statistiques et des études économiques », « assure la formation du personnel spécialisé »* et *« observe l'évolution de la situation économique dans la métropole, la France d'Outre-mer et l'étranger »* (alinéa 4).

Ces ambitions initiales deviennent irréalistes avec l'échec du Grand ministère de l'Économie nationale porté par Mendès France et le rattachement de l'Insee au sous-secrétariat d'État à l'Économie nationale, puis au secrétariat d'État aux Affaires économiques qui ouvre une période de vaches maigres jusqu'en 1961. Le rattachement au ministère des Finances apporte ensuite à l'Insee les moyens budgétaires de ses ambitions initiales *(Closon, 1971)*. Ce rattachement sera aussi l'occasion d'une intégration partielle du Service des études économiques et financières (SEEF) et donc de l'arrivée des comptables nationaux à l'Insee. Comme l'écrira Closon quelques années après son départ de l'Insee : « *L'essentiel de l'édifice était bâti et la statistique, dans son acception la plus large, avait acquis droit de cité. Il manquait encore un éclat et un complément nécessaire, celui que pouvait apporter le calcul économique, et une participation plus active aux comptes de la Nation, un lien plus étroit avec la réflexion du ministère des Finances et du Plan, la lente montée de nouvelles équipes. La tâche n'était pas terminée en 1961, mais suffisamment engagée pour que la voie fût non seulement indiquée, mais ouverte. La période du risque vital était franchie ».*

▶ Un auditoire volontairement limité : la *Note verte*

Inspiré par la revue Nord-américaine *Fortunes* et disposant d'une équipe de conjoncturistes héritée du Service national des statistiques, Closon décide de publier une note de conjoncture intitulé la *Note verte* qui serait envoyée à une centaine de décideurs triés sur le volet *(Touchelay, 1993)*. Les questions d'actualité, telles que le *ralentissement de l'activité économique* ou les *mécanismes de ce que l'on a appelé la relance* en 1952, font l'objet de réflexion commune des spécialistes qui diffusent leurs comptes rendus dans ce document confidentiel. À partir de 1953, elle est dépouillée de tout commentaire des statistiques qui sont intégrés au *Bulletin mensuel de statistiques.* Elle ne contient plus que des observations générales et devient vraiment confidentielle puisque son tirage est limité à 150 exemplaires. La liste des destinataires est établie par Closon en accord avec le ministre de tutelle. La *Note* est le domaine réservé de Closon, particulièrement exigeant à l'égard des auteurs.

Ses exigences se manifestent dans des directives relativement sèches comme celle de 1957 qui critique le manque de synthèse des études et qui souligne que *« tout doit être bon car l'Institut ne peut pas se payer le luxe de sortir des papiers médiocres »*. Closon considère qu'une *« trop grande diffusion aurait pour lamentable résultat une diminution très importante de notre liberté d'expression »*, ce qui explique qu'il contrôle très strictement la liste des destinataires. En octobre 1952, il refuse la demande d'abonnement de la Chambre de commerce de Mulhouse en précisant que cette publication n'est *« diffusée qu'à titre gratuit et qu'elle est destinée particulièrement aux membres du gouvernement »*. Entre avril et octobre 1955, le tirage de la *Note* augmente de cent exemplaires ***(Tableau)***. Closon souligne à plusieurs reprises que ce volume risque d'atténuer son caractère confidentiel et il lui arrive fréquemment de refuser de la communiquer.

▶ Tableau : Diffusion de la note verte de 1955 à début 1957

1955		1956		1957	
Janvier	405	Janvier	550	Février	510
Avril	386	Juin	520		
Octobre	500	Décembre	520		

Le succès de la *Note* est international puisqu'un correspondant du *Sunday Times* à Londres réclame qu'elle lui soit communiquée régulièrement à partir d'août 1959. Il devra se contenter de consulter *Étude et conjoncture* quatre fois par an à la suite du refus de Closon. Un mois plus tard, le directeur des finances et trésorier de Esso Standard SA France, se heurte au même refus. Lorsque les destinataires font un usage du document qui ne correspond pas à son caractère confidentiel, son envoi est interrompu. Closon suspend l'envoi à Jean Monnet à la CECA[3] à Luxembourg après deux années, à la suite de sa découverte dans la bibliothèque de la Haute Autorité en 1954. Il n'envisage de revenir sur sa décision que si Monnet confirme que *« ce document sera traité comme il doit l'être »*. En revanche, l'accès accordé à Pierre Locardel du *Figaro* est définitivement stoppé à partir du moment où celui-ci utilise des renseignements contenus dans la *Note* pour un article sur les ventes d'automobiles neuves publié en juillet 1954.

Grâce à la *Note verte*, l'Institut confirme sa fonction d'observateur de la conjoncture au service d'une poignée de décideurs et conforte sa réputation. Le succès du document révèle l'existence d'un besoin réel d'informations quantifiées. Compte tenu des moyens techniques disponibles, il représente une très lourde tâche, rédaction, impression et diffusion par voie postale, pour une institution peu dotée et qui fait des choix.

3 CECA : la Communauté européenne du charbon et de l'acier était une organisation internationale fondée sur le traité de Paris, entré en vigueur le 23 juillet 1952 pour une durée de cinquante ans. Elle n'existe plus depuis le 23 juillet 2002.

Une institution en prise avec les débats démocratiques : la question des fichiers

En août 1946, la loi confie à l'Insee la gestion du fichier électoral. La question des fichiers n'est pas réglée pour autant puisqu'un inspecteur général est chargé d'étudier tous ses aspects en janvier 1947. En avril, une instruction de la direction générale précise aux directeurs régionaux que *« le principe du fichier démographique n'est pas condamné mais qu'il n'y a pas lieu d'aller plus loin pour le moment »*. En revanche, *« le fichier des établissements est non seulement maintenu, mais c'est sur lui que portera le premier effort dans le domaine des fichiers » (Chevry, 1948)*. Le mois suivant, dans un rapport adressé au ministre de l'Économie nationale, Closon suggère de *« conserver les fichiers en limitant au maximum les renseignements portant sur les personnes »*[4]. La politique de l'Insee en la matière est ainsi définie clairement. Cela ne suffit pas à désamorcer les critiques. Dans un article publié par la *Revue Défense Nationale* en février 1947, Alfred Sauvy, entré à la Statistique générale de la France en 1922, nommé directeur du nouvel Ined[5] à la Libération, s'en prend violemment aux *« instruments d'oppression »* que pourraient constituer les fichiers de l'Insee. La réponse de Closon clôt la polémique[6] ***(encadré « Quarante ans de controverse »)***. Reconnaissant qu'il s'agit d'un « problème très délicat » et précisant qu'il a *« mûrement réfléchi à l'instrument d'oppression que pouvaient constituer ces appareils »*, le directeur général précise que ces fichiers *« sont une nécessité pour l'État moderne et qu'il est préférable que leur gestion soit confiée à l'Institut dont l'indépendance est connue plutôt qu'à un ministère politique »*. Des critiques continuent à paraître dans la presse généraliste jusqu'en 1950. Closon transmet alors une mise au point au directeur de l'Agence France-Presse qui contribue à les atténuer en posant une question de fond : *« Doit-on n'organiser un pays que dans la mesure où cette organisation ne présentera pas d'inconvénient en cas d'occupation étrangère ou au contraire l'équiper en vue d'éviter cette occupation ? »* Il fait aussi remarquer que *« des documents aussi menaçants que les inventaires de l'Insee comme les listes électorales ou celles de l'état civil existaient en 1940 et qu'ils n'ont pas été utilisés par l'occupant »*. Un an après cette mise au point, la loi sur le secret statistique mettra fin aux critiques sur les fichiers.

Une institution en quête de confiance : la loi sur le secret, la coordination et l'obligation statistique

La loi crée une commission paritaire chargée d'élaborer le programme des enquêtes statistiques et de suivre son exécution – le Comité de Coordination des Enquêtes Statistiques (COCOES) –, impose le secret statistique et l'obligation de réponse ; elle assortit de sanctions l'obligation de réponse et crée un comité de contentieux. Elle introduit le principe de l'agrément accordé aux organisations professionnelles intermédiaires dans les enquêtes qui va permettre d'organiser les enquêtes par branches d'activité. Ses décrets d'applications ne sont pas publiés avant le 15 septembre 1952, ce qui montre le peu d'empressement des décideurs politiques à l'égard de la statistique.

4 SAEF, Rapport sur l'INSEE au ministre de l'Économie Nationale, Closon, mai 1947.
5 Ined : Institut national d'études démographiques.
6 SAEF, Réponse de Closon à Sauvy, décembre 1947.

Le COCOES définit le programme annuel des recensements et des enquêtes statistiques du secteur public, examine leur bien-fondé pour limiter l'émission des questionnaires et éviter les doubles emplois, met au point la rédaction des questionnaires et soumet ce programme au ministre de tutelle de l'Insee qui publie l'arrêté d'application. Chaque opération agréée par le Comité comporte deux visas : celui du ministère enquêteur et celui de l'Insee. Compte tenu de l'emprise réelle de ce Comité sur la quantification publique,

▶ Encadré. Quarante ans au prisme des controverses

Après avoir assis son rôle au sein de l'État, le système statistique public s'est tourné vers l'extérieur pour être plus à l'écoute des critiques formulées à son encontre, et diffuser de façon plus régulière et plus efficace l'ensemble des statistiques produites. Comme le disait Malinvaud en faisant un retour sur le colloque sur la statistique dans une société pluraliste et décentralisée organisé par le Conseil national de la statistique (CNS) en 1983 : « Les statisticiens ne travaillent pas seulement pour la science mais aussi pour les acteurs sociaux, et notamment les pouvoirs publics, qui doivent pouvoir fonder leurs décisions sur des études statistiques. C'est même un des aspects de la démocratie et en dernière analyse l'intérêt bien compris du corps social. » *(Malinvaud, 1983)*. Mais que cela soit dans la phase de construction ou de consolidation, l'Insee a dû faire face à de nombreuses controverses : nous en retiendrons deux qui restent d'actualité.

La première controverse est celle des libertés individuelles en lien avec le développement des fichiers informatiques. La loi de 1951 sur le secret statistique était censée apaiser les tensions sur les données individuelles stockées par l'Insee. Mais le projet de Système automatisé pour les fichiers administratifs et répertoires des individus (dont l'acronyme est SAFARI) relance la polémique. Il s'agit d'un projet d'interconnexion des fichiers nominatifs de l'administration française, notamment par le biais du numéro individuel d'identité. La polémique est lancée par un article du Monde de Philippe Boucher, intitulé « Safari, la chasse aux Français ». Cette polémique mènera à un arrêt du projet et à la création de la Cnil en lien avec la loi Informatique et Libertés. Cette loi et les conséquences de la création de la Cnil ont été mal perçues dans un premier temps par les statisticiens. Comme le rappelle Padieu* : « Gardiens d'une déontologie reposant en partie sur la loi et élaborée pour le reste comme expression d'une éthique professionnelle, ils [Les statisticiens publics] ont assez mal reçu la législation « informatique et libertés », en 1978. Elle leur semblait les soupçonner de crimes qu'ils s'attachaient à ne pas commettre, faisant bon marché du combat pour l'intégrité qu'ils avaient le sentiment d'avoir mené ». Depuis, les tensions avec la Cnil se sont apaisées.

La deuxième concerne l'indice des prix. Au milieu des années 1950, des débats existent sur le « panier de la ménagère » et sa composition. La polémique est l'occasion pour l'Insee de montrer son indépendance en résistant aux pressions gouvernementales pour contrôler ce panier. L'indexation du salaire minimum garanti sur l'indice des prix de détail de l'Insee le met au cœur de la tourmente (Jany-Catrice, 2019 ; Touchelay, 2014). Le gouvernement, par l'intermédiaire de la direction des prix, contrôle les blocages de prix d'un certain nombre de produits. La « mise au secret » de la liste des variétés de produits composant l'indice et la baisse de l'inflation avec le Plan Giscard de 1963 mettent un terme aux tensions. Dans les années 1970, la critique vient des syndicats, en particulier de la Confédération générale du travail (CGT) qui fabrique son propre indice des prix, concurrent de celui de l'Insee, à partir de janvier 1972. Cette fois, ce sont les choix méthodologiques qui sont critiqués et en particulier la correction de l'effet qualité. Syndicats et Insee se répondent par brochure interposée : « Indice Insee, Indice truqué », puis « Pour comprendre l'indice des prix » et enfin « Pour combattre l'indice des prix ». La position de l'Insee, « qui reste la même aujourd'hui » est que l'indice des prix de consommation (IPC) n'est pas un indice du coût de la vie et que la question n'est pas de celle de l'outil, mais de l'usage qui en est fait.

* *Padieu (1991),* « La déontologie des statisticiens », *Sociétés contemporaines, n° 7, pp. 35-62.*

sa composition est l'objet de nombreuses négociations. Le décret du 15 septembre 1952 (n° 52-1059), dressant la liste des administrations et des organisations professionnelles représentées au Comité, est modifié plusieurs fois.

Ce Comité devient l'interlocuteur permanent du patronat en matière de statistiques publiques. Il examine toute initiative susceptible de faciliter la coopération des entreprises et des services enquêteurs ainsi que toute contestation sur telle ou telle investigation administrative. En 1953, l'assemblée générale du CNPF[7] dresse un bilan très positif de la réforme de la statistique publique qui a favorisé la *« coopération entre les services administratifs et les professionnels »* et elle fait le constat que la *« contribution des producteurs à l'élaboration des statistiques officielles »* peut leur procurer de grands avantages[8].

À partir de cette réforme, toute opération statistique résulte d'une concertation officielle, qui parfois aboutit à un rejet[9].

La loi de 1951[10] améliore les relations entre les statisticiens et les milieux professionnels *(Touchelay, 2000)*, mais elle ne suffit pas à éliminer les réticences à l'égard des enquêtes statistiques. L'attitude du journaliste Robert Lazurick, qui assimile la statistique à un *« détraquement de la vie publique »* et qui dénonce *« ces abus de paperasse qui mobilisent un personnel abusivement nombreux »*, dans le quotidien *L'Aurore* en 1953 n'est pas exceptionnelle[11]. Le recensement démographique au printemps 1954 suscite même une véritable campagne d'hostilité. La formule employée par Robert Escarpit dans les colonnes du quotidien *Le Monde* deux jours avant la diffusion des questionnaires est révélatrice : *« Français de naissance ou pas, le nombre de mes enfants, mes domiciles successifs, mes moyens d'existence – qu'est-ce que ça peut bien Leur faire ? »*[12]. En 1957, le groupe parlementaire Union et fraternité française, proche des poujadistes, propose de supprimer la loi de juin 1951[13]. Bien que ces députés affirment ne pas contester l'intérêt des statistiques, ils justifient leur démarche en précisant que la loi *« permet de recourir à des méthodes inquisitoriales touchant à la vie professionnelle et même à la vie privée des individus »*[14]. Ils considèrent en outre qu'elle *« multiplie la paperasserie »,* avant de conclure que *« le temps passé à ce genre de travail est perdu pour la production »*. L'échec de leur tentative le 18 avril 1957 marque le reflux de la vague poujadiste et le succès des éléments modernisateurs du patronat[15].

> ***« Français de naissance ou pas, le nombre de mes enfants, mes domiciles successifs, mes moyens d'existence – qu'est-ce que ça peut bien Leur faire ? »***

7 CNPF : Conseil national du patronat français, prédécesseur du MEDEF (Mouvement des entreprises de France).
8 « Compte-rendu de la quatorzième assemblée générale du CNPF », Bulletin du CNPF, 5 février 1953, p. 16.
9 « Activité du CNPF, questions économiques intérieures : Programme de travaux statistiques intéressant l'industrie et le commerce pour 1956 », Bulletin du CNPF, décembre 1955, p. 3 ; Matheron G, *« Programme des enquêtes statistiques pour 1956 »*, Bulletin du CNPF, mars 1956, p. 22-24. L'auteur préside la commission de l'organisation professionnelle du CNPF.
10 Voir les références juridiques en fin d'article.
11 SAEF, H 1579, n° 309/920, 3 juin 1953, lettre de Closon à Lazurick.
12 SAEF, H 1579, n° 292/920, 19 mai 1954, note de Closon, 3 pages.
13 SAEF, H 1580, n° 232/920, 2 mai 1956, lettre de Closon pour Berger-Perrin, président de l'Association de l'entreprise à capital personnel.
14 SAEF, H 1580, copie de l'exposé des motifs de la proposition de loi n°1534 déposée par René Icher du groupe Union et fraternité française et transmise aux directeurs régionaux de l'INSEE, 12 mai 1956, 2 pages.
SAEF, H 1580, n° 258/920, 18 mai 1956, lettre de Closon pour Edouard Ramonet, président de la commission des affaires économiques de l'Assemblée Nationale.
15 Travaux des commissions présentés dans le Bulletin du CNPF :

Enquêtes ménages et comptabilité nationale : répondre au souhait de planification des décideurs

Pendant cette période de construction, la Statistique publique développe sa production afin de répondre avant tout aux besoins de planification des décideurs politiques. Cette volonté d'extension des enquêtes sera dans un premier temps, freinée par les moyens budgétaires.

La connaissance des conditions de vie s'améliore à partir de la première enquête *« budget des familles »* organisée auprès de salariés modestes dans la région parisienne en 1946. L'opération devient régulière, les échantillons se stabilisent et les questionnaires se diversifient. Elle est complétée par des enquêtes plus ciblées comme celles de 1949 sur les vacances des Français et sur les conditions de vie des personnes âgées. La première enquête emploi est organisée en 1950, suivie par une enquête sur la mobilité professionnelle et sociale en 1952, sur la consommation des ménages en 1953, le budget des familles en collaboration avec le Centre de recherche et de documentation sur la consommation (CREDOC) en 1956, les revenus fiscaux en 1958, etc. Les enquêtes se multiplient pour répondre à la demande de statistiques sociales et documenter la question des inégalités. La collaboration du CREDOC permet d'améliorer les connaissances statistiques sur le logement des Français.

> ***Les statisticiens ont le souci d'analyser « les comportements des milieux sociaux, et [d'étudier] les conditions de vie au sens large »***

Le recensement en 1946 franchit une étape en posant des questions sur le confort. Celui de 1954, puis de 1962, utilise une feuille de logement élargie et fournit des données complémentaires[16]. En 1955, l'Insee et le CREDOC enquêtent le logement des ménages non agricoles puis de l'ensemble des ménages en 1961. De fait, il y a un boom des enquêtes conditions de vie dans le cadre du programme d'investissement prioritaire du V^e^ plan. Ainsi, dans les années 1960 l'Insee s'intéresse aux vacances, aux loisirs, au transport, à la santé, au budget-temps et réalise la première enquête formation qualification professionnelle. Les enquêtes sont conçues pour alimenter les exercices de la planification et plus précisément les prévisions d'emploi, puis elles explorent d'autres champs comme la mobilité sociale et professionnelle, les itinéraires migratoires, etc. *(Monso, Thévenot, 2010)*. Le développement des enquêtes, l'augmentation des moyens budgétaires et l'arrivée des comptables nationaux à l'Insee *(Desrosières, 1998)* correspondent au besoin de la planification. Selon Jacques Desabie, les statisticiens ont aussi le souci d'analyser *« les comportements des milieux sociaux, et [d'étudier] les conditions de vie au sens large » (Insee, 1996, p.83)*.

Cette période est aussi celle de l'arrivée des comptables nationaux à l'Insee. Au début des années 1960, la planification est à son zénith et s'appuie sur les travaux des comptables nationaux pour asseoir ses prévisions. Cette comptabilité nationale est produite par le SEEF dirigé par Gruson à la direction du Trésor. Le SEEF, composé d'une cinquantaine de personnes a une capacité limitée pour produire l'information statistique nécessaire à ses travaux. L'arrivée de Gruson,

– Commission de politique économique générale, mai 1955, p. 2.
– Commission de l'organisation professionnelle, questions économiques intérieures, quinzième assemblée générale du CNPF le 3 juillet 1953, 20 juillet 1953.
– « Compte-rendu des travaux de la commission de l'organisation professionnelle présidée par Georges Matheron », Bulletin du CNPF, février 1959

16 La particularité de ce recensement est la grande automatisation de son exploitation (Insee, 1996).

comme directeur général, et d'une partie des équipes du SEEF à l'Insee en 1961 permet de surmonter cette difficulté. Cependant, l'intégration à l'Insee n'est pas si aisée, car statisticiens et comptables nationaux doivent trouver le bon *modus operandi* pour travailler ensemble. Cette complémentarité est nécessaire pour produire *« une information destinée non seulement à décrire et à expliquer des situations et des évolutions, mais aussi à éclairer des actions qui ne sont jamais purement techniques, qui sont sociales et politiques » (Gruson, 1971).* Cela passera par une réorganisation de l'Insee et la création d'une direction des synthèses économiques en 1962. Au-delà des questions relatives aux comportements des ménages et des groupes sociaux qui sont au cœur des préoccupations des statisticiens, il s'agit de produire les grands agrégats économiques, de mieux suivre la consommation par grands postes, de pouvoir éclairer les décisions de régulation économique et de planification *(Desrosières et al., 1976).*

Dans cette période, l'Insee expérimente les premiers essais de comptes trimestriels et de régionalisation des comptes, le développement de modèles macro-économiques complexes, etc.

C'est aussi une période de transformation de la statistique industrielle avec la mise en place en complément des enquêtes de branche, d'enquêtes de secteurs reposant sur l'unité statistique entreprise, les premières expérimentations d'enquête annuelle d'entreprises, avec surtout une collecte assurée par l'Administration pour ces nouveaux dispositifs, malgré des oppositions du CNPF. C'est également la banalisation de l'utilisation des données fiscales comme les bénéfices industriels et commerciaux, dont la normalisation selon le plan comptable général de 1957 est établie par décret le 28 octobre 1965, et que la Direction générale des impôts transmet officiellement à l'Insee à partir de 1967.

Au terme du mandat de Gruson, l'Insee et plus généralement la statistique publique a développé son offre d'information statistique mais reste à l'écart du débat démocratique et du « grand » public. Il y a eu certes, en 1965, la création du Comité de liaison entre l'Insee, les administrations économiques, et les organisations professionnelles, syndicales et sociales, dans le cadre de la mise en place d'une politique des revenus, mais le chantier ambitieux d'une démocratisation du système d'information statistique reste à faire *(Bardet, 2000).*

▶ Période II : la consolidation par l'ouverture

Quand Gruson fait le bilan des vingt-cinq premières années de l'Insee, il insiste sur le côté scientifique de l'Institut et sur le fait qu'il est passé d'une activité de *« cueillette à la moisson »* : *« En règle générale, l'Insee ne récolte plus que ce qu'il a semé — ce qu'il a semé, non en début de saison, mais plusieurs années à l'avance. C'est pourquoi la construction du système d'information économique est un exemple typique d'activité dans laquelle l'ordre ne se maintient que par la planification, c'est-à-dire à condition de modeler chaque décision sur une conception précise de l'avenir à long terme. » (Gruson, 1971).* Jean Ripert[17] fait un constat similaire mais en parlant des utilisateurs : *« Le statisticien doit souvent rappeler aux utilisateurs les délais nécessairement longs — ils se comptent toujours en années —*

17 Jean Ripert, nommé directeur général de l'Insee en 1967.

> ***« Il importe qu'un dialogue constructif et confiant s'établisse, entre producteurs et utilisateurs d'informations statistiques, pour déterminer des priorités. »***
> *(Ripert, 1971).*

qu'exige la mise en place de nouveaux instruments statistiques. » (Ripert, 1971). Au regard des défis qu'il pressent pour l'Institut, il insiste sur le développement d'une relation entre producteurs et utilisateurs pour améliorer le système : *« il importe qu'un dialogue constructif et confiant s'établisse, entre producteurs et utilisateurs d'informations statistiques, pour déterminer des priorités. » (Ripert, 1971)*. Il va donc s'agir pendant les années suivantes de consolider les acquis d'une institution dont le rôle est reconnu par la création d'espaces d'échanges permettant d'orienter le programme statistique et de mieux répondre à la demande sociale.

L'ouverture au niveau local : la création des observatoires économiques et régionaux (OER)

Le contexte politique des années 1960 contribue à ouvrir l'Insee à de nouveaux publics. Tout d'abord, il y a une forte demande d'informations économiques et sociales au niveau local. Créée au début des années 1960, la Délégation interministérielle à l'aménagement du territoire et à l'action régionale (Datar)[18] souhaite entraîner l'Insee vers la création d'Observatoires économiques régionaux (OER) pour répondre à la demande croissante d'informations locales. En effet, au milieu des années 1960, une réforme entraîne la création des régions administratives, et des « missions régionales » qui regroupent entre autres des représentants administratifs, économiques et universitaires. Par ailleurs, à l'Insee, des discussions s'engagent sur ses missions et sur son rôle. Elles se traduiront en juin 1967 par un colloque sur l'information économique à Villemetrie. Puis viennent les « événements » de mai 1968 qui seront un temps « d'assemblées générales permanentes » au sein de l'Institut où les questions posées sont : *« Comment le travail est-il vécu et organisé ? À quoi les statistiques servent-elles ? Au bénéfice de qui les produit-on ? » (Insee, 1996 p.102)*. Enfin, en 1969, il y a le référendum sur les régions et la réorganisation du Sénat. Autant d'éléments qui ont joué sur la participation et le rôle de l'Insee dans la production et la diffusion de l'information économique et sociale locale *(Bardet, 2000)*.

L'ambition de la Datar consiste à créer dans chaque région des lieux rassemblant l'ensemble des producteurs et des utilisateurs d'information statistique, à la fois pour mieux coordonner la diffusion de cette information et pour *« introduire du pluralisme dans sa production »*. Il s'agit d'aller à la rencontre des utilisateurs, de compléter les données de l'Insee par d'autres sources locales et de faire remonter les demandes d'informations statistiques.

La position de l'Insee sur son implication, sa participation et le contour de l'information évoluent. Ainsi, les premiers OER sont créés à Lille et Marseille en 1967 et d'autres voient le jour jusqu'au milieu des années 1970. Mais la volonté première d'une ouverture totale, sur un programme statistique régional piloté par une commission mixte réunissant

18 La Datar était une administration française chargée, de 1963 à 2014, de préparer les orientations et de mettre en œuvre la politique nationale d'aménagement et de développement du territoire.

producteurs et utilisateurs n'aboutira pas. Toutefois, la question d'un lieu d'échanges entre les utilisateurs et les producteurs d'information statistique n'est pas écartée au niveau national.

Dans le cadre de la réorganisation de l'Insee du début des années 1970, les OER sont rattachés aux directions régionales (DR) de l'Insee[19] et contribuent à une *« prise en compte de la dimension régionale et locale dans l'élaboration des statistiques nationales »* souhaitée par Edmond Malinvaud dès le milieu des années 1970 (Insee, 1996). Par ailleurs, le nombre de cadres A affectés aux études double entre la fin des années 1970 et le début des années 1980, pour répondre à ce besoin d'études régionales et locales.

L'ouverture au niveau national : la création du Conseil national de la statistique (CNS) puis du Conseil national de l'information statistique (Cnis)

Les réflexions sur le rôle local de l'Insee s'accompagnent d'un questionnement sur l'information économique au niveau national avec un double enjeu d'organisation du système statistique et des sujets à traiter ou à mieux traiter. Ces questions ne sont évidemment pas absentes du colloque de Villemetrie et des débats de mai 1968. Ainsi, lors de son discours de politique générale le 16 septembre 1969 devant l'Assemblée nationale, le premier ministre Jacques Chaban-Delmas invite à instaurer une « nouvelle société » dans laquelle la politique de l'information économique sera « repensée » *(Bardet, 2000)*. Reprenant une des propositions du colloque de Villemetrie, le gouvernement crée une commission de l'information économique dans le cadre de la préparation du VI[e] plan. Conscient de l'importance de cette commission pour l'Insee, Ripert veille à la présence de cadres de l'Insee en son sein. Ainsi, le rapporteur général de cette commission est le directeur adjoint des synthèses économiques de l'Insee, Philippe Berthet, et de nombreux collaborateurs de Ripert participent aux groupes de travail de cette commission *(Berthet, 1971)*. Les axes de réflexion retenus sont : prise en compte des besoins des différents utilisateurs de l'information économique ; transparence avec la mise en place de cellules de diffusion dans les centres de production ; formation ; accroissement de l'efficacité du système d'information économique et social. De fait, les appels à un pluralisme ou une démocratisation du service statistique public portés par les revendications externes et internes sont entendus et affirmés dans les travaux de la commission : *« Il est essentiel que chaque membre du corps social, chaque groupe, chaque collectivité puisse jouer son rôle dans la vie sociale, et ceci implique en particulier de veiller à ce que le niveau d'information économique et sociale de chacun lui permette effectivement de jouer ce rôle. » (Berthet, 1971)*. Une des recommandations phares de la commission est la création d'un Conseil national de la statistique qui associe des représentants des administrations, des organisations professionnelles et syndicales, des chercheurs et les producteurs de statistiques. Il s'agit de faire débattre les utilisateurs et les producteurs, en amont de la mise en place des opérations statistiques, des nomenclatures, des répertoires, etc., afin de s'assurer de l'adéquation entre demande sociale et offre, d'éviter des redondances entre les opérations et ainsi limiter les enquêtes. Cette recommandation aboutit à la création du CNS en 1972. Mais au début des années 1980, les partenaires sociaux critiquent le fonctionnement

19 Les directions régionales existent depuis 1941, avant la création de l'Insee en 1946.

de cette instance, car ils ont le sentiment de ne pas être écoutés. Ils affirment *« que le CNS relève d'une manipulation qui permet à l'Insee, prétextant des divergences entre partenaires, de s'arroger implicitement le droit de décider de son programme de travail, alors que le but du CNS est précisément de permettre aux partenaires sociaux d'avoir prise sur ce programme » (Spenlehauer, 1998)*. André Vanoli, ancien secrétaire général du CNS puis du Cnis, constatait une *« double revendication d'une information statistique qui soit à la disposition de tous les acteurs sociaux, et non de manière trop privilégiée au service du seul gouvernement ou des organisations professionnelles, et d'une association des partenaires sociaux à certains des mécanismes déterminant le développement et le contenu de cette information » (Vanoli, 1989)*. Ces réactions combinées au changement de majorité politique conduisent à la création d'un groupe de travail sur la réforme du CNS qui aboutit à la création du Cnis par le décret de 1984. Parmi les changements, on note la création d'un bureau préparant les travaux et composé de cinq représentants des confédérations syndicales de salariés, cinq représentants des organisations représentatives des entreprises, trois représentants de l'administration [le directeur général de l'Insee, le gouverneur de la Banque de France (BdF) et le commissaire au Plan] et deux membres élus par les autres catégories de membres du Cnis[20]. Cette transformation permet de préciser que la concertation porte sur l'ensemble des statistiques publiques[21], que le producteur appartienne ou non au service statistique public, de la production à la diffusion des données. Les thématiques couvertes s'élargissent également avec par exemple, l'apparition des données sur le système financier qui explique la présence d'un représentant de la BdF au Bureau. Le comité du secret statistique est créé pour permettre l'accès des chercheurs ou des organismes publics ou parapublics aux données détaillées des entreprises. Des liens entre le Cnis et la Commission nationale informatique et libertés (Cnil) créée en 1978 dans le cadre de la loi informatique et libertés sont aussi établis.

L'ouverture vers le grand public : la création du département de la diffusion

> ***Une des recommandations phares de la commission est la création d'un Conseil national de la statistique.***

Un des axes de travail évoqués par la commission de l'information économique du VI[e] plan est l'accroissement de l'efficacité du système d'information économique et social. Ripert décide de prendre les devants et lance un audit de l'Insee réalisé par le cabinet Mc Kinsey avant la rédaction de la synthèse des travaux de cette commission. Les recommandations de l'audit font écho aux propositions déjà contenues dans les conclusions du colloque de Villemetrie et de la commission sur l'information économique. Au-delà de la nécessité d'être attentif à la demande sociale, il est nécessaire de diffuser largement les résultats produits.

Rappelant le cœur de métier de l'Insee – produire, analyser et diffuser des statistiques – l'audit propose une réorganisation pour mieux réaliser ces missions. La direction de la production fait son retour à la direction générale, des chefs de service production

20 Voir l'article du courrier des statistiques N6 : *« Le Conseil national de l'information statistique : la qualité des statistiques publiques passe aussi par la concertation »*, Isabelle Anxionnaz et Françoise Maurel.
21 Au sens précisé par Michel Isnard (2018), *« Qu'entend-on par statistique(s) publique(s) », Courrier des statistiques,* N1.

sont introduits en DR. Il s'agit de mieux coordonner les travaux de production en DR et de dégager des marges de manœuvre pour se consacrer à des travaux d'études. Mais la mesure la plus emblématique est la création d'un département de la diffusion. C'est l'occasion de consolider et poursuivre les travaux de refonte des publications et d'asseoir le rôle du bureau de presse créé à la fin des années 1960. Les revues sont spécialisées afin de mieux tenir compte de la diversité des publics visés. Il y a aussi la volonté que certaines revues ne soient pas seulement destinées à des spécialistes de l'information économique et sociale, mais qu'elles soient « accessibles par toute personne cultivée s'intéressant à l'économie » *(Insee, 1996)*. Ainsi, la revue *Économie et statistique* apparue en 1969 bénéficie rapidement d'un rédacteur en chef chargé de réécrire les articles pour les rendre accessibles à un large public. Fin 1973, le premier numéro de *Données sociales* est publié. Il répond à une forte demande de statistiques sociales. Cette publication rencontre un vrai succès éditorial et dépasse les 10 000 exemplaires vendus en 1987. Pour accompagner ce succès, l'Insee crée une division études sociales qui renforce l'équipe de rédaction, utilise et valorise les nombreuses sources du département Population ménages.

> ***La mesure la plus emblématique est la création d'un département de la diffusion.***

Autre succès éditorial, *les tableaux de l'économie française* (TEF) : en 1976, une refonte de cette publication lui assure un large public et dépasse les 24 000 exemplaires vendus en 1987. L'élan des publications nationales se retrouve au niveau régional. Les maquettes des revues régionales sont rénovées, les auteurs formés aux techniques rédactionnelles, les TEF sont déclinés au niveau régional et obtiennent aussi un succès éditorial. Et en 1987, après une initiative de l'OER d'Aquitaine, l'Insee crée un vidéotex national grand public : c'est la naissance du 3615-Insee.

Le développement de ces publications et du bureau de presse donne plus de visibilité aux travaux de l'Insee. On lit dans la presse *« L'Insee dit que… »* plutôt que *« Selon les statistiques officielles… »*. Cela conduit aussi à mettre en place la notion d'embargo qui fixe le moment de la diffusion publique, en imposant aux quelques acteurs informés de façon anticipée (agences de presse, autorités), de respecter l'heure prévue pour la diffusion publique, afin que chacun reçoive l'information en même temps. C'est aujourd'hui un critère important de l'indépendance au regard du code de bonnes pratiques de la statistique européenne.

Au cours des deux périodes de fondation et d'ouverture, l'Insee est parvenu à fournir des statistiques qui répondent mieux aux besoins d'informations des décideurs et d'un public élargi. Le passage de la mécanographie à l'informatique raccourcit le délai entre l'enquête et la publication de ses résultats. Des publications symboliques comme *Données sociales* destinées aux étudiants et aux lycéens des sections de sciences humaines et sociales marquent l'ouverture de l'Institut du service des décideurs à celui d'un plus vaste public et du débat démocratique. S'ouvre une nouvelle période de démocratisation de l'information économique et sociale, de concurrence entre public et privé et de dispersion de la demande, dans un contexte d'internationalisation, en particulier au niveau européen avec l'importance que prend la commission européenne dans les programmes statistiques nationaux. Mais ceci est une autre histoire…

▶ Fondements juridiques

- Loi n°46-854 du 27 avril 1946 modifiée portant ouverture et annulation de crédits sur l'exercice 1946 : les articles 32 et 33 de cette loi créent l'Insee. In : *site de Légifrance*. [en ligne] [Consulté le 15/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000687377*.
- Loi n° 46-1889 du 28 août 1946 relative au contrôle des inscriptions sur les listes électorales et à la procédure des inscriptions d'urgence. In : *site de Légifrance*. [en ligne] [Consulté le 15/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000693137*.
- 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 15/03/2023]. 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 15/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460/*.
- Décret n°46-1432 du 14 juin 1946 modifié portant règlement d'administration publique pour l'application des articles 32 et 33 de la loi de finances du 27 avril 1946 relatifs à l'institut national de la statistique et des études économiques pour la métropole et la France d'outre-mer. In : *site de Légifrance*. [en ligne] Mise à jour le 12 juin 1989. [Consulté le 15/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000872629*.
- Décret n° 52-1059 du 15 septembre 1952 portant application de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistique. In : *site de Légifrance*. [en ligne] [Consulté le 15/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000327099*.
- Décret no 63-112 du 14 février 1963 créant une délégation à l'aménagement du territoire et à l'action régionale et fixant les attributions du délégué. In : *site de Légifrance.* [en ligne] [Consulté le 03/07/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000704036*.
- Décret n°65-968 du 28 octobre 1965 fixant les conditions d'application de l'article 54 du code général des impôts relatif aux renseignements que les entreprises industrielles et commerciales doivent fournir en même temps que la déclaration prévue à l'article 53 du même code et édictant des définitions et des règles d'évaluation auxquelles les dites entreprises sont tenues de se conformer. In : *site de Légifrance*. [en ligne] [Consulté le 15/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000674769*.
- Décret n°84-628 du 17 juillet 1984 relatif au conseil national de l'information statistique (Cnis) et portant application de la loi 51711 du 07-06-1951 sur l'obligation, la coordination et le secret en matière de statistique. In : *site de Légifrance*. [en ligne] [Consulté le 15/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000701777*.

▶ Bibliographie

- AMOSSÉ, Thomas et DE PERETTI, Gaël, 2011. Hommes et femmes en ménage statistique : une valse à trois temps. In : *Travail, genre et sociétés* [en ligne]. 2011/2 (n° 26), pp. 23-46. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *https://www.cairn.info/revue-travail-genre-et-societes-2011-2-page-23.htm*.

- AMOSSÉ, Thomas et DE PERETTI, Gaël, 2011. *Men and Women in Household Statistics : A Piece In Three Acts.* In : *Travail, genre et sociétés* [en ligne]. 2011/2 (n° 26), pp. 23-46. [Consulté le 15/03/2023].Disponible à l'adresse suivante : *https://www.cairn-int.info/abstract-E_TGS_026_0023--men-and-women-in-household-statistics.htm?contenu=article*.

- BARDET, Fabrice, 2000. La statistique au miroir de la région, éléments pour une sociologie historique des institutions régionales du chiffre en France depuis 1940. Thèse de science politique, Paris I Panthéon Sorbonne.

- BERTHET, Philippe, 1971. Information économique. Rapports des commissions du VIe Plan (1971-1975), Paris, La documentation française.

- BRUNAUD, Françoise, THÉLOT, Claude et TOUCHELAY, Béatrice, 2020, Des statisticiens racontent..., Comité pour l'histoire économique et financière de la France. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *https://www.economie.gouv.fr/igpde-editions-publications/des-statisticiens-racontent*.

- CHEVRY, Gabriel, 1948. Un nouvel instrument de travail statistique : le fichier des établissements industriels et commerciaux. Journal de la Société de statistique de Paris, Tome 89 (1948), pp. 245-262. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *http://www.numdam.org/item/JSFS_1948__89__245_0/*.

- CLOSON, Francis-Louis, 1971. Les difficultés d'un commencement. In : *Économie et statistique. supplément pour le vingt-cinquième anniversaire de l'Insee*, n°24, juin 1971, pp. 5-7. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *https://www.persee.fr/doc/estat_0336-1454_1971_num_24_1_6669*.

- DESROSIÈRES, Alain. Les spécificités de la statistique publique en France : une mise en perspective historique. In : *Courrier des statistiques,* janvier 1989, Insee, Paris.

- DESROSIÈRES, Alain. MAIRESSE Jacques et VOLLE, Michel, 1976. Les temps forts de l'histoire de la statistique française. In : *Économie et statistique*, n°83, Novembre 1976. pp.19-28. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *https://www.persee.fr/doc/estat_0336-1454_1976_num_83_1_2401*.

- DESROSIÈRES, Alain, 1998. *The Politics of Large Numbers: A History of Statistical Reasoning* 15 novembre 1998 Édition en anglais de Alain Desrosieres (Auteur), Camille Naish (Traduction) Éditeur : Harvard University Press (15 novembre 1998). 380 pages. ISBN 9-78-0674009691.

- GRUSON, Claude, 1971. Jeunesse d'une institution. In : *Économie et statistique*, n° 24, Juin 1971. pp. 10-11. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *https://www.persee.fr/doc/estat_0336-1454_1971_num_24_1_6672*.

- INSEE, 1996. Cinquante d'Insee… ou la conquête du chiffre. ISBN 9-78-2110663993.

- ISNARD, Michel, 2018. Qu'entend-on par statistique(s) publique(s) ? In : *Courrier des statistiques*, N1, [Consulté le 15/03/2023] Disponible à l'adresse suivante : *https://www.insee.fr/fr/information/3646978?sommaire=3647035.*

- JANY-CATRICE, Florence, 2019. L'indice des prix à la consommation, La Découverte, collection Repères n°717. 3 janvier 2019. ISBN : 9782707199317.

- MALINVAUD, Edmond, 1983. Conseil national de la statistique, colloque « La statistique dans une société pluraliste et décentralisée ». In : *Courrier des statistiques.* Insee n° 27.

- MONSO, Olivier et THÉVENOT, Laurent, 2010. Les questionnements sur la société française pendant quarante ans d'enquêtes Formation et Qualification Professionnelle. In : *Économie et Statistique,* n° 431-432 pp. 13-36. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *https://www.persee.fr/doc/estat_0336-1454_2010_num_431_1_8072.*

- PADIEU, René, 1991. La déontologie des statisticiens, Sociétés contemporaines, n° 7, p. 35-62.Disponible à l'adresse suivante : *https://www.persee.fr/docAsPDF/socco_1150-1944_1991_num_7_1_1008.pdf.*

- PORTER, Theodore, 1995. *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life. Princeton University Press.* ISBN 9-78-0691208411.

- RIPERT, Jean, 1971. Des progrès substantiels à notre portée, n° 24, Juin 1971. pp. 12-13. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *https://www.persee.fr/doc/estat_0336-1454_1971_num_24_1_6673.*

- SPENLEHAUER, Vincent, 1998. L'évaluation des politiques publiques, avatar de la planification. Thèse de sciences politiques à l'Université Pierre Mendès-France de Grenoble.

- TOUCHELAY, Béatrice, 1993. L'Insee des origines à 1961 : évolution et relation avec la réalité économique, politique et sociale. Thèse d'histoire contemporaine, université de Paris 12.

- TOUCHELAY, Béatrice, 2000. Le service central de la statistique publique et l'entreprise française jusqu'aux années 1960 : un jeu de cache-cache ? In : *Anne Pezet, Nicolas Berland* (dir.), JHCM, Faculté Jean Monnet – PESOR et AFC, pp. 363-391.

- TOUCHELAY, Béatrice, 2014. Les ordres de la mesure des prix. Luttes politiques, bureaucratiques et sociales autour de l'indice des prix à la consommation (1911-2012). In : *Politix,* vol. 27, n°105, pp. 117-138. [Consulté le 15/03/2023]. Disponible à l'adresse suivante : *https://www.cairn.info/revue-politix-2014-1-page-117.htm.*

- VANOLI, André, 1989. Le Cnis. In : *Courrier des statistiques*. Insee n° 52, pp. 11-18.

Comptes nationaux distribués : une nouvelle manière de distribuer la croissance

Une expérience innovante au service du débat public

Mathias André[*], *Jean-Marc Germain*[**] *et Michaël Sicsic*[***]

Les comptes économiques distributionnels sont utilisés pour étudier la répartition du revenu national entre les ménages. Ils permettent notamment d'analyser conjointement l'effet redistributif des services publics, des prestations monétaires et des prélèvements. Sur la base de travaux existants et d'innovations récentes, l'Insee a développé une nouvelle approche désignée par « comptes nationaux distribués » (CND). Les CND sont fondés sur une méthode qui relie les données individuelles de la statistique sociale aux agrégats macroéconomiques de la comptabilité nationale. Ils permettent de décrire la distribution du revenu national et de son évolution entre deux années. Ils quantifient également la réduction des inégalités opérée par l'ensemble des transferts publics, versés ou perçus par les ménages. Plusieurs grilles de lecture des résultats sont possibles en regroupant les ménages par groupe de revenus, cohortes d'âge, catégories socio-professionnelles, diplômes ou tailles des territoires. Un apport majeur permettant la comparabilité internationale est d'élargir le champ de la redistribution en intégrant les transferts en nature, comme la santé et l'éducation, ainsi que les dépenses collectives, comme la police et la justice. Cet article présente l'histoire, la méthode, les principaux résultats déjà publiés ainsi que les perspectives des CND.

Distributive accounts are used to study the distribution of national income among households. Redistributive impact of monetary benefits and taxes can be analysed together with the one of public services. Based on existing work and recent innovations, INSEE has developed a new approach referred to as distributive national accounts (DNA). The DNA is based on a method that connects individual data from social statistics to macroeconomic aggregates of national accounts. They provide a description of the distribution of national income and an estimate of the reduction of inequalities achieved thanks to all public transfers received or paid by households. Results can be presented by grouping households by income groups, age cohort or socio-professional categories and sub-national zone. The inclusion in the redistribution field of transfers in kind, such as health and education, and collective expenses, such as police, justice and local communities, is a major contribution. It enables international comparability and plays a major role in this expanded redistribution. The history, the method, the main results already published and the perspectives of DNA are presented in this article.

* À la date de rédaction de l'article, Chargé d'études – Division Études macroéconomiques, Dese, Insee, *mathias.andre@insee.fr*

** Conseiller technique, DG, Insee, *jean-marc.germain@insee.fr*

*** À la date de rédaction de l'article, Chargé d'études – Division Revenus des ménages, DSDS, Insee, *michael.sicsic@insee.fr*

À qui bénéficie la croissance ? Comment se distribue le revenu national entre les ménages ? Quel est l'effet redistributif de l'ensemble des services publics ? Répondre à ces trois grandes questions au centre des débats économiques est l'objet des comptes économiques distributionnels. Ces derniers s'inscrivent dans une histoire longue dans laquelle l'Insee joue un rôle important. Une première étape a consisté à rapprocher les données microéconomiques de la statistique sociale des résultats de la comptabilité nationale afin d'étudier les différences de revenus et de consommation selon l'hétérogénéité des situations des ménages. Il s'agissait alors de construire la décomposition du compte des ménages par catégories de ménages[1]. Récemment, une approche innovante a été développée par l'Insee afin de prolonger et de compléter ces comptes par catégorie de ménage. Elle vise à répartir entre les seuls ménages l'ensemble du revenu national, tous secteurs institutionnels confondus. Elle s'appuie sur les comptes nationaux distribués (CND), construits dans le cadre d'un rapport d'un groupe d'experts *(Insee méthodes n° 138, février 2021)*. Ces travaux s'inscrivent dans une littérature statistique et universitaire cherchant à élargir le champ de la redistribution et à rapprocher données microéconomiques et approche comptable ***(encadré 1)***. Ainsi, les CND permettent d'analyser la réduction des inégalités opérée par l'ensemble des transferts publics, dans une optique dite élargie de la redistribution ***(encadré 2)***. Cette approche élargie de la distribution primaire des revenus et de la redistribution repose sur une méthode présentée dans cet article.

▶ Encadré 1 : Histoire des comptes distribués

La méthode des comptes distribués s'appuie sur des réflexions anciennes et une littérature riche.

Cette dernière s'est d'abord intéressée à la décomposition des comptes des ménages par catégories usuellement produite par l'Insee *[Accardo et Billot, 2020]*. De multiples travaux au sein de la statistique publique ont cherché dès les années 1980 à compléter l'approche microéconomique de la redistribution monétaire par une décomposition des comptes nationaux (historique dressé par *Accardo*, 2020). Entre 1980 et 1985, l'Insee a publié annuellement un compte de revenus pour plusieurs dizaines de types de ménages afin de donner une image du budget d'un ménage en fonction de ses caractéristiques sociodémographiques. Plus récemment, *Accardo et al.,* (2009), ont proposé une décomposition du compte des ménages par catégorie portant sur l'année 2003, en combinant l'approche des comptes nationaux et les statistiques microéconomiques sur les inégalités (document de travail de *Bellamy et al.*, 2009). Le revenu disponible et la consommation des comptes nationaux sont décomposés selon quatre critères socio-économiques : niveau de vie, composition du ménage, âge ou catégorie socioprofessionnelle de la personne de référence. Cela permet de déduire le taux d'épargne selon ces différentes caractéristiques. Cette approche a été reprise par *Le Laidier* (2009) et plus récemment par *Billot* et *Bourgeois* (2019), afin, notamment, de comparer les évolutions annuelles des comptes par catégorie de ménages.

Par ailleurs, l'Insee a publié des travaux élargissant le champ de la redistribution en intégrant les transferts « en nature » et les services publics individualisables *[Amar et al., 2008]*, mais sans couvrir les services publics collectifs, les taxes sur la production et la consommation, et l'imposition des sociétés.

Certains travaux universitaires récents, notamment ceux du *World Inequality Lab [Bozio et al., 2020 ; Alvaredo et al., 2020]*, intègrent également les impôts sur la production et sur les produits mais reposent sur des hypothèses générales pour les prestations « en nature » et les dépenses collectives revenant à neutraliser leur effet sur la redistribution. Ces travaux s'appuient sur un réseau de chercheurs qui s'attache à mettre en œuvre dans différents pays du monde un concept proche des CND, les DINA (*distributional international national accounts*). Ces travaux ont émergé à la suite d'une publication influente de *Thomas Piketty*, *Emmanuel Saez*, et *Gabriel Zucman* (2018).

1 Les comptes dits « par catégories » correspondent à un compte distribué du secteur des ménages (CDSM) et visent surtout à déterminer les variations du taux d'épargne entre les différentes catégories de ménages. Ils se limitent en conséquence au secteur institutionnel des ménages (S14) et à leur revenu disponible brut, qu'ils mettent en regard de leur consommation.

Après avoir détaillé le cadre comptable, les enjeux à rapprocher les statistiques sociales individuelles des données macroéconomiques et à adopter une vision complète de la redistribution sont présentés. Le fonctionnement et les conclusions du groupe d'experts ont permis de construire une méthode de comptabilité distributionnelle expliquée ensuite. Pour conclure, les principaux enseignements ainsi que les perspectives de travaux futurs sont détaillés.

De la comptabilité nationale à la redistribution élargie

La comptabilité nationale adopte un ensemble de conventions concernant les concepts de revenus et de production des agents économiques. Il s'organise autour de secteurs institutionnels (sociétés, ménages, administrations dans une version simplifiée[2]) qui vont produire et échanger des biens et des services. Ce cadre comptable permet ainsi de construire les grands agrégats macroéconomiques tels que la valeur ajoutée de l'économie nationale, le revenu disponible brut des ménages, le patrimoine économique, etc. *(Vanoli, 2002)*. Intégrés dans le tableau économique d'ensemble, les échanges entre secteurs sont décrits de façon détaillée et sous la forme d'emplois ou de ressources. Fruits d'un travail titanesque de recueil de données et de réconciliation entre les différentes sources, les outils comptables définissent le revenu national d'une économie, en retranchant au PIB le solde des importations et des exportations *(Lequillier et Blades, 2014)*.

Encadré 2 : Notions et définitions de la comptabilité distributionnelle

- Le **niveau de vie** (ou niveau de vie usuel) est égal au revenu disponible du ménage divisé par le nombre d'unités de consommation (UC). Le niveau de vie est donc le même pour tous les individus d'un même ménage. Les unités de consommation sont calculées selon l'échelle d'équivalence dite « de l'OCDE modifiée » qui attribue 1 UC au premier adulte du ménage, 0,5 UC aux autres personnes de 14 ans ou plus et 0,3 UC aux enfants de moins de 14 ans. Si on ordonne une distribution de niveaux de vie, les **déciles** sont les valeurs qui partagent cette distribution en dix parties égales. Les individus ainsi classés appartiennent à des **dixièmes** de niveau de vie : les 10 % les plus modestes constituent le premier dixième.
- Un **prélèvement** est un transfert versé par les ménages aux administrations publiques et aux institutions sans but lucratif au service des ménages (ISBLSM). Une **prestation** est un transfert reçu par les ménages. Elle peut être en espèces ou « en nature ».
- La **redistribution élargie** intègre l'ensemble des transferts publics des différents secteurs institutionnels de la comptabilité nationale, y compris les services publics collectifs. Afin de mesurer les effets de l'ensemble des prélèvements, des prestations et des dépenses collectives, elle compare par différence les « revenus **avant transferts** », aux **revenus après transferts** dits « **niveaux de vie élargis** ». La **redistribution usuelle monétaire** se concentre sur la prise en compte des transferts monétaires (et non des prestations contributives comme la retraite et les cotisations sociales).
- Le **revenu national net** est obtenu en retranchant la consommation de capital fixe (CCF), qui correspond au coût d'usure du capital, au revenu national brut. Le revenu national brut est la somme des revenus primaires perçus par les unités économiques résidentes, elles-mêmes ventilées au sein des secteurs institutionnels. Il est égal au produit intérieur brut (PIB) diminué des revenus primaires versés à des unités économiques non résidentes et augmenté des revenus primaires reçus du reste du monde par des unités résidentes.

2 Les différents acteurs de la vie économique sont regroupés selon leurs comportements économiques en cinq secteurs institutionnels résidents : les sociétés non financières (SNF), les sociétés financières (SF), les administrations publiques (APU), les ménages, les institutions sans but lucratif au service des ménages (ISBLSM).

> ***La première étape essentielle à la distribution de l'ensemble du revenu national aux ménages résidents est de considérer les ménages comme destinataires finaux des revenus des autres secteurs institutionnels.***

La première étape essentielle à la distribution de l'ensemble du revenu national aux ménages résidents[3] est de considérer les ménages comme destinataires finaux des revenus des autres secteurs institutionnels ***(figure 1)***. Les entreprises sont possédées par les ménages, soit directement en tant que patrimoine professionnel, soit indirectement par le patrimoine financier et l'épargne. De la même manière, les administrations publiques sont *in fine* attribuées aux ménages.

À partir de la mesure fine des différents revenus des secteurs institutionnels et les transferts qu'ils opèrent entre eux, il est alors possible de définir la réduction des inégalités organisée par les transferts publics. L'objectif premier est de tenir compte du fait que tout ce qui est fourni par la collectivité est financé directement ou indirectement par la population, et profite *in fine* à celle-ci, de nouveau de manière directe ou indirecte. En outre, et c'est un avantage majeur de cette méthode, seule l'exhaustivité des revenus et transferts pris en compte permet des comparaisons robustes entre pays ou entre périodes pour un même pays *(cf. infra)*. La décomposition des différentes composantes, comme les dépenses de retraites ou de santé, permet notamment une comparabilité entre les différents systèmes internationaux, dont le caractère socialisé ou public peut différer. Pour cela, l'ensemble du revenu national est attribué et distribué aux ménages résidant en France.

Ainsi, deux notions essentielles du revenu sont introduites :

- Le revenu avant transfert détermine qui « reçoit » le revenu primaire. Il se distingue principalement du revenu primaire des ménages par l'attribution aux ménages des profits réinvestis dans les entreprises, c'est-à-dire l'épargne des sociétés, mais aussi du revenu primaire des administrations publiques dont l'essentiel est constitué des taxes sur la production et les produits ;
- Le revenu après transferts mesure qui « bénéficie » des transferts publics (ou y « contribue »). Il ajoute notamment au revenu des ménages une valorisation des services rendus par les administrations publiques.

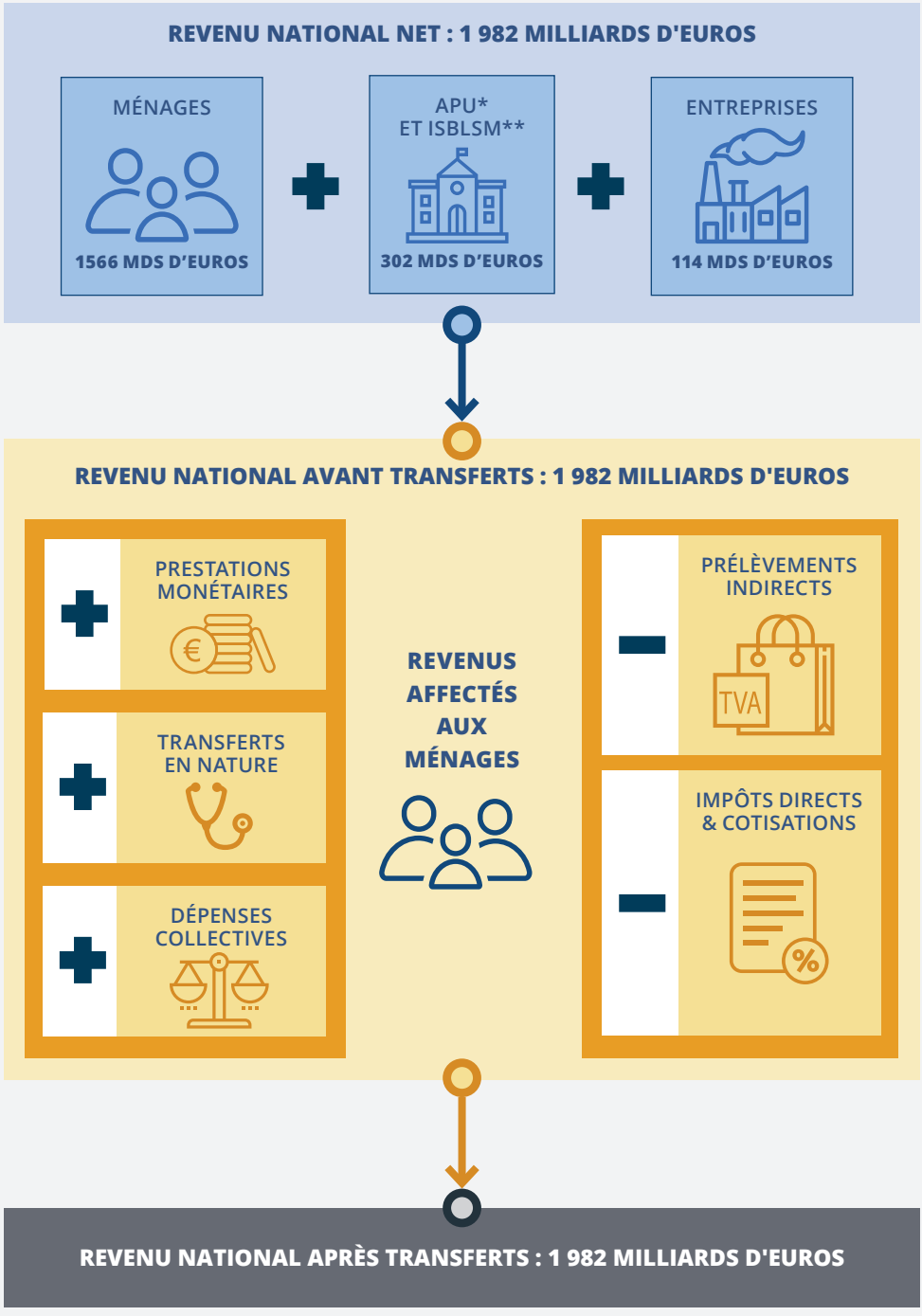
La redistribution élargie se mesure pour chaque ménage par différence entre ces deux concepts centraux et diffère de la mesure usuelle dite monétaire ***(figure 2)***. La redistribution monétaire examine les transferts monétaires les plus directs, c'est-à-dire, les prélèvements comme l'impôt sur le revenu, la contribution sociale généralisée (CSG) et les prestations sociales monétaires (prestations familiales ou minima sociaux par exemple)[4].

En distribuant aux ménages résidant en France l'ensemble du revenu national net, la redistribution élargie tient compte des prélèvements qui affectent d'une manière indirecte les ménages, comme la taxe sur la valeur ajoutée (TVA) ou encore les droits d'accise (sur le tabac et l'alcool par exemple).

3 Une correction est effectuée avec le reste du monde pour l'épargne des entreprises possédée par les ménages non-résidents et par les résidents dans les entreprises à l'étranger.

4 Dans certaines analyses de la redistribution monétaire, les prestations contributives (retraites et assurance chômage) et les cotisations sociales peuvent être prises en compte.

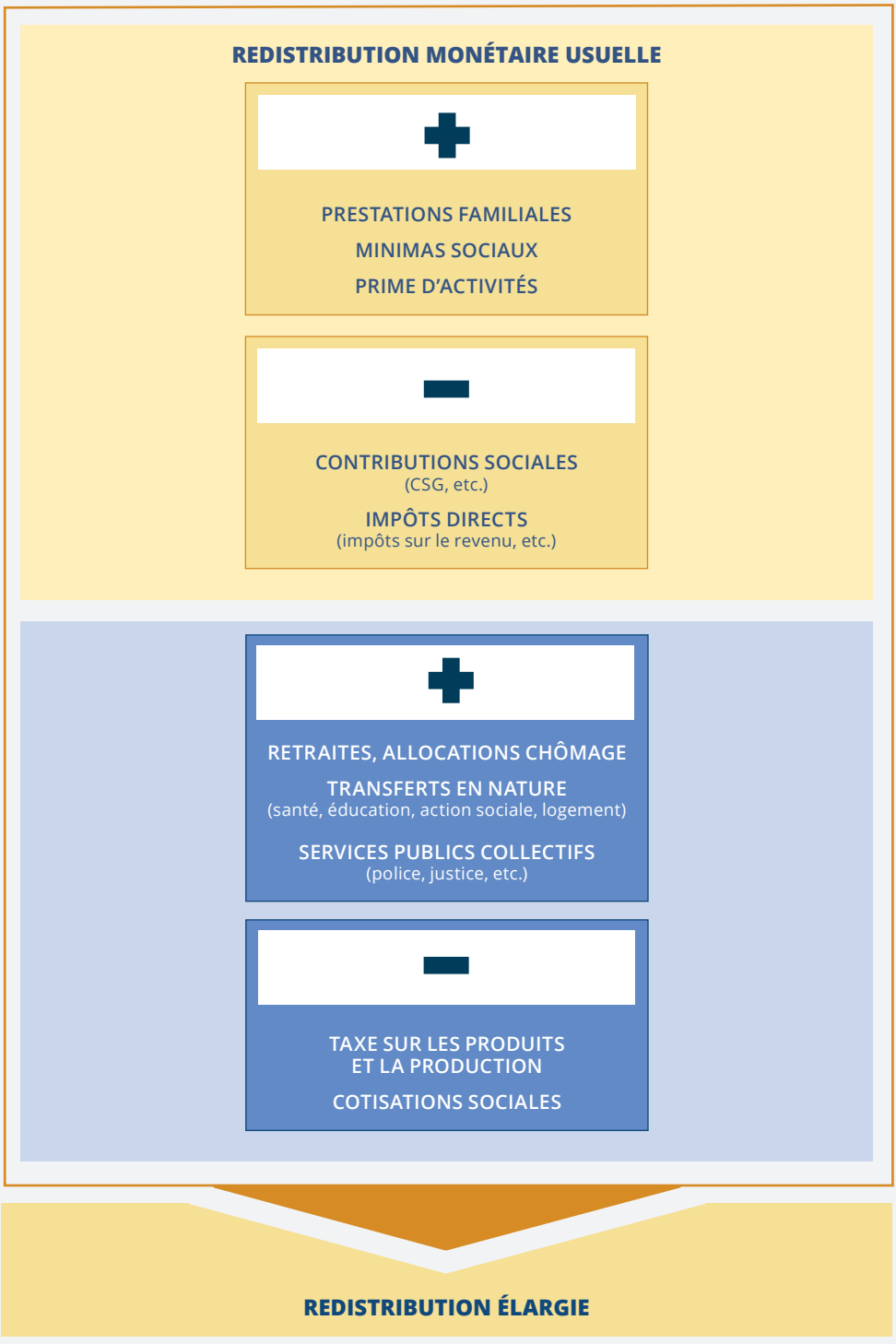
Figure 1 - Du cadre comptable à la redistribution élargie



** APU : Administrations Publiques*

*** ISBLSM : Institution Sans But Lucratif au Service des Ménages*

Figure 2 - De la redistribution monétaire usuelle à la redistribution élargie



Elle inclut également les transferts sociaux dits « en nature » par la comptabilité nationale, et les services publics rendus par les dépenses collectives (dépenses des administrations publiques dans des domaines tels que la santé et l'éducation, mais aussi les services publics comme la police, la justice ou les services d'administration, centrale ou locale). Ce deuxième groupe de transferts est moins directement comptable que le premier mais sa quantification n'en apparaît pas moins essentielle au regard de la réalité qu'elle représente et de son ampleur, afin d'analyser précisément le caractère redistributif du système socio-fiscal.

Le cadre exhaustif de la redistribution élargie a ainsi l'avantage de "prélever" autant qu'il "verse", et réciproquement. En ce sens, elle est dite équilibrée et permet d'effectuer des comparaisons temporelles mais aussi internationales. En effet, dans le cas par exemple d'une réforme visant à augmenter ou baisser la TVA en contrepartie d'une baisse ou d'une hausse de l'impôt sur le revenu ou des cotisations sociales, la non prise en compte de la TVA dans l'analyse biaiserait les comparaisons temporelles et internationales. Néanmoins, à ce stade, ce cadre élargi attribue aux ménages l'ensemble du revenu sans fournir d'information sur l'hétérogénéité des revenus ou transferts au-delà des secteurs institutionnels. Comme l'a notamment souligné le rapport *Stiglitz-Sen-Fitoussi*, « aller au-delà du PIB » (2009) est une demande sociale forte, et ancienne. Il s'agit alors de dépasser les seuls agrégats comptables qui, divisés par la population, fournissent des moyennes, et de construire des distributions des revenus et de transferts à différentes catégories de ménages. La méthode de la comptabilité distributionnelle associée aux outils des comptes nationaux distribués mesure la déformation de la distribution des revenus avant et après transferts.

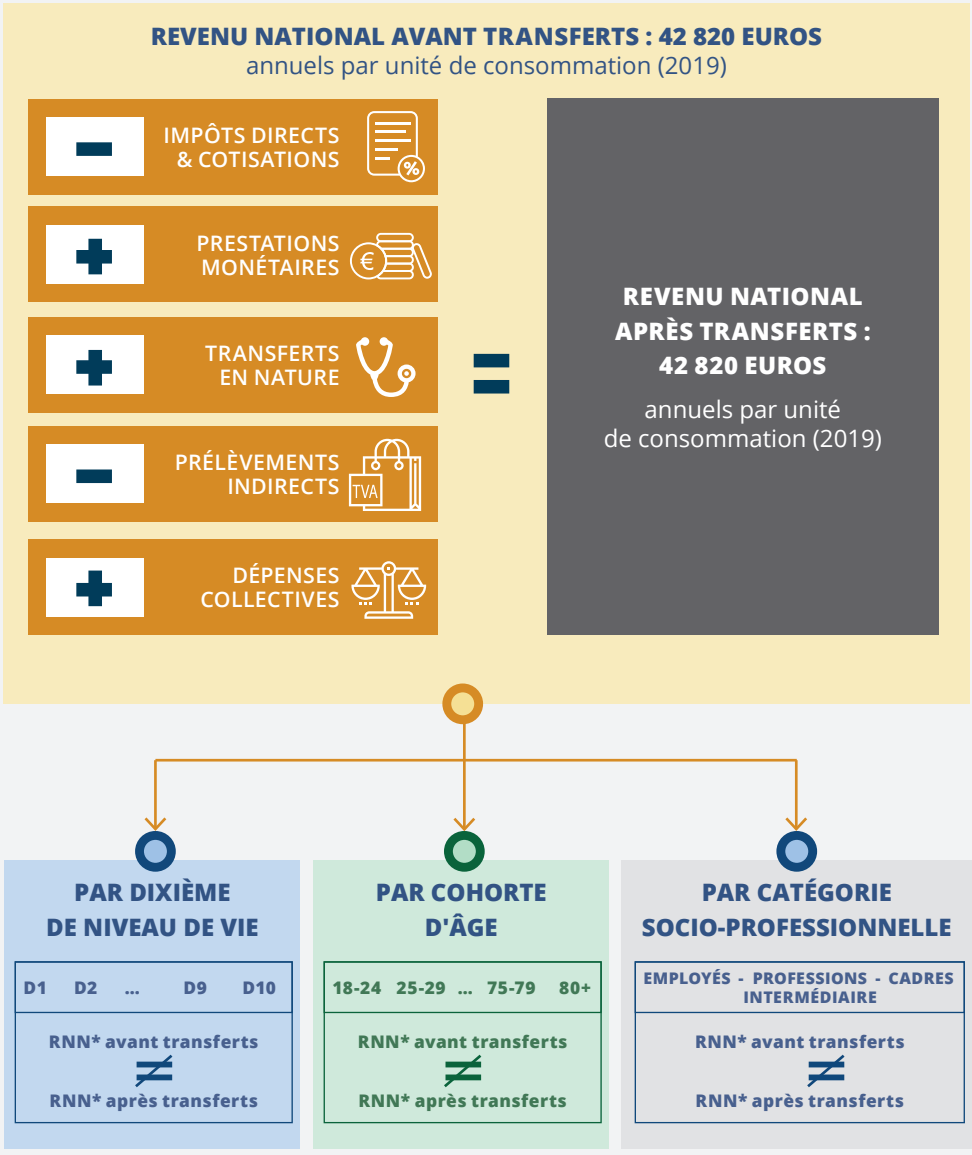
▶ Réconcilier la comptabilité nationale et les statistiques sociales

> ***La comptabilité économique distributionnelle croise les informations comptables macroéconomiques avec des données individuelles de la statistique sociale.***

Partant de la représentation générale comptable ainsi agrégée dans un seul secteur institutionnel, les ménages, la comptabilité économique distributionnelle a pour objet d'effectuer une répartition entre les ménages de tout ou partie de ces agrégats. Elle documente ainsi l'hétérogénéité des situations économiques de différentes catégories de ménages, et plus encore leur inégalité, sur la base des concepts de la comptabilité nationale.

Pour ce faire, elle croise les informations comptables macroéconomiques avec des données individuelles de la statistique sociale. Sous la forme des comptes nationaux distribués, elle prend la forme d'un tableau distributionnel d'ensemble, inspiré du tableau économique d'ensemble qui décrit la séquence de passage du revenu distribué avant transferts au revenu distribué après transferts. Les transferts sont constitués de l'ensemble des prélèvements, sous la forme d'impôts, cotisations et taxes, ou bien de prestations monétaires et services publics, en nature s'ils sont individualisables comme les dépenses de santé ou d'éducation, ou collectifs comme

Figure 3 - Différentes représentations de la redistribution élargie par groupes de ménages



RNN : Revenu National Net

les dépenses de justice, police ou de défense. Distribué à chaque catégorie de ménages, le profil avant et après transferts définit ainsi la redistribution élargie entre chaque groupe de ménages ***(Figure 3 et encadré 3)***. Il est alors possible de comparer les inégalités avant et après, et par différence, la redistribution élargie, c'est-à-dire la contribution des transferts à l'atténuation des inégalités primaires.

La première étape consiste à réconcilier les différences de champ et de concepts entre les agrégats macroéconomiques de la comptabilité nationale et les différentes sources de données microéconomiques. Ainsi, les pensions de retraites et les allocations chômage sont considérées comme des revenus différés dans la statistique sociale et entrent dans la définition du revenu primaire *(Sicsic, 2021)*. En comptabilité nationale, comme en comptabilité distributionnelle, elles sont considérées comme des transferts. Cette convention peut contribuer à modifier les inégalités primaires, notamment pour les ménages retraités aisés en classant les individus en fonction de leur revenu primaire. Ce n'est pas le cas lorsque le revenu disponible est la variable utilisée pour le classement entre individus et que celui-ci reste inchangé comme préconisé par le rapport d'experts Insee (2021).

Autre élément important, la comptabilité nationale intègre les loyers imputés aux ménages propriétaires, c'est-à-dire la valorisation du service de logement qu'ils se rendent à eux-mêmes. Un ménage propriétaire de son logement ne paie pas de loyer, à la différence d'un ménage propriétaire qui serait locataire de son logement et propriétaire bailleur d'un autre. La comptabilité nationale vise ainsi à égaliser ces deux situations en ajoutant un revenu fictif aux propriétaires occupants égal au montant du loyer qu'ils auraient à payer s'ils étaient locataires. Ceci permet les comparaisons internationales, notamment entre les pays qui ont des proportions de ménages propriétaires et locataires différents. Cette mesure n'est pas intégrée dans le revenu primaire de la statistique sociale et n'est donc pas incluse dans le taux de pauvreté.

▶ Encadré 3 : Distribuer les transferts selon d'autres variables que le revenu : c'est possible.

L'attribution des revenus et transferts de la comptabilité nationale est réalisée au niveau de chaque ménage du modèle Ines*, qui comprend une grande diversité d'informations socio-économiques. Cette méthode autorise par divers regroupements, l'analyse des inégalités et de la redistribution sous différents angles.

La plus classique et intuitive est celle selon le niveau de vie. Dans ce cas, les individus sont classés selon leur niveau de revenu disponible par unité de consommation du ménage (définition dans l'***encadré 2***), en dix ou vingt groupes égaux. Chaque transfert est ensuite affecté à un individu selon des hypothèses d'incidence fiscale, sans modifier le classement initialement réalisé sur la base du niveau de vie. D'une manière générale, l'hypothèse faite est la suivante : l'individu qui paie un impôt est celui dont le montant de l'impôt dépend indirectement. Les cotisations employeurs portent ainsi sur les salariés, car elles sont assises sur la masse salariale.

Il est également possible de classer les individus selon d'autres caractéristiques que le niveau de vie : l'âge ou le diplôme de la personne de référence du ménage, la configuration familiale du ménage, son lieu d'habitation (selon la taille d'unité urbaine par exemple). Des croisements de variables sont aussi envisageables (comme l'âge et le niveau de vie) : il est alors nécessaire de restreindre les catégories pour s'assurer d'avoir assez d'individus dans les échantillons.

** Ce modèle s'appuie sur les données de l'ERFS, ce qui permet de simuler finement un grand nombre de transferts socio-fiscaux, notamment les prestations sociales dans le bas de la distribution (ainsi que certains transferts non monétaires) et les cotisations sociales. Il permet aussi de s'assurer de la cohérence des revenus et transferts utilisés en utilisant une source centrale, une des recommandations du rapport d'experts (Insee, 2021).*

Il existe également des écarts dans la définition du revenu disponible entre les deux approches. Les allocations logement versées aux ménages locataires à revenus modestes sont considérées comme des prestations monétaires par la statistique sociale mais comme un transfert en nature dans le cadre comptable. À ce titre, elles ne sont pas ajoutées comptablement dans le revenu disponible des ménages. La comptabilité distributionnelle suit cette dernière convention. À signaler enfin, concernant les montants les plus importants, que la comptabilité nationale, et donc la comptabilité distributionnelle, intègrent aux revenus primaires une valorisation de fraude des entreprises comme du travail au noir, ce qui n'est pas le cas des statistiques sociales.

▶ Une démarche inédite entre statisticiens et universitaires

Plusieurs travaux de recherche ou institutionnels ont construit une comptabilité distributionnelle ***(encadré 4)***. Face à la diversité des approches, aboutissant parfois à des conclusions contradictoires, l'Insee a pris l'initiative de proposer un cadre de travail commun aux différentes équipes engagées dans ces démarches, composées d'universitaires ou de statisticiens et d'experts de l'administration[5].

▶ Encadré 4 : Des initiatives internationales complémentaires

Les travaux de l'Insee s'inscrivent dans un paysage international dans lequel plusieurs acteurs institutionnels s'intéressent à la distribution des agrégats comptables. Plusieurs pays avancent dans une mise en production récurrente de distribution de certains agrégats comptables. Les échanges ont principalement lieu dans le cadre de l'OCDE, Eurostat* ou UNStats**.

En raison d'une forte demande sociale de travaux sur ces sujets et l'apport de telles statistiques, le paysage international évolue rapidement sur le plan institutionnel. Dans le cadre de la révision du Système des comptes nationaux (SNA 2025) en cours au niveau mondial, un chapitre sera dédié à la distribution des comptes.

Dans le même temps, différents groupes de travail et institutions multilatérales abordent la distribution des agrégats comptables. Tout récemment, Eurostat a procédé au lancement d'une *task force* dédiée visant à accélérer et généraliser la production de comptes distributionnels.

L'OCDE a également mené des travaux allant dans ce sens. *L'Expert Group on Micro Statistics on Income, Consumption and Wealth* (EG ICW) porte essentiellement sur la cohérence microéconomique des données, mais travaille en lien avec *l'Expert Group on Disparities in National Accounts* (EG DNA), une autre initiative de l'OCDE qui porte sur la cohérence microéconomique et macroéconomique des statistiques distributionnelles. Plusieurs instituts statistiques produisent des statistiques expérimentales sur le sujet (*Statistics Netherlands*, 2014 ; Eurostat, 2018 ; *Statistics Canada*, 2018 ; *Australian Bureau of Statistics*, 2019). À ce stade, la plupart de ces statistiques sont fondées sur des enquêtes et ne couvrent qu'une part du revenu national. Les États-Unis, l'Italie ou le Mexique mènent des réflexions similaires sur le rôle des transferts en nature ou les séries temporelles de comptes distribués.

Enfin, dans le cadre du programme européen PAS2 (*Panafrican Statistics*), des moyens ont été alloués à l'Insee par Eurostat afin d'appuyer certains pays africains dans l'instruction et le développement de comptes économiques distributionnels à l'horizon 2024.

** Eurostat est l'institut statistique communautaire, associé à la Commission européenne.*

*** UNStats est la Division de statistique des Nations unies.*

5 Ils sont issus des services statistiques ministériels concernés, de l'OCDE et des équipes universitaires du *World Income Lab* (WIL) et l'Institut des politiques publiques (IPP) à l'École d'économie de Paris, l'Observatoire français des conjonctures économiques (OFCE), ainsi que le Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP) à Sciences-Po.

Le groupe d'experts s'est attaché à identifier les sources d'écarts entre les travaux existants, qui sont : les sources de données utilisées (enquête ou base administrative), les écarts de méthode pour mesurer les revenus (ménages ou individus, échelles d'équivalence) et les concepts plus ou moins élargis de revenu, avant comme après transferts. Ce sont ces derniers, surtout, qui conduisent à des différences notables. L'approche usuelle est celle de la redistribution dite « monétaire » ; elle prend en compte les impôts directs, les cotisations sociales et les prestations en espèces ; les travaux du *World Income Lab (WIL)* ajoutent les impôts sur la production et sur les produits, dont la TVA ; l'OCDE exclut ces derniers mais ajoute les prestations sociales en nature et les services publics individualisables, que l'Insee intègre aussi à ses analyses mais de manière plus occasionnelle. Aucune de ces approches ne prend en compte les dépenses publiques intégralement collectives.

> ***Tout ce qui est fourni par la collectivité est financé directement ou indirectement par la population et profite in fine à celle-ci, de nouveau de manière directe ou indirecte.***

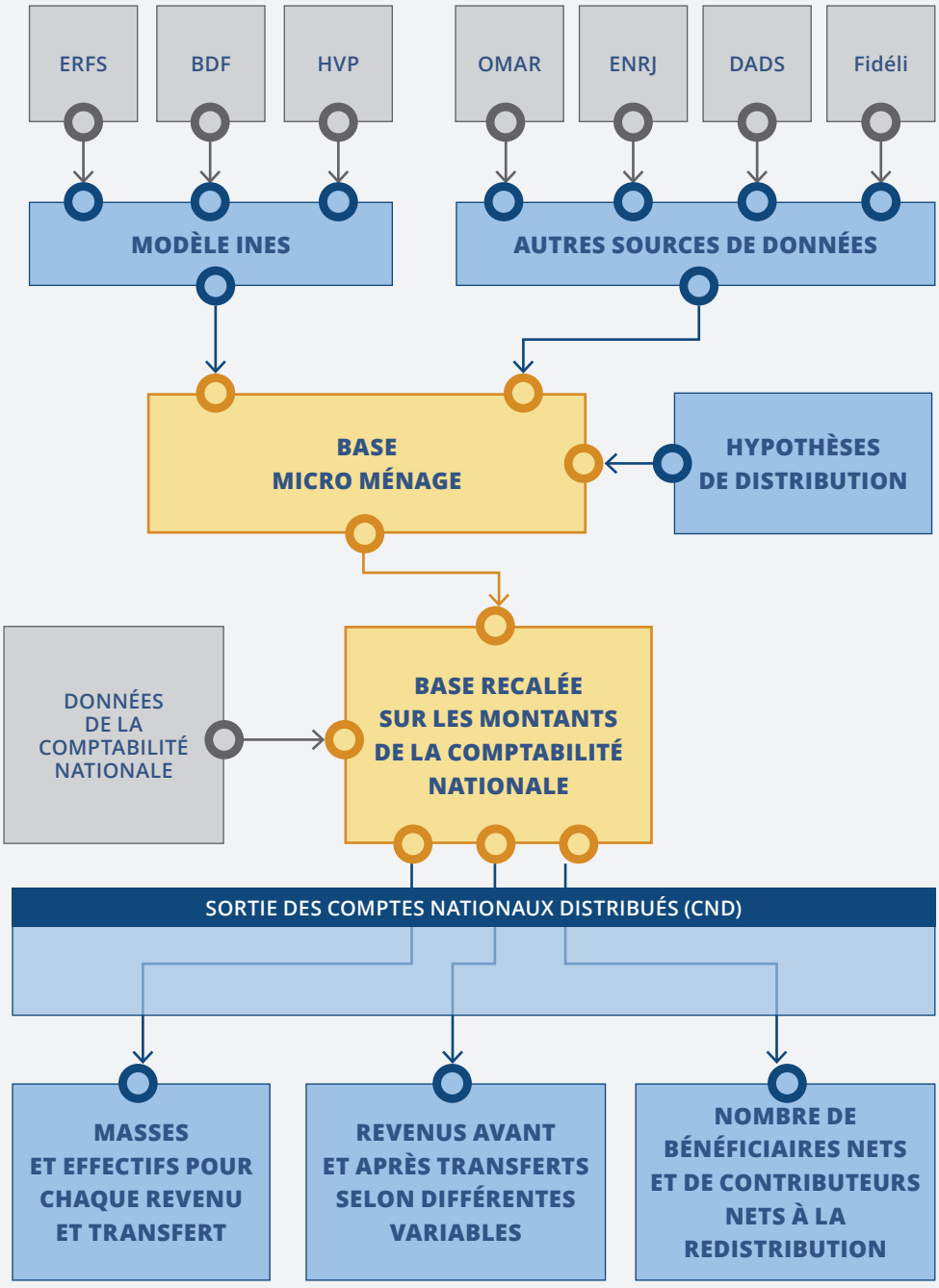
Ces travaux ont abouti à un rapport publié par l'Insee *(Insee méthodes n° 138, février 2021)*, regroupant des recommandations visant à développer un cadre unifié pour construire des comptes nationaux distribués, ainsi qu'un prototype portant sur l'année 2016. Ce rapport recommande notamment d'adopter autant que possible une vision exhaustive de la redistribution réconciliant les différentes approches, en incluant tous les modes de financement et tous les types de prestations ou services publics. Ainsi, tout ce qui est fourni par la collectivité est financé directement ou indirectement par la population et profite *in fine* à celle-ci, de nouveau de manière directe ou indirecte. Plus les concepts de revenus et de redistribution sont élargis, plus les hypothèses d'imputation sont nombreuses. Aucune de ces approches ne saurait prétendre à remplacer les autres ; elles doivent être regardées comme différents « halos » éclairant de manière complémentaire les inégalités et la redistribution.

Ce rapport a été suivi de publications par l'Insee *(cf. infra)* reprenant l'ensemble du cadre de distribution des comptes. Les conclusions du groupe d'experts ont conduit à une meilleure prise en compte des transferts en nature dans les publications universitaires du WIL. D'autres instituts statistiques tels que celui des Pays-Bas ont depuis repris la méthode de distribution des dépenses en nature et l'ensemble du cadre de distribution des comptes *(Bruil et al., 2022)*.

▶ Le cadre général de distribution des revenus et transferts à partir de données individuelles

Dans le cadre de l'estimation des CND par l'Insee, la distribution des revenus et des transferts est principalement réalisée grâce au modèle de microsimulation Ines et la base de données de l'enquête Revenus fiscaux et sociaux (ERFS) ***(figure 4)***. Cette dernière est constituée à partir des données administratives fiscales et sociales et de l'enquête emploi. Le modèle Ines est le principal outil de distribution en raison du grand nombre de revenus et de transferts qu'il contient. Il permet de distribuer entre les ménages l'ensemble des prélèvements et prestations monétaires (cotisations sociales, impôts directs et indirects, minima sociaux, prime d'activité, etc.) et certaines prestations en nature comme le chèque énergie.

Figure 4 - Les sources mobilisées



- *ERFS : Enquête Revenus Fiscaux et Sociaux*
- *BDF : Enquête Budget De Famille*
- *HVP : Enquête Histoire de Vie et Patrimoine*
- *OMAR : Outil de Microsimulation pour l'Analyse des Restes à charge*
- *ENRJ : ENquête sur les Ressources des Jeunes*
- *DADS : Déclaration Annuelle de Données Sociales*
- *Fidéli : Fichier démographique sur les logements et les individus*

> ***Une source unifiée permet de meilleures estimations lorsqu'un nombre élevé de revenus et transferts sont distribués.***

Il est alimenté principalement par l'ERFS, mais aussi par d'autres bases de données afin de simuler le maximum de transferts. La documentation du modèle *(Fredon et Sicsic, 2020)*, regroupant notamment les écarts aux données externes agrégées (nombre de ménages concernés et total des transferts simulés) permet d'être transparent sur les résultats des comptes distribués.

En outre, une source unifiée permet de meilleures estimations lorsqu'un nombre élevé de revenus et transferts sont distribués. En effet, une base de données unique permettrait, par exemple, d'éviter des hypothèses d'appariement statistique, mécaniquement imparfaites ; une trop grande diversité de sources impliquerait d'avoir des estimations bruitées. Cela serait d'autant plus dommageable puisqu'il s'agit de croiser des variables et de les distribuer à des catégories de ménages. Les CND privilégient ainsi la base du modèle Ines qui intègre par construction une grande information sur les ménages qui la composent.

Chaque transfert est affecté à un individu selon des hypothèses d'incidence fiscale à partir des informations fournies par le modèle Ines. Par exemple, les cotisations sociales sont attribuées et donc payées par les salariés, et les dividendes sont affectés et payés par les actionnaires. Les revenus et transferts inclus dans l'ERFS et Ines sont ensuite calés sur des données de la comptabilité nationale grâce au tableau de correspondance établi par le groupe d'experts sur les inégalités et la redistribution. Lors de cette étape, il est nécessaire de combler les écarts de couverture des données individuelles avec le champ complet de la comptabilité nationale. Par construction, notamment en raison du plan de sondage des enquêtes ménages à l'Insee, l'ERFS couvre 93,5 % de la population française[6]. Afin de combler ce manque, les revenus et transferts de la population hors champ d'enquête sont distribués en reproduisant la distribution connue dans l'ERFS. Cette hypothèse est satisfaisante dans la mesure où la couverture des revenus ou des principaux prélèvements par l'ERFS s'établit majoritairement autour de 90 %.

À titre illustratif, le modèle Ines permet de simuler, et donc de distribuer, environ 430 Mds € de cotisations sociales, contre un total de 480 Mds € dans les comptes nationaux, soit environ 90 %. Cette couverture partielle est liée d'abord à la différence de champ et dans une moindre mesure à une simulation imparfaite des cotisations. Les 10 % non simulés par le modèle Ines sont, selon l'hypothèse retenue, distribués de la même manière que les 90 % simulés par le modèle[7].

L'ensemble de ces travaux permet d'aboutir à une large base de données ménages, incluant les revenus primaires et les transferts dont la somme est égale à la valeur établie par la comptabilité nationale, et de pouvoir distribuer ces revenus et transferts selon différentes variables ***(encadré 3)***.

6 Les DOM, les logements non ordinaires ou les ménages dont la personne de référence est étudiante ou déclare un impôt sur le revenu négatif ne sont pas couverts.

7 Cette opération est réalisée à un niveau plus fin de cotisations. L'écart le plus important entre les montants simulés par le modèle Ines et les comptes nationaux porte sur les cotisations de régimes particuliers de protection sociale (qui représentent 2 % du RNN) puisque ces dernières ne sont pas simulées par Ines. Les autres cotisations sont bien simulées et donc l'hypothèse est moins forte. La distribution de ces cotisations connues est répliquée pour la catégorie des régimes particuliers de protection sociale.

▶ Compléter les données individuelles afin de distribuer l'ensemble du revenu

Certains transferts ne sont toutefois pas présents dans les données de l'ERFS ou du modèle Ines par manque d'informations. C'est notamment le cas de l'impôt sur les sociétés (IS) et des profits non distribués des entreprises. Les distribuer nécessite donc des hypothèses d'imputation ou de simulation. Une première hypothèse se fonde sur la notion de revenus tirés de la propriété des entreprises et suppose de les distribuer comme les dividendes versés (variable présente dans les données de l'ERFS) en suivant la règle d'incidence évoquée précédemment. En effet, les entreprises étant détenues par les actionnaires, l'hypothèse revient à leur attribuer les profits réalisés et l'IS payé par les entreprises. Idéalement, il serait souhaitable de directement lier les revenus fiscaux des individus aux caractéristiques et résultats des entreprises qu'ils possèdent mais cela nécessite des travaux plus ambitieux[8] que les estimations initiales des comptes nationaux distribués. Des travaux de la statistique publique sont en cours afin de tenir compte des actionnaires détenteurs d'entreprises avec une importante valorisation boursière mais ne versant pas de dividendes. D'autres hypothèses de redistribution peuvent être envisagées, fondées sur les revenus du patrimoine au sens large ou le patrimoine détenu par exemple : elles modifient toutefois peu le profil redistributif global *[Piketty, Saez et Zucman, 2018]*. Certains transferts spécifiques comme les droits de mutation ou des taxes à faible rendement, ou encore les activités culturelles et associatives, d'un montant moindre et non mesuré par les données individuelles, sont distribués selon le profil de transferts de la même catégorie comptable mais dont le profil est connu *via* les données microéconomiques[9].

> ***La santé, l'éducation et les dépenses collectives font ainsi l'objet d'une attention particulière.***

Enfin, un point important mis en évidence par ces analyses et qui n'avait pas été traité avec précision par certains travaux universitaires est la prise en compte précise des services publics. La santé, l'éducation et les dépenses collectives font ainsi l'objet d'une attention particulière du fait : (i) de l'importance des montants en jeu, (ii) des enjeux méthodologiques et conceptuels particuliers qu'ils soulèvent, (iii) de la spécificité des données mobilisées, différentes de celles l'ERFS/Ines, (iv) de leur caractère novateur (auparavant, la littérature n'a jamais cherché à les distribuer conjointement à d'autres transferts). En effet, de nombreux travaux ont estimé l'effet redistributif de la santé ou de l'éducation, de façon séparée, mais aucune contribution n'a estimé l'effet redistributif de l'ensemble des transferts en nature et des dépenses collectives.

8 Il s'agit notamment de rapprocher des sources administratives sur les ménages et les entreprises ainsi que les informations relatives à la propriété des entreprises.

9 C'est le cas de certains transferts d'action sociale, hors allocation personnalisée d'autonomie et complément de mode de garde qui sont simulés avec Ines et qui sont distribués comme les prestations familiales, ou les taxes spéciales sur les conventions d'assurance qui sont distribuées comme les taxes sur les primes d'assurance (simulées avec *Ines*).

▶ Mieux prendre en compte les services publics : l'innovation sur les transferts en nature et dépenses collectives

Du fait de leur importance dans le revenu national net (RNN) en France (30 %), la prise en compte des dépenses en nature et de consommation collective dans la redistribution est cruciale. Que ce soient les dépenses de santé prises en charge par la Sécurité sociale ou bien la gratuité de l'enseignement primaire et secondaire, les transferts en nature sont quotidiens dans la vie des ménages en France. En 2018 par exemple, les dépenses de consommation collective représentent 10 % du RNN et 4 130 euros en moyenne annuelle par unité de consommation (UC). Concernant les prestations en nature, les principales dépenses, celles de santé et d'éducation, représentent respectivement 9 % et 5 % du RNN, soit 3 950 euros et 2 260 euros en moyenne par UC ***(figure 5)***. À l'inverse, aux États-Unis, l'école privée est courante et il n'existe pas de Sécurité sociale comme en France : les dépenses de santé et d'éducation sont principalement financées directement par les ménages. Ignorer les services publics éducatifs et de santé conduirait à sous-estimer la redistribution dans les pays européens où ces services sont particulièrement étendus.

> ***Compte tenu des montants associés à ces transferts, l'enjeu est donc de valoriser rigoureusement la distribution des services publics.***

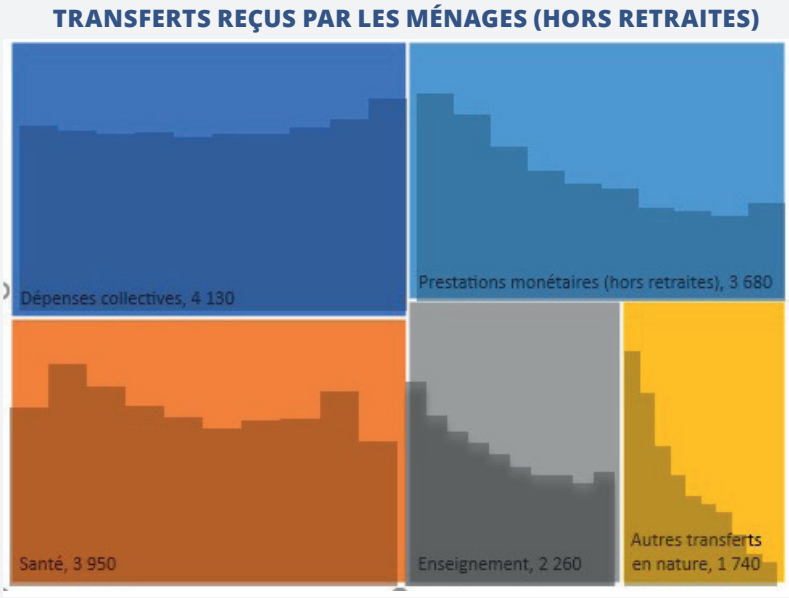
Compte tenu des montants associés à ces transferts, l'enjeu est donc de valoriser rigoureusement la distribution des services publics. Sans information précise sur les individus, des hypothèses doivent être formulées concernant les distributions de ces dépenses : par exemple, une distribution semblable à celles des revenus primaires observés, de façon uniforme *(Piketty, Saez, Zucman, 2018 ; Bozio et al., 2020)*, ou encore en utilisant des données agrégées *(OCDE, 2013)*. Dans les premières publications des comptes distribués de l'Insee en France *(Accardo et al., 2021, André, Germain et Sicsic, 2023)*, ces dépenses ont été simulées dans le détail en utilisant des données microéconomiques complémentaires.

Les **dépenses de santé** sont distribuées en fonction des remboursements de l'assurance maladie obligatoire, et celles des complémentaires santé en utilisant le modèle Ines-OMAR (Outil de microsimulation pour l'analyse des restes à charge) de la Drees[10] *[Lardellier et al., 2012]*. Ce modèle s'appuie sur l'enquête Santé et protection sociale (ESPS-EHIS), appariée aux données administratives fournissant des données détaillées de consommation de soins (Système national de données de santé, SNDS), ainsi que sur l'enquête de la Drees sur les contrats les plus souscrits auprès des organismes complémentaires et l'enquête Statistique sur les ressources et les conditions de vie (SRCV), notamment son module concernant la couverture complémentaire santé et l'état de santé perçu. Le modèle est rapproché du modèle Ines, ce qui permet de ventiler les montants de remboursement de l'assurance maladie selon le niveau de vie des individus ou d'autres caractéristiques (catégories socio-professionnelles, diplôme, âge, etc.), et d'inclure le financement de la branche maladie de la Sécurité sociale. Avec cette méthode de distribution des dépenses de santé, les transferts reçus par les ménages en fonction des revenus sont quasi stables ***(Figure 5)***.

10 La Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) est le service statistique ministériel dans les domaines de la santé et du social.

Les **dépenses d'éducation** sont distribuées en plusieurs étapes. La première consiste à multiplier le coût moyen par élève de l'élémentaire au collège issu du compte de l'éducation, par le nombre d'enfants concernés, à partir de l'âge et du nombre d'enfants renseignés dans l'ERFS. Les enfants et étudiants de plus de 14 ans présents dans le ménage indiquent leur type de formation de façon précise dans l'ERFS, ce qui permet de distinguer le lycée général

Figure 5 - Montants (en euros par unité de consommation) et profils des transferts reçus par les ménages



Lecture : *Les dépenses d'enseignement représentent en moyenne 2 260 euros par unité de consommation (UC). Avec cette méthode de distribution des dépenses d'éducation, les transferts reçus par les ménages décroissent en fonction des revenus. Les ménages les plus modestes reçoivent plus en moyenne par UC, du fait notamment d'un nombre d'enfants plus élevé.*

Source : *Insee, Comptes Nationaux distribués 2018, auteurs.*

Note : *La surface des rectangles de couleur est proportionnelle aux montants des transferts qu'ils représentent.*

ou professionnel, les sections de technicien du supérieur (STS), les classes préparatoires aux grandes écoles (CPGE) et l'université. Les coûts moyens de chaque formation sont utilisés pour déterminer les dépenses d'éducation reçues par chaque individu de façon relativement précise. Les étudiants non cohabitants sont rattachés aux ménages de leurs parents à partir de l'enquête sur les ressources des jeunes (ENRJ) réalisée par l'Insee et la Drees en 2014 afin de prendre en compte les transferts intrafamiliaux. La non prise en compte de ces transferts pourrait en effet générer un surplus de dépenses en bas de la distribution, du fait d'un classement erroné des étudiants de famille aisés en bas de la distribution ***(Figure 5)***.

Les **dépenses de consommation collective** sont séparées en deux. Les dépenses de consommation collective dites « localisables » (à l'inverse de la défense ou des affaires étrangères par exemple) sont distribuées à partir de la masse salariale des fonctionnaires concernés (hors hôpital et enseignement qui ont été comptabilisés pour les dépenses de santé et d'éducation). À partir des déclarations annuelles de données sociales (DADS) puis des Déclarations sociales nominatives (DSN), les services publics sont localisés pour chaque bassin de vie. Le ratio entre la masse salariale de ces fonctionnaires et le nombre d'individus est affecté à chaque ménage de l'ERFS afin de mesurer les dépenses collectives sur son bassin de vie. Cette méthode conduit à un profil relativement uniforme selon les niveaux de vie : cependant, les dépenses sont un peu plus élevées pour les plus modestes et les plus aisés (qui vivent davantage dans les grandes villes et disposant de services publics plus importants). D'autre part, les **dépenses de consommation collective** d'attribution nationale comme la défense, les affaires étrangères et la fonction publique de l'État des administrations générales sont distribuées forfaitairement, ces dépenses bénéficiant à l'ensemble de la population. Selon *Accardo* et *al., (2021),* distribuer ces dépenses collectives « nationales » proportionnellement aux revenus conduit à une baisse relative des revenus de 500 à 800 euros par unité de consommation (suivant le revenu pris en compte pour la redistribution de ces dépenses) pour les 10 % les plus modestes et une hausse de 1 000 à 1 500 euros pour les 10 % les plus aisés. Cet effet n'est pas neutre mais demeure contenu à l'échelle du profil redistributif élargi, car les dépenses d'attribution nationale ne représentent que 20 % environ des dépenses collectives. L'impact serait nettement plus important si l'hypothèse de proportionnalité était retenue pour l'ensemble des dépenses collectives.

▶ Quelques leçons sur la redistribution

Les comptes nationaux distribués apportent un nouveau regard sur les inégalités et la redistribution en France. Ils constituent un outil complémentaire aux dispositifs existants, permettant de mieux appréhender qui reçoit et qui verse quoi en France. Il est toutefois important de garder à l'esprit que les comptes nationaux distribués s'appuient sur des concepts abstraits, éloignés du ressenti des individus. Le revenu avant transferts, c'est-à-dire « primaire », est par exemple plus élevé que ce que les individus perçoivent réellement sur leurs comptes en banque. Il comprend non seulement les cotisations employeurs mais également les profits non distribués des entreprises, les loyers imputés aux propriétaires et certaines taxes indirectes sur la production et la consommation selon les hypothèses adoptées. De même, les transferts sociaux en nature et les services publics rendus par les dépenses collectives sont gratuits ou quasi gratuits et diffèrent ainsi de la réalité monétaire que représentent les prestations monétaires comme les allocations familiales par exemple.

Cette valorisation monétaire des services publics définit un revenu après redistribution ainsi élargie. Celui-ci ne correspond pas à une réalité monétairement tangible pour les ménages. Ces quantifications sont essentielles pour analyser avec rigueur le caractère redistributif du système socio-fiscal, mais ce travail reflète une valorisation comptable, et non un montant que les ménages peuvent effectivement percevoir.

Ces quantifications sont essentielles pour analyser avec rigueur le caractère redistributif du système socio-fiscal, mais ce travail reflète une valorisation comptable, et non un montant que les ménages peuvent effectivement percevoir.

Le premier exercice grand public des comptes nationaux distribués en fonction du niveau des ménages a porté sur l'année 2018 *(Accardo et al., 2021)*. La réduction des inégalités liée à la redistribution est à un niveau deux fois plus important dans l'approche élargie que dans l'approche monétaire usuelle. Le caractère redistributif du système socio-fiscal français provient avant tout des transferts en nature, comme l'éducation, la santé et le logement, qui contribuent pour 50 % à la réduction des inégalités. Les services publics collectifs contribuent également de façon significative à la réduction des inégalités par leur distribution relativement uniforme dans la population.

Dans l'exercice sur l'année 2019 *(André et al., 2023),* des nouveaux résultats sont présentés selon différentes variables : l'âge, le diplôme, la configuration familiale, le genre et ou encore la catégorie socio-professionnelle. Environ 60 % des individus sont bénéficiaires nets de la redistribution, au sens où ils reçoivent plus qu'ils ne contribuent. Ces bénéficiaires nets de la redistribution élargie sont surtout les retraités et les plus modestes, mais également les familles avec enfants, les ménages moins diplômés et les ouvriers dans une moindre mesure. À l'inverse, les personnes aisées, sans enfant, urbains (et notamment vivant dans l'agglomération parisienne), en bonne santé ou encore entre 40 ans et 60 ans reçoivent en moyenne moins qu'elles ne versent de prélèvements. Les comptes distribués sont également exploités pour étudier la redistribution selon la catégorie de commune et leur type d'aires urbaines *(André, 2022)*.

▶ Perspectives

Les comptes distribués tels que construits par l'Insee sont appelés à se développer car ils répondent à une demande sociale forte, exprimée notamment dans les travaux du Cnis[11] mais aussi par l'intermédiaire des travaux de recherche qui contribuent à l'élargissement de la notion d'inégalités de revenus. Les comptes économiques distributionnels ont ainsi vocation à intégrer les normes internationales de comptabilité nationale : des travaux sont en cours à ce sujet sous l'égide de l'Organisation des Nations unies (ONU) et concernent la décomposition du secteur des ménages par catégories (UN-STATS).

11 Le Conseil national de l'information statistique (Cnis) assure la concertation entre les producteurs et les utilisateurs de la statistique publique.

À l'Insee, des travaux visent à développer des séries temporelles de comptes distribués afin de pouvoir distribuer les fruits de la croissance (du revenu national, très proche de la croissance du PIB) entre catégories de ménages, tout en assurant une comparabilité dans le temps et dans l'espace. Cela pourra permettre de déterminer les principaux bénéficiaires de la croissance sur différents périodes (annuelles ou entre deux dates, une décennie par exemple). Une mise en production annuelle des comptes distribués est prévue, à la fois sous la forme privilégiée par le cadre international, à savoir la distribution des agrégats du secteur des ménages *(Accardo et Billot, 2020)* mais aussi sous la forme élargie des comptes nationaux distribués tels que présentés dans ce dossier. Pour intégrer ces méthodes aux outils de la comptabilité nationale, l'Insee a décidé d'approfondir l'expertise technique. Une analyse rétrospective permettra d'examiner la temporalité adéquate de fourniture de tels comptes, en aval de la statistique sociale. En outre, une attention particulière sera apportée à la lisibilité et à la cohérence des messages de l'Institut concernant les inégalités et leurs évolutions. La mise en production de ces comptes distribués permettra de les pérenniser et d'assurer des publications régulières. Des perspectives ambitieuses de plus long terme viseraient à élaborer des distributions trimestrielles des inégalités *(Blanchet, Saez et Zucman, 2021)*. Cela se heurte pour l'instant à de nombreuses difficultés méthodologiques, au premier rang desquelles figure l'enjeu d'utilisation des données administratives infra annuelles sur les revenus et les transferts.

Pour mener à bien ces projets et améliorer la qualité des comptes distribués, il pourrait être envisagé de compléter le modèle Ines en mobilisant les sources Fidéli, Filosofi et l'EDP-santé[12] afin d'améliorer la connaissance aux extrémités de la distribution[13] et des dépenses de santé. Un projet en cours entre l'Insee et la DGFiP vise à rapprocher les données ménages et les données entreprises afin de mieux mesurer l'incidence des taxes sur la production (IS notamment) et de formuler moins d'hypothèses d'imputation sur la distribution des profits non distribués. D'autres instituts nationaux mènent également des travaux afin d'élargir le champ usuel de l'analyse des inégalités, en intégrant notamment les transferts en nature. Il pourrait aussi être envisagé d'améliorer les informations disponibles sur l'éducation avec des données de dépenses localisées pour les enfants et les étudiants associées au ménage des parents, et de s'appuyer sur la DSN afin d'améliorer la distribution des dépenses publiques collectives locales. Enfin, dans le prolongement des travaux déjà produits par l'Insee sur la décomposition du revenu national, une prochaine étape pourrait être la constitution d'un compte distributionnel de patrimoine des ménages en clarifiant la cohérence entre les concepts et données de la comptabilité nationale, les données fiscales et l'enquête Patrimoine.

12 Fidéli (Fichier démographique sur les logements et les individus) est un répertoire statistique de données fiscales, sociales et de revenus *(Lamarche et al., 2021)*. Filosofi (Dispositif sur les revenus localisés sociaux et fiscaux) est une base de données sur les revenus. L'EDP-santé (EDP *Robert-Bobée et al.,* 2021) est un enrichissement de l'échantillon démographique permanent par les données du système national des données de santé (SNDS).

13 Un inconvénient d'utiliser l'ERFS et le modèle Ines concerne la granularité des résultats, du fait de l'aléa de sondage qui peut être théoriquement important aux extrémités de la population en raison notamment de la grande variance des revenus. Ainsi, les analyses par centième doivent faire l'objet d'une attention particulière. Dans le cadre du rapport d'experts (Insee, 2021), les statistiques sur le top 1 % dans les comptes distribués ont été comparées entre la méthode de l'Insee basée sur l'ERFS et la méthode des équipes de la *World income Database* (WID) qui se base sur des données exhaustives : les résultats se sont révélés très proches, à champ et méthode identiques.

▶ Bibliographie

- ACCARDO Jérôme, BELLAMY Vanessa, CONSALES Georges, FESSEAU Maryse, LE LAIDIER Sylvie et RAYNAUD Émilie, 2009. Les inégalités entre ménages dans les comptes nationaux : une décomposition du compte des ménages. In : *L'économie française*, coll. « Insee Référence » [en ligne]. 25 juin 2009. pp. 77-101. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.insee.fr/fr/statistiques/1372352?sommaire=1372361*.

- ACCARDO Jérôme, 2020. Supplementing GDP: Some Recent Contributions from Official Social Statistics. In : *Economie et Statistique / Economics and Statistics*, n°517-518-519. 8 octobre 2020 [en ligne]. pp. 25-39. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://doi.org/10.24187/ecostat.2020.517t.2016*.

- ACCARDO Jérôme et BILLOT Sylvain, 2020. Plus d'épargne chez les plus aisés, plus de dépenses contraintes chez les plus modestes. *Insee Première n° 1815*, [en ligne]. Septembre 2020. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.insee.fr/fr/statistiques/4764600.*

- ACCARDO Aliocha, ANDRÉ Mathias, BILLOT Sylvain, GERMAIN Jean-Marc et SICSIC Michaël, 2021. Réduction des inégalités : la redistribution est deux fois plus ample en intégrant les services publics. In : *Revenus et patrimoine des ménages*, coll. « Insee Références » [en ligne]. 27 mai 2021, pp 77-96. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.insee.fr/fr/statistiques/5371275?sommaire=5371304*.

- AMAR Élise, BEFFY Magali, MARICAL François et RAYNAUD Émilie, 2008. Les services publics de santé, éducation et logement contribuent deux fois plus que les transferts monétaires à la réduction des inégalités de niveau de vie, In « Vue d'ensemble — Redistribution », France, portrait social, coll. « Insee Références ». [en ligne]. 1er novembre 2008. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.insee.fr/fr/statistiques/1372948?sommaire=1372956.*

- AUSTRALIAN BUREAU OF STATISTICS, 2019. *Australian National Accounts: Distribution of Household Income, Consumption and Wealth*, 2003-04 to 2017-18. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.abs.gov.au/statistics/economy/national-accounts/australian-national-accounts-distribution-household-income-consumption-and-wealth/2003-04-2017-18.*

- ANDRÉ Mathias, 2022. Les prélèvements obligatoires au regard des enjeux redistributifs. In : *rapport particulier n°3 pour le Conseil des prélèvements obligatoires*. [en ligne]. Février 2022. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.ccomptes.fr/sites/default/files/2022-02/20220209-rapport-particulier-redistribution.pdf.*

- ANDRÉ Mathias, GERMAIN Jean-Marc et SICSIC Michaël, 2023. 'Do I get my money back?': A Broader Approach to Inequality and Redistribution in France With a Monetary Valuation of Public Services. In : *Documents de travail*. N°2023-07 [en ligne]. 8 mars 2023. Insee. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.insee.fr/en/statistiques/6964929.*

- BELLAMY Vanessa, CONSALES Georges, FESSEAU Maryse, LE LAIDIER Sylvie et RAYNAUD Émilie, 2009. Une décomposition du compte des ménages de la comptabilité nationale par catégorie de ménage en 2003, In : *Documents de travail*. [en ligne]. N° G2009/11. [en ligne]. 1 novembre 2009. Insee. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.insee.fr/fr/statistiques/1380884*.

- BILLOT Sylvain et BOURGEOIS Alexandre, 2019. Quelle(s) mesure(s) du pouvoir d'achat ?, In : *L'économie française – Comptes et dossiers*, coll. « Insee Références ». [en ligne]. 28 juin 2019 [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://www.insee.fr/fr/statistiques/4181441?sommaire=4180914*.

- BLANCHET Thomas, SAEZ Emmanuel and ZUCMAN Gabriel, 2022. *Real-Time Inequality (July 2022). NBER Working Paper No. W30229* [en ligne]. 12 juillet 2022 [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4159144#*.

- BOZIO Antoine, GARBINTI Bertrand, GOUPILLE-LEBRET Jonathan, GUILLOT Malka et PIKETTY Thomas, 2020. Predistribution vs. Redistribution: *Evidence from France and the U.S World Inequality Lab – Working Paper N° 2020/22*. [en ligne]. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://wid.world/document/predistribution-vs-redistribution-evidence-from-france-and-the-u-s/*.

- BRUIL Arjan, VAN ESSEN Céline, LEENDERS Wouter, LEJOUR Arjan, MOHLMANN Jan and RABATÉ Simon, 2022. *Inequality and Redistribution in the Netherlands. CPB Discussion paper* [en ligne]. [consulté le 21 mars 2023]. Disponible à l'adresse suivante : *https://wouterleenders.eu/Bruiletal2022DP.pdf*.

- EUROSTAT, 2018. *Income and consumption: social surveys and national accounts*. In : *site de Eurostat*. [en ligne] [Consulté le 21 mars 2023]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/web/experimental-statistics/ic-social-surveys-and-national-accounts*.

- FREDON, Simon et SICSIC, Michaël, 2020. Ines, le modèle qui simule l'impact des politiques sociales et fiscales. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee. N° N4, pp. 42-61. [Consulté le 21 mars 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497070?sommaire=4497095*.

- INSEE « Rapport du groupe d'experts sur la mesure des inégalités et de la redistribution », sous la direction de J.-M. Germain (rapporteurs : André, M. et Blanchet, T.), Insee méthodes n°138, février 2021. [Consulté le 3 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5020893*.

- LAMARCHE, Pierre et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 28-46. [Consulté le 22 juin 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398683/courstat-6-art-2.pdf.*

- LARDELLIER Rémi, LEGAL Renaud, RAYNAUD Denis et VIDAL Guillaume, 2012. Un outil pour l'étude des dépenses de santé et des « restes à charge » des ménages : le modèle Omar. In : *Économie et statistique. [en ligne].* 30 novembre 2012. Insee. N°450, pp. 47-77. [Consulté le 21 mars 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1377828?sommaire=1377832*.

- LE LAIDIER Sylvie, 2009. Les transferts en nature atténuent les inégalités de revenus. *[en ligne]. Novembre 2009.* Insee Première n° 1264. *[Consulté le* 21 mars 2023*]. Disponible à l'adresse : https://www.bnsp.insee.fr/ark:/12148/bc6p0702087.r=Les%20transferts%20en%20nature%20att%C3%A9nuent%20les%20in%C3%A9galit%C3%A9s%20de%20revenus?rk=21459;2*.

- PIKETTY Thomas, SAEZ Emmanuel et Zucman Gabriel, 2018. Distributional National Accounts: Methods and Estimates for the United States. In : Quarterly Journal of Economics, 133 (2), pp. 553-609, [en ligne]. 10 octobre 2017. [Consulté le 21 mars 2023]. Disponible à l'adresse : *https://academic.oup.com/qje/article/133/2/553/4430651*.

- ROBERT-BOBÉE, Isabelle et GUALBERT, Natacha Gualbert, 2021. L'échantillon démographique permanent : en 50 ans, l'EDP a bien grandi ! In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 47-63. [Consulté le 22 juin 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398685/courstat-6-art-3.pdf.*

- STATISTICS CANADA, 2021. Distributions of Household Economic Accounts, estimates of asset, liability and net worth distributions, 2010 to 2021, technical methodology and quality report. [en ligne]. 3 août 2022. [Consulté le 21 mars 2023]. Disponible à l'adresse : *https://www150.statcan.gc.ca/n1/pub/13-604-m/13-604-m2022002-eng.htm.*

- STATISTICS NETHERLANDS, 2014. *Measuring Inequalities in the Dutch Household Sector.*

- STIGLITZ Joseph, SEN Amartya et FITOUSSI Jean-Paul, 2009. Rapport de la Commission sur la mesure des performances économiques et du progrès social. Éditions Odile Jacob, 13 novembre 2009. ISBN 978-2-7381-9380-3.

- VANOLI André, 2002. Une histoire de la Comptabilité nationale. Paris, La Découverte, 2002, p 656.

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.

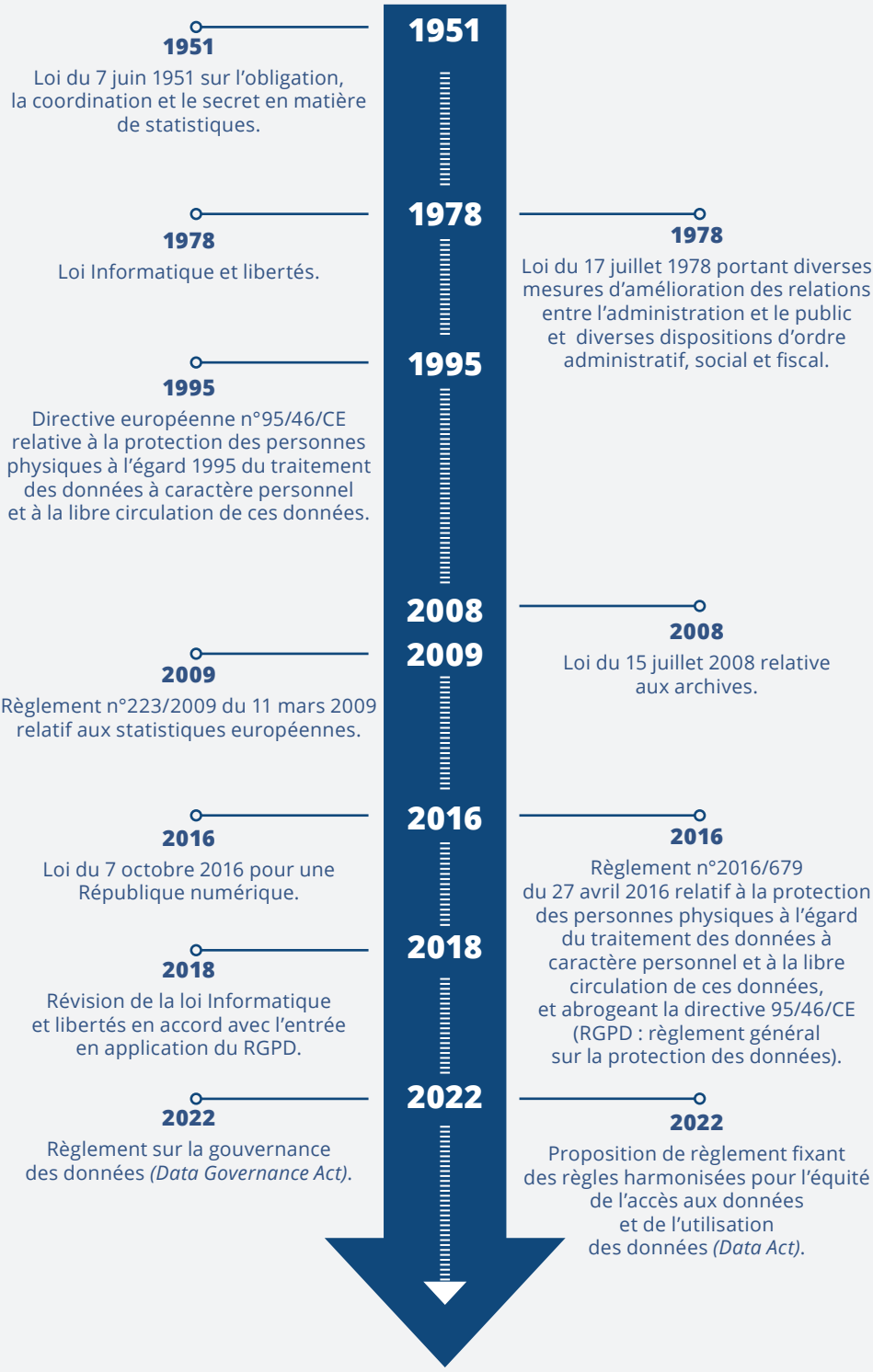
Toutes ces mesures, pour développées et sophistiquées qu'elles soient, sont insuffisantes sans la maîtrise du facteur humain, qui passe par les moyens que le SSP consacre à la formation et à la sensibilisation de ses agents aux enjeux de la confidentialité.

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*.

Le Code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers

Yves-Laurent Bénichou[*], *Lionel Espinasse*[**] *et Séverine Gilles*[***]

Les appariements de fichiers permettent d'accroître considérablement les possibilités d'études des phénomènes économiques et sociaux. Le Code statistique non signifiant (CSNS) a été défini par la loi pour une République numérique de 2016 afin de permettre la mise en œuvre d'appariements de fichiers à des fins statistiques en limitant l'usage du NIR (ou numéro de sécurité sociale), et garantir ainsi un niveau élevé de protection des données à caractère personnel. Le principe général est d'utiliser une clé d'appariement calculée à partir d'un chiffrement irréversible du numéro de sécurité sociale. Ce nouveau service offert par l'Insee aux organismes du service statistique public s'applique à une grande diversité de fichiers administratifs ou issus d'enquêtes. Une méthode innovante a été conçue pour identifier des personnes à partir de leurs traits d'identité. À l'issue du processus, la fiabilité de l'identification est mesurée par des indicateurs de qualité.

Les premières utilisations sont prometteuses. Le CSNS permet, par exemple, de contribuer à l'analyse de l'insertion des jeunes diplômés en facilitant l'appariement des données du système éducatif et du ministère du Travail. Il aide aussi à mesurer l'impact de la transition écologique selon les catégories de ménage en rapprochant les données du répertoire de véhicules et les informations sur les revenus.

File matching considerably increases the possibilities of studying economic and social phenomena. The Non-SignifIcant Statistical Code (CSNS) was defined by the Law for a Digital Republic of 2016 in order to allow the implementation of file matching for statistical purposes without using the NIR (or national insurance number), thus ensuring a high level of personal data protection. The general principle is to use a matching key calculated from an irreversible encryption of the national insurance number. This new service offered by INSEE to official statistical services applies to a wide variety of administrative or survey files. An innovative method has been developed to identify people on the basis of their identity. At the end of the process, the reliability of the identification is measured by quality indicators.

The first uses are promising. For example, the CSNS can contribute to the analysis of the integration of young graduates by facilitating the matching of data from the education system and the Ministry of Labour. It also helps to measure the impact of the ecological transition according to household categories by matching vehicle register data and income information.

* Expert en Data science, Unité SSP Lab, Insee
yves-laurent.benichou@insee.fr
** Adjoint à la cheffe du département de la Démographie, DSDS, Insee,
lionel.espinasse@insee.fr
*** À la date de la rédaction de l'article, cheffe de projet statistique (CSNS), DSDS, Insee
severine.gilles@insee.fr

Avec le développement de systèmes d'information performants dans de nombreux secteurs, les appariements[1] de fichiers deviennent un mode d'enrichissement des données très puissant. Ils permettent de mettre en relation des informations variées qui se trouvent dans des univers différents, et sont moins coûteux que de réaliser des enquêtes sur le terrain. Le Code statistique non signifiant (CSNS) a été défini par la loi pour une République numérique de 2016[2] pour permettre la mise en œuvre de ces appariements tout en préservant la confidentialité, en limitant l'usage du NIR[3].

> ***En reliant des informations collectées par des organismes différents, les appariements de fichiers permettent ainsi d'accroître considérablement les possibilités d'études des phénomènes économiques et sociaux.***

Deux exemples peuvent illustrer les enrichissements permis par les appariements : pour appréhender le devenir des étudiants du supérieur après leur formation, il est pertinent de croiser les données relatives à l'enseignement, détenues par le ministère de l'Enseignement supérieur avec les données d'emploi détenues par le ministère du Travail. Ou dans un autre registre, mettre en regard les données sur le parc de véhicules automobiles et leur consommation d'énergie avec le revenu de leurs propriétaires permet de déterminer l'impact de la hausse des prix des carburants sur les différentes catégories de population. En reliant des informations collectées par des organismes différents, les appariements de fichiers permettent ainsi d'accroître considérablement les possibilités d'études des phénomènes économiques et sociaux.

Pour apparier des fichiers, la situation la plus favorable est celle où l'on dispose d'un identifiant pour chaque individu, commun aux deux fichiers[4]. On peut ainsi associer les enregistrements correspondant aux mêmes individus et regrouper alors l'information les concernant issue des deux fichiers. Dans l'idéal, cet identifiant a un statut formel, est conservé dans un référentiel et permet de caractériser chaque individu sans ambiguïté et de manière unique. Le NIR, numéro d'identification au répertoire national d'identification des personnes physiques (RNIPP)[5] communément appelé « numéro de sécurité sociale » est un très bon exemple d'identifiant *(Espinasse et Roux, 2022)*.

Mais encore faut-il respecter certaines règles et notamment les attentes des citoyens relatives à la protection de leurs données personnelles. Avant 2018, pour répondre à cette attente, les appariements de fichiers sur la base du NIR pour les besoins de la statistique publique nécessitaient en particulier, pour chaque traitement, un décret en Conseil d'État après avis publié et motivé de la Commission nationale de l'informatique et des libertés (Cnil)[6]. Face à la croissance de la demande de données appariées pour les besoins

1 Les appariements constituent des interconnexions de sources de données statistiques tierces, fondées sur les données à caractère personnel et permettent de créer de nouveaux fichiers comprenant tout ou partie des variables de chacun des fichiers des sources d'origine.
2 Voir les fondements juridiques en fin d'article.
3 Le NIR est le numéro d'inscription au Répertoire national d'identification des personnes physiques (RNIPP), il est plus communément appelé « numéro de sécurité sociale ».
4 Il existe d'autres méthodes, notamment en rapprochant les traits d'identité *(voir infra),* mais celles-ci sont plus complexes à mettre en œuvre dans un processus standardisé et industrialisé.
5 Voir les fondements juridiques en fin d'article.
6 Voir les fondements juridiques en fin d'article.

de conception et d'évaluation des politiques publiques, il est devenu nécessaire d'imaginer une procédure plus simple juridiquement qui respecte le principe de minimisation des données et le statut particulier conféré au NIR par la loi Informatique et Libertés. L'idée a ainsi été émise de ne plus utiliser directement le NIR comme identifiant d'appariement, mais un identifiant non signifiant, le « NIR haché » conservant les propriétés techniques d'un identifiant tout en rendant impossible de revenir directement à l'identité des personnes.

▶ Une innovation de la loi pour une République numérique : le CSNS, un service rendu par l'Insee

> ***Le CSNS est un service de l'Insee ayant pour finalité de faciliter les appariements de fichiers de personnes physiques au sein du service statistique public (SSP), tout en garantissant un niveau de protection des données à caractère personnel plus élevé qu'avec l'usage du NIR.***

Dans ce contexte, un dispositif spécifique a été mis en place avec la création du Code statistique non signifiant (CSNS)[7]. Au-delà d'un simple code, le CSNS est aussi un service de l'Insee ayant pour finalité de faciliter les appariements de fichiers de personnes physiques au sein du service statistique public (SSP) garantissant un niveau de protection des données à caractère personnel plus élevé qu'avec l'usage du NIR.

La loi pour une République numérique précise que les institutions appartenant au service statistique public peuvent bénéficier du CSNS. Ainsi, ce service rendu par l'Insee s'adresse aussi aux Services statistiques ministériels (SSM). Les appariements peuvent être réalisés entre les fichiers de deux SSM, d'un SSM et de l'Insee ou de deux unités de l'Insee. Ces configurations peuvent être facilement étendues à trois ou quatre partenaires.

La prestation de l'Insee comporte une dimension technique pour l'identification des personnes et le chiffrement irréversible du NIR et une dimension organisationnelle avec la mise à disposition d'une application permettant aux utilisateurs de déposer leurs demandes et récupérer leurs résultats.

Tous les traitements réalisés sont enregistrés dans les registres des activités de traitement prévus par l'article 30 du Règlement général sur la protection des données (RGPD)[8] et doivent être rendus publics. Chaque traitement doit également être communiqué au Conseil national de l'information statistique (Cnis) qui pourra évaluer ce nouveau dispositif. Cette liste des traitements est diffusée publiquement sur le site du Cnis.

Le principe de minimisation des données fait aussi l'objet d'une attention particulière : seules les informations strictement nécessaires au calcul du CSNS sont échangées entre les organismes partenaires et l'Insee, puis l'Insee détruit sans délai les données qui lui ont été confiées dès que ce calcul est achevé.

7 Le CSNS est prescrit par l'article 34 de la loi pour une République numérique du 7 octobre 2016 et ses modalités opérationnelles sont précisées par un décret en Conseil d'État (n°2016-1930 du 28 décembre 2016) et par un arrêté du ministre chargé de l'économie (du 28 septembre 2020) *(voir Fondements juridiques).*

8 Voir les fondements juridiques en fin d'article.

Chiffrer le numéro de sécurité sociale : le CSNS en tant que code

Le CSNS a vocation à servir de multiples usages dans des domaines variés. Il ne s'agit pas de limiter son utilisation à l'appariement de quelques fichiers spécifiques ; il doit pouvoir être utilisé par tous les services statistiques ministériels, pour toutes leurs sources de données. Il prend ainsi appui sur un référentiel de population qui couvre toute la population vivant en France, le Répertoire national d'identification des personnes physiques (RNIPP) qui centralise l'ensemble des numéros de sécurité sociale.

Le principe général du CSNS est d'appliquer au numéro de sécurité sociale (NIR), une opération de chiffrement irréversible pour obtenir une clé d'appariement en garantissant l'impossibilité d'identifier individuellement les personnes concernées.

Chaque personne ayant un NIR unique, le calcul du CSNS donnera toujours le même résultat quel que soit le fichier sur lequel il est appliqué et permettra ainsi des appariements sans que l'on ait besoin de connaître l'identité des personnes.

Enfin, le calcul du CSNS repose sur un processus entièrement automatisé. Ce principe a été posé dès la conception du projet. Les propriétaires de fichiers ont accès à une application dédiée et sont autonomes pour faire leurs demandes et récupérer leurs résultats. Ils obtiennent des fichiers de CSNS produits sans intervention humaine, même pour le traitement des cas les plus complexes (voir *infra*). Cette automatisation présente l'avantage de rendre le coût de fonctionnement supportable pour l'Insee et de pouvoir ainsi rendre ce service gratuit, mais aussi de gagner en temps de traitement. En contrepartie, elle implique une standardisation du processus d'identification des personnes et du processus d'évaluation de la qualité des résultats. Elle conduit aussi à ce que le travail de préparation des données, propre aux spécificités de chaque fichier, relève de la responsabilité des demandeurs.

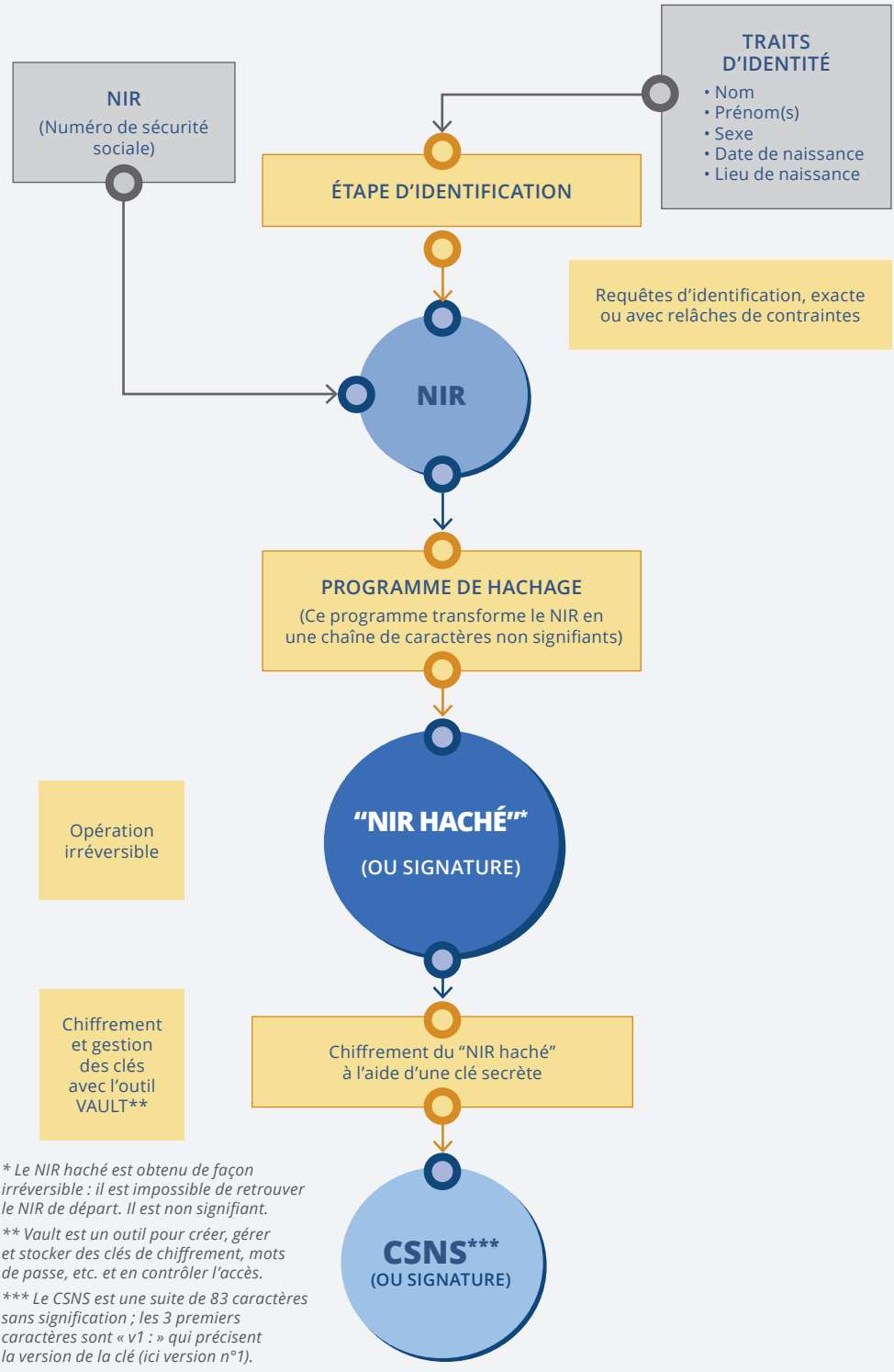
Deux possibilités pour s'adapter aux besoins

La richesse et la qualité des informations sur les identités des personnes sont variables selon les fichiers à apparier. Certains disposent déjà du NIR, d'autres ne l'ont pas mais rassemblent des traits d'identité comme le nom, les prénoms, le sexe, la date et le lieu de naissance. Deux **possibilités** sont alors offertes ***(figure 1)***. Le calcul du CSNS peut se faire soit à partir du NIR, soit à partir des traits d'identité.

Dans le premier cas, le calcul se résume simplement à chiffrer le NIR de manière irréversible. Dans le deuxième cas, le plus fréquent, une étape supplémentaire est nécessaire. Il faut d'abord retrouver le NIR des personnes à partir de leurs traits d'identité. Cette phase est appelée « étape d'identification ». Il suffit ensuite de chiffrer les NIR ainsi retrouvés. Cette deuxième façon de procéder est plus complexe à mettre en œuvre et restitue des résultats avec des niveaux de fiabilité qui dépendent de la qualité des données d'identité en entrée.

Le fonctionnement est le suivant : chaque propriétaire de fichier transmet au service CSNS de l'Insee les NIR ou les traits d'identité des individus de son fichier.

Figure 1 - Comment est obtenu le CSNS ?



** Le NIR haché est obtenu de façon irréversible : il est impossible de retrouver le NIR de départ. Il est non signifiant.*

*** Vault est un outil pour créer, gérer et stocker des clés de chiffrement, mots de passe, etc. et en contrôler l'accès.*

**** Le CSNS est une suite de 83 caractères sans signification ; les 3 premiers caractères sont « v1 : » qui précisent la version de la clé (ici version n°1).*

Cette transmission s'effectue de façon sécurisée, avec une application dédiée et accessible aux seuls demandeurs habilités du service statistique public. En retour, chaque demandeur reçoit un CSNS calculé pour chacun des individus de son fichier ***(encadré 1)***. Pour un même individu, le CSNS sera toujours le même, quelle que soit la source où figure cet individu et quelle que soit l'année où le CSNS est calculé. Avec cette clé unique et pérenne[9], les partenaires peuvent ensuite apparier leur fichier.

La technique de chiffrement répond à un haut niveau d'exigence de sécurité et comporte deux étapes. Le calcul du CSNS correspond d'abord à un hachage du NIR, programme qui transforme le NIR en une chaîne de caractères non signifiants de façon irréversible, puis à un chiffrement de ce « NIR haché » à l'aide d'une clé secrète. Cette double opération permet de garantir qu'il est impossible de retrouver le NIR à partir du CSNS. Même si la clé de chiffrement était malheureusement dévoilée, elle ne permettrait de revenir qu'au NIR haché mais pas au NIR.

▶ Encadré 1 : Le CSNS en pratique

Le service de calcul du CSNS à partir du NIR est ouvert depuis octobre 2021 et celui proposant une identification sur traits d'identité depuis octobre 2022. Il est réservé au service statistique public.

Pour y avoir accès, chaque service statistique doit au préalable signer avec l'Insee un contrat de sous-traitance qui précise les droits et obligations de chacun et indique les principes de fonctionnement du processus. Ce contrat est valable 5 ans. Une fois signé, chaque service peut effectuer autant de demandes qu'il le souhaite. Le service est gratuit.

La phase de calcul du CSNS n'est toutefois qu'une des phases du processus d'appariement. Elle s'inscrit dans une démarche plus générale où deux propriétaires de fichiers se rapprochent pour apparier leurs données, exploiter le résultat de cet appariement et le diffuser. L'Insee en tant qu'opérateur du CSNS offre un service de sous-traitance à ces propriétaires de fichiers pour la production d'une clé d'appariement commune, mais n'intervient pas sur les autres aspects.

En particulier, il convient qu'au moins un des propriétaires soit responsable de traitement au sens du RGPD et réalise toutes les démarches requises à ce titre, notamment la production d'une analyse d'impact relative à la protection des données (AIPD) si nécessaire.

La démarche de mise en œuvre d'un traitement CSNS comprend alors en général six étapes :

- établissement d'une convention entre les deux propriétaires de fichiers fixant les conditions de réalisation de l'appariement de leurs données et de leur utilisation ;
- déclaration du traitement par le ou les responsables de traitement ;
- inscription du traitement dans le programme de travail transmis au Cnis ;
- demande par chaque propriétaire d'un calcul de CSNS à l'Insee et restitution par l'Insee à chaque propriétaire du résultat du calcul avec des indicateurs de qualité ;
- appariements des données par les propriétaires de fichiers selon les modalités qu'ils auront définies dans leur convention ;
- conservation de ses données par chaque propriétaire selon des règles définies par la réglementation et rappelées dans le contrat de sous-traitance.

Les modalités pratiques pour effectuer une demande de calcul de CSNS sont simples. Après vérification des habilitations, le demandeur dépose son fichier de NIR ou de traits d'identité dans une application dédiée ouverte en ligne. Puis il reçoit en retour les CSNS et leurs indicateurs de qualité. Chaque demandeur peut réitérer ses demandes autant de fois qu'il le souhaite, notamment s'il apporte des améliorations à son fichier en entrée.

9 Les CSNS calculés sont valables 10 ans à compter de 2022, sauf si une faille de sécurité est détectée. Au bout de 10 ans (ou plus tôt en cas de faille), les clés de chiffrement des NIR seront modifiées. Le processus de renouvellement produira une table de correspondance entre les nouveaux et les anciens CSNS.

Après appariement, les CSNS doivent être conservés par chaque propriétaire de manière sécurisée et isolée dans un fichier qui ne comprend aucune variable socio-démographique, aucun NIR et aucun trait d'identité. Seul un numéro d'ordre permettra à l'avenir de faire à nouveau correspondre le CSNS avec les individus, pour de nouveaux appariements. Par ailleurs, la durée de conservation doit être proportionnée aux besoins actuels et éventuellement futurs.

Lorsque le demandeur ne dispose pas du NIR mais uniquement de traits d'identité, une étape d'identification est nécessaire avant de réaliser ce processus de hachage-chiffrement du NIR.

▶ Retrouver les NIR à partir des traits d'identité : l'étape d'identification

L'opération d'identification du NIR à partir de traits d'identité est effectuée avec un moteur spécifiquement développé pour le calcul du CSNS. Il existe déjà un processus d'identification au RNIPP, mais celui-ci est utilisé pour des besoins à finalité administrative (pour les services fiscaux par exemple). Dans ces configurations qui induisent des conséquences administratives sur la vie des personnes, aucune erreur n'est admise. L'identification n'est confirmée que si la correspondance entre les traits d'identité à vérifier et les traits d'identité du RNIPP est certaine. De ce fait, l'identification échoue parfois ; la recherche de l'exactitude de la correspondance se solde par un taux d'échec d'appariement parfois relativement élevé.

> ***Les besoins statistiques sont différents des besoins administratifs.***

Or, les besoins statistiques sont différents des besoins administratifs. Sur une population de plusieurs milliers ou centaines de milliers de personnes, quelques erreurs d'identification n'auront pas de conséquences importantes sur les résultats statistiques finaux (et aucune conséquence sur les individus eux-mêmes). Il peut donc être intéressant d'accepter quelques approximations dans la correspondance des traits d'identité si cela permet d'augmenter le taux d'identification, sous contrôle d'un taux d'erreur acceptable mais faible. Tout l'enjeu est ainsi de trouver le point d'équilibre entre maximiser l'identification et minimiser les erreurs. Ceci est d'autant plus important lorsque les sources statistiques utilisées sont des enquêtes. Les personnes enquêtées renseignent en général moins scrupuleusement leurs données d'identité dans une enquête que dans un formulaire administratif. A fortiori, les processus de lecture optique de questionnaires papier, comme dans le recensement de la population, peuvent ajouter de l'incertitude sur la qualité de la saisie des données d'identité. C'est pour cela qu'un moteur d'identification spécifique a été développé pour le CSNS.

La construction de ce moteur d'identification s'est appuyée en partie sur la théorie des appariements *(cf. Christen, 2012 ; Fellegi et al. 2014)*, mais en tenant compte de la spécificité du référentiel de population utilisé. La phase de préparation des données relève de la responsabilité de l'organisme demandeur et non de l'équipe CSNS de l'Insee, même si quelques actions sont nécessaires en début de processus de calcul afin d'adapter au mieux les données au fonctionnement du moteur.

Ensuite, le très gros volume du référentiel de population (le RNIPP comprend 130 millions d'occurrences) a conduit à concevoir trois étapes pour l'identification d'un NIR avec le souci d'optimiser les temps de traitement selon la difficulté des cas à traiter. Pour cela, le principe de traiter les cas simples avec des processus peu gourmands en temps, et de réserver les processus complexes aux cas qui le nécessitent vraiment, a été retenu.

Un enchaînement d'étapes, de la plus simple à la plus complexe

Un enchaînement de trois étapes est conçu pour retrouver un NIR, adaptées aux différents niveaux de complexité de la recherche. ***(figure 2)***.

La première étape est une requête dite « exacte ». Les éléments d'identification (nom, prénoms, date de naissance, code géographique du lieu de naissance) sont recherchés dans le RNIPP, de façon exacte. L'ensemble des prénoms doit être exact également. À cette étape, l'identification suit un principe fondamental : elle ne peut avoir lieu que si un seul écho[10] est candidat à l'identification. À noter également que les recherches sont aussi effectuées sur les anciens noms (lorsque les personnes en ont changé) ou les noms d'usage (marital souvent lorsqu'il figure au RNIPP).

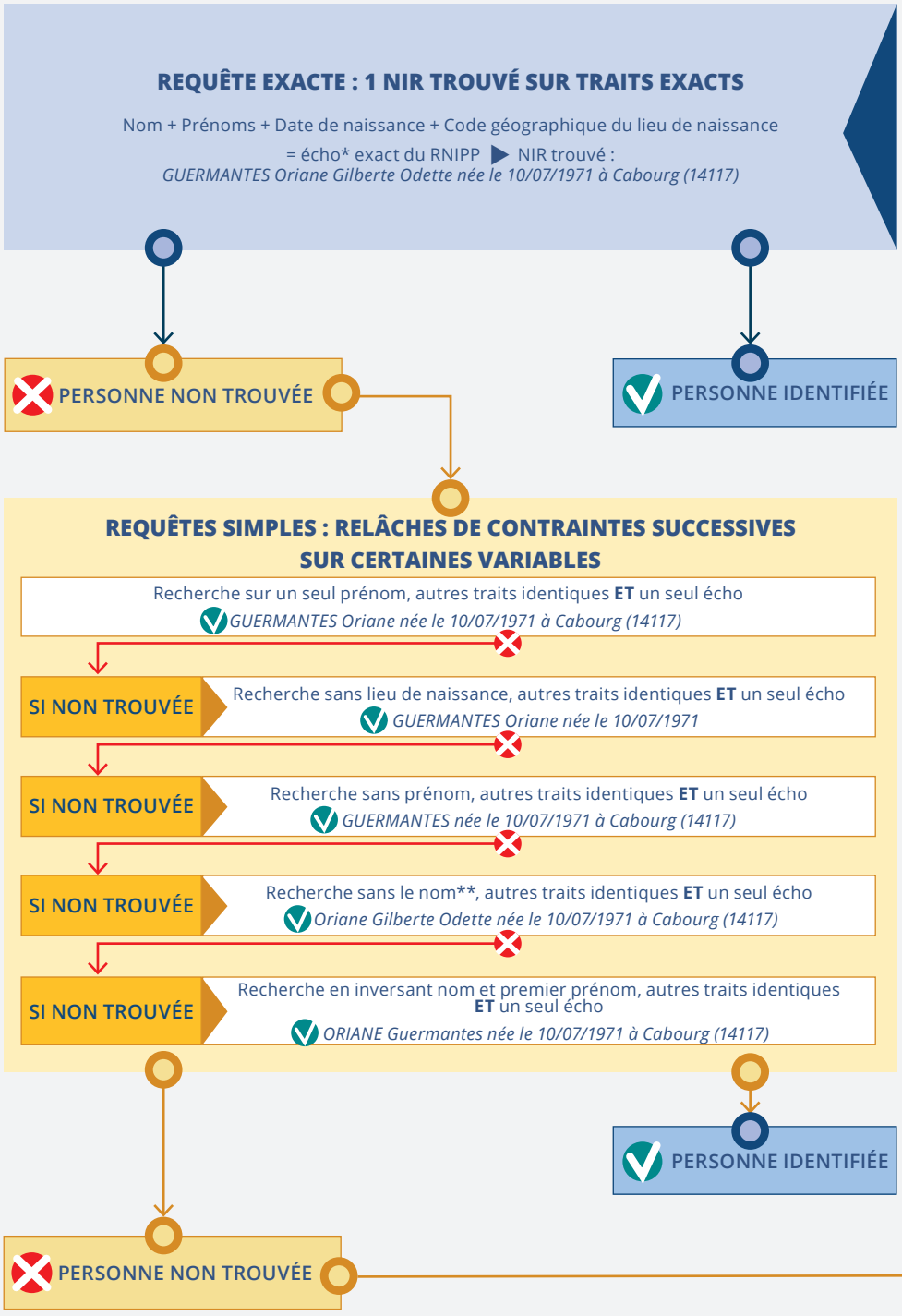
Vient ensuite une deuxième étape dite de « requêtes simples ». On autorise alors quelques « relâches de contraintes »[11] sur certaines variables. Là encore, ne sont retenus que les échos sans concurrent. Cinq assouplissements successifs des contraintes d'identification sont mis en œuvre :

- une relâche simple sur les prénoms : les éléments d'identification sont tous identiques sauf les prénoms pour lesquels on autorise une identification sur un seul et non sur tous ;
- une relâche simple sur le code géographique du lieu de naissance : les éléments d'identification sont tous identiques, à l'exception du code géographique du lieu de naissance ; pour les prénoms, l'exactitude est exigée sur le premier ;
- une relâche simple et totale sur les prénoms : les éléments d'identification sont tous identiques, à l'exception des prénoms ;
- une relâche simple et totale sur le nom : les éléments d'identification sont tous identiques à l'exception du nom ; pour les prénoms, l'exactitude est exigée sur le premier prénom. Cette relâche sur le nom peut paraître surprenante au premier abord, le nom semblant être l'élément déterminant d'une identité. Mais en fait, les tests ont montré que de nombreux échecs d'identification étaient dus à l'usage du nom marital dans les fichiers des utilisateurs alors que celui-ci pouvait être inconnu du RNIPP (le nom marital n'est pas une variable obligatoire du RNIPP). Cette requête permet ainsi d'identifier de nombreuses femmes mariées qui ne déclarent pas leur nom de naissance dans les enquêtes mais leur nom d'usage ;

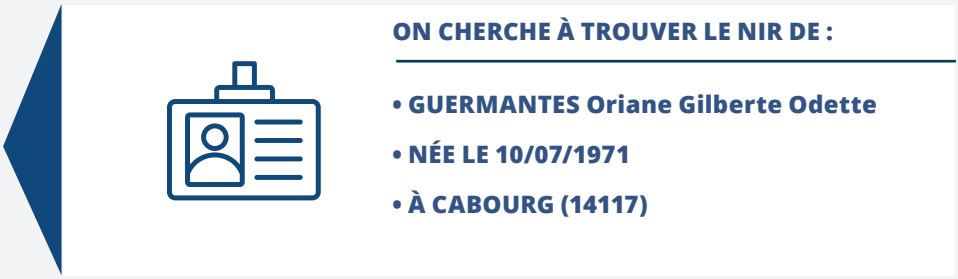
10 Un écho est un retour d'information à la suite d'une recherche sur les traits d'identité d'une personne. Si les traits d'identité recherchés sont incomplets ou très fréquents, plusieurs personnes peuvent correspondre et dans ce cas, plusieurs échos peuvent être restitués.

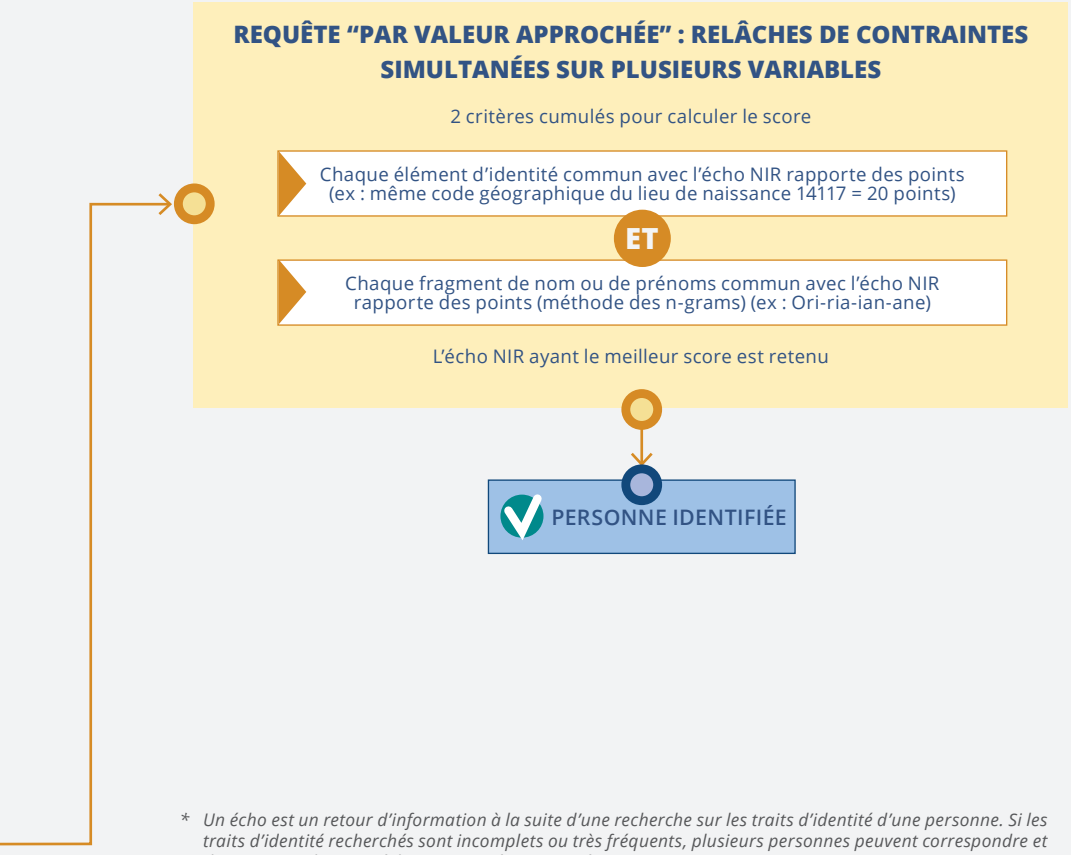
11 « Relâcher des contraintes » : réduire le nombre de composantes concordantes de l'identité nécessaires à la validation d'une identification.

Figure 2 - Exemple (fictif) pour retrouver un NIR (ou numéro de sécurité sociale)



à partir des traits d'identité





* *Un écho est un retour d'information à la suite d'une recherche sur les traits d'identité d'une personne. Si les traits d'identité recherchés sont incomplets ou très fréquents, plusieurs personnes peuvent correspondre et dans ce cas, plusieurs échos peuvent être restitués.*

** *Cette requête permet de retrouver les personnes qui se déclarent avec leur nom marital.*

- une identification par interversion des nom et premier prénom : les éléments d'identification sont tous identiques, à l'exception près que les champs nom et premier prénom sont intervertis pour la recherche d'identification.

Ces relâches successives sur les variables permettent en général de compléter significativement l'identification de l'étape de requête exacte. Pour les fichiers administratifs testés et avec une bonne qualité de remplissage des traits d'identité, l'identification sur l'ensemble des deux premières étapes permet en général de retrouver plus de 95 % des individus. En revanche, pour les fichiers d'enquêtes où la qualité de remplissage des traits d'identité est moins bonne, ces deux étapes donnent des résultats moins complets (80 %).

> ***Pour les fichiers administratifs testés et avec une bonne qualité de remplissage des traits d'identité, l'identification sur l'ensemble des deux premières étapes permet en général de retrouver plus de 95 % des individus.***

Après ces deux premières étapes, il reste toutefois à identifier les cas les plus complexes. Certains ont été délibérément exclus de la sélection des requêtes précédentes, leurs données d'identités étant sujettes à caution : il s'agit des personnes déclarées nées les 1er janvier et 31 décembre (ces dates étant parfois des dates « par défaut » indiquées lorsque l'information exacte n'est pas connue[12]) ou avec une date de naissance (jour/mois) à (00/00). Sont également exclues des sélections précédentes les identifications aboutissant à une personne de sexe différent de celui déclaré. Les autres cas complexes correspondent simplement à toutes les personnes non identifiées précédemment.

▶ Une méthode innovante pour les cas les plus complexes

Pour traiter ces cas, une troisième et dernière requête, « par valeur approchée », permet de relâcher des contraintes sur plusieurs variables de façon simultanée. La personne à identifier correspond à plusieurs échos dans le RNIPP et il s'agit de choisir celui qui présente le maximum de caractéristiques communes avec elle. Le choix s'appuie sur le classement des échos, réalisé en calculant un score. Le score est une somme de « points » attribués en fonction des éléments communs entre l'écho du RNIPP et la personne à identifier.

Ces scores sont d'abord attribués sur des éléments d'identité exactement identiques. Par exemple, le fait d'avoir le même code géographique du lieu de naissance donne 20 points ; le fait d'avoir les mêmes jours et mois de naissance donne 20 points ; le fait d'avoir le même nom donne 10 points ; le fait d'avoir le même premier prénom donne 10 points ; etc.

Mais cette méthode ne suffit pas pour bien discriminer les échos et éviter d'aboutir à des scores *ex æquo*. On ajoute alors un autre critère pour préciser les recherches, et calculer des scores plus fins.

12 Les personnes véritablement nées le 1er janvier ou le 31 décembre seront alors identifiées lors de la dernière étape.

Ce second critère de recherche consiste à filtrer les échos du RNIPP en fonction des fragments de noms ou de prénoms qu'ils ont en commun avec la personne à identifier. Ces fragments sont des tuiles de caractères, plus ou moins longues (3 à 5 caractères). Par exemple, le prénom Justine présente 5 tuiles de 3 caractères : *jus – ust – sti – tin – ine*. Cette technique appelée « méthode des n-grams » permet de discriminer de manière efficace les différents échos ayant des points communs avec les traits d'identité d'une personne à retrouver. Elle permet également de s'affranchir de quelques erreurs de saisie ou d'orthographe. Ensuite, chaque tuile commune apporte des points qui s'ajoutent au score. Ces opérations sont réalisées à l'aide du logiciel Elasticsearch. ***(encadré 2)***.

L'écho du RNIPP ayant obtenu le score le plus élevé est retenu et le CSNS sera alors calculé par chiffrement du NIR correspondant à cet écho comme aux étapes précédentes.

▶ Une obligation : mesurer la qualité de l'identification

Les dispositions réglementaires prévoient également que le service CSNS produise une mesure de la qualité de l'identification des personnes. Celle-ci est indispensable afin que les utilisateurs puissent apprécier la fiabilité de l'appariement à venir et adapter en conséquence d'éventuels traitements statistiques. La qualité de l'identification s'apprécie au regard de deux axes : la qualité générale du fichier des traits d'identité en entrée du processus et la qualité de l'identification par enregistrement individuel.

La qualité du fichier en entrée relève de la responsabilité du demandeur, mais le service CSNS lui fournit des informations pour l'aider à l'améliorer. Pour chacune des variables « nom, prénoms, année de naissance, jour et mois de naissance et code géographique du lieu de naissance », des indicateurs de taux d'anomalie et de taux de valeur manquante sont communiqués. Ces informations sont fournies directement par l'application mise à disposition des utilisateurs. Ceux-ci sont alors autonomes pour tester plusieurs versions de leur fichier. Après une première analyse, ils peuvent ainsi repérer les variables pour lesquelles des actions d'amélioration sont nécessaires. Il est possible de recommencer cette analyse autant de fois que nécessaire.

La qualité de l'identification de chaque enregistrement relève de la responsabilité du service CSNS. Cette qualité est appréhendée sous l'angle de la mesure des faux-positifs. Un faux-positif est une identification qui aboutit à trouver un NIR différent de celui de la personne recherchée. On se trompe donc de personne et l'appariement rapprochera les données de deux personnes différentes. À l'inverse, les faux-négatifs sont des personnes qui n'ont pas été identifiées alors qu'elles auraient dû l'être. La mesure de la qualité de l'identification individuelle de chaque enregistrement est présentée sur la base du taux de faux-positifs. Plus ce taux est faible, meilleure est la qualité supposée.

▶ Encadré 2 : Un moteur de recherche au cœur du moteur d'identification au RNIPP

Ces dernières années, la palette d'outils à disposition des statisticiens s'est considérablement enrichie : langages de programmation tels que R, Python et Julia, bibliothèques de modules très performants pour traiter le nettoyage, la transformation et l'analyse des données, *machine learning*, *deep learning,* visualisation, traitement du langage…

Ces évolutions se sont accompagnées d'avancées technologiques toutes aussi innovantes dans la façon de stocker, organiser et traiter les données : démocratisation de l'utilisation des systèmes de gestion de bases de données SQL, bases NoSQL, nouveaux formats de stockage, répartition des traitements. Autant d'avancées technologiques, majoritairement mises à disposition sous licences libres qui permettent de traiter efficacement des données massives.

Parmi ces outils, les moteurs de recherche ont largement bénéficié de toutes ces évolutions récentes, ont participé aux succès des grands acteurs de l'Internet et sont désormais incontournables dans la vie quotidienne. Nous les utilisons tous sciemment plusieurs fois par jour en ouvrant notre navigateur et tout autant inconsciemment, puisque les applications de médias sociaux, de commerce électronique, de cartographie, de transport, de diffusion de musique ou de vidéos utilisent toutes leurs capacités de recherche et d'analyse en temps réel à très grande échelle.

Pour les besoins d'identification au RNIPP où il s'agit d'apparier, avec des temps de traitements raisonnables, des fichiers de traits d'identité qui peuvent contenir jusqu'à plusieurs millions de lignes avec le répertoire des personnes qui lui en contient plus de 130 millions, l'équipe de projet CSNS s'est rapidement orientée vers une solution utilisant le moteur de recherche Elasticsearch pour plusieurs raisons :

1) Une architecture technique performante et évolutive

– l'architecture d'*Elasticsearch* est distribuée : plusieurs instances (ou nœuds) peuvent être lancées sur un ensemble de serveurs (ou *ferme de serveurs*) et collaborer. Les données, selon leur volume, peuvent être découpées en plusieurs partitions puis distribuées et répliquées sur les différents nœuds afin d'assurer performance et sécurité grâce aux dispositifs de répartition de charge et de haute disponibilité.

En cas de défaillance d'un ou plusieurs nœuds de la ferme, le système continue à fonctionner en mode dégradé, sans autre conséquence que des temps de réponse plus longs, et ceci tant que les nœuds restants peuvent accéder à au moins une version opérationnelle des données.

– l'architecture d'*Elasticsearch* est extensible : la puissance de calcul offerte par la ferme peut être adaptée à l'évolution des besoins ; il est possible d'étendre la ferme dynamiquement en ajoutant des serveurs.

2) Des fonctions de traitement et d'analyse de texte avancées

Elasticsearch offre nativement des fonctionnalités telles que la recherche en texte intégral, des analyseurs pour traiter et normaliser le texte, la recherche sur synonymes, l'analyse de données géospatiales, ou encore la segmentation en unités lexicales qui permet de découper les phrases en mots ou en n-grams de mots ou de caractères.

3) Une architecture flexible

Elasticsearch s'intègre sans difficulté dans un projet informatique, car il est facile de l'interroger à partir de tous les langages de programmation tels que Java, R ou Python pour ne citer que ceux qui figurent au schéma directeur informatique de l'Insee.

En complément de tous les mécanismes et fonctionnalités précédemment décrits, la force du moteur de recherche *ElasticSearch* est sa capacité à retrouver l'information parmi des millions de lignes en temps réel. Et pour ce faire, il "triche" un peu : tout est pré-calculé lors de la phase dite d'indexation pendant le chargement des données ; par "tout" il faut entendre les calculs de tous les n-grams possibles, les variantes des termes sans les caractères spéciaux, sans les majuscules, sans les mots vides de sens, etc. Cette phase prend beaucoup de temps (6 heures pour le RNIPP) et est réalisée une fois par mois dans le cas du CSNS pour intégrer les récentes mises à jour du RNIPP.

Puis, lorsqu'une requête de recherche est soumise, le serveur qui la reçoit la distribue sur les serveurs de la ferme et un score de pertinence pour chaque enregistrement correspondant à la requête est calculé. Ce score de pertinence est basé sur le module de similarité utilisé par Elasticsearch et paramétré par l'Insee pour évaluer les similarités entre les termes recherchés et les termes indexés. Les enregistrements avec les scores les plus élevés sont considérés comme plus pertinents et sont proposés en premier dans les résultats de recherche.

En conclusion, la solution retenue pour le CSNS permet de bénéficier de l'efficacité d'un moteur de recherche intégrant la combinaison de plusieurs facteurs : pré calcul des valeurs de tous les champs de recherche, répartition de ces informations sur plusieurs serveurs, distribution des requêtes sur tous ces serveurs pour augmenter la puissance de calcul et enfin un moteur de similarité pour produire les scores de pertinence des résultats.

> ***La stratégie consiste à minimiser les erreurs d'identification (avoir peu de faux-positifs) et indiquer de manière transparente le risque pris pour chaque enregistrement.***

La stratégie consiste à minimiser les erreurs d'identification (avoir peu de faux-positifs) et indiquer de manière transparente le risque pris pour chaque enregistrement. Le choix a ainsi été fait d'identifier tout le fichier en entrée, exceptés les quelques cas extrêmes où les données sont trop parcellaires, et d'indiquer pour chaque enregistrement une estimation de la probabilité d'avoir un faux-positif. Cette estimation prend la forme d'un indicateur de qualité en 7 modalités, allant de « parfaitement fiable » (1) à « non fiable » (7)[13].

L'utilisateur reste maître de son choix : retenir ou non l'identification et le CSNS afférent proposé. Selon ses objectifs, il peut éventuellement reconsidérer les cas qu'il juge trop incertains, soit en améliorant l'identification avec d'autres informations (par exemple l'adresse qui ne figure pas dans le processus CSNS), soit en refusant ces identifications et en traitant alors ces données avec des techniques de redressement analogues à celles du traitement de la non-réponse.

On aurait pu imaginer une autre approche consistant à livrer uniquement des CSNS pour les identifications jugées de bonne qualité. Mais la concertation préalable avec les futurs utilisateurs a mis en évidence le besoin de disposer du maximum d'informations, même de qualité moindre, pour conserver l'opportunité de les retraiter et de les améliorer, dès lors qu'une évaluation du niveau de fiabilité est fournie pour chaque enregistrement.

Une mesure de la qualité adaptée aux différentes méthodes d'identification

Les modalités pratiques du calcul des indicateurs de qualité doivent tenir compte de deux contraintes. D'une part, les méthodes d'identification sont différentes selon les étapes du processus, allant d'une identification exacte à une identification par valeur approchée. D'autre part, le calcul réel d'un taux de faux-positifs requiert que l'on dispose des NIR des personnes à identifier à comparer avec les NIR trouvés par le moteur. Or, ces fichiers ne sont pas très nombreux et ne constituent pas la majorité des fichiers pour lesquels le service CSNS sera utilisé. Par ailleurs, si un utilisateur dispose du NIR, le calcul des CSNS sera réalisé par simple hachage-chiffrement sans passer par l'étape d'identification sur traits d'identité.

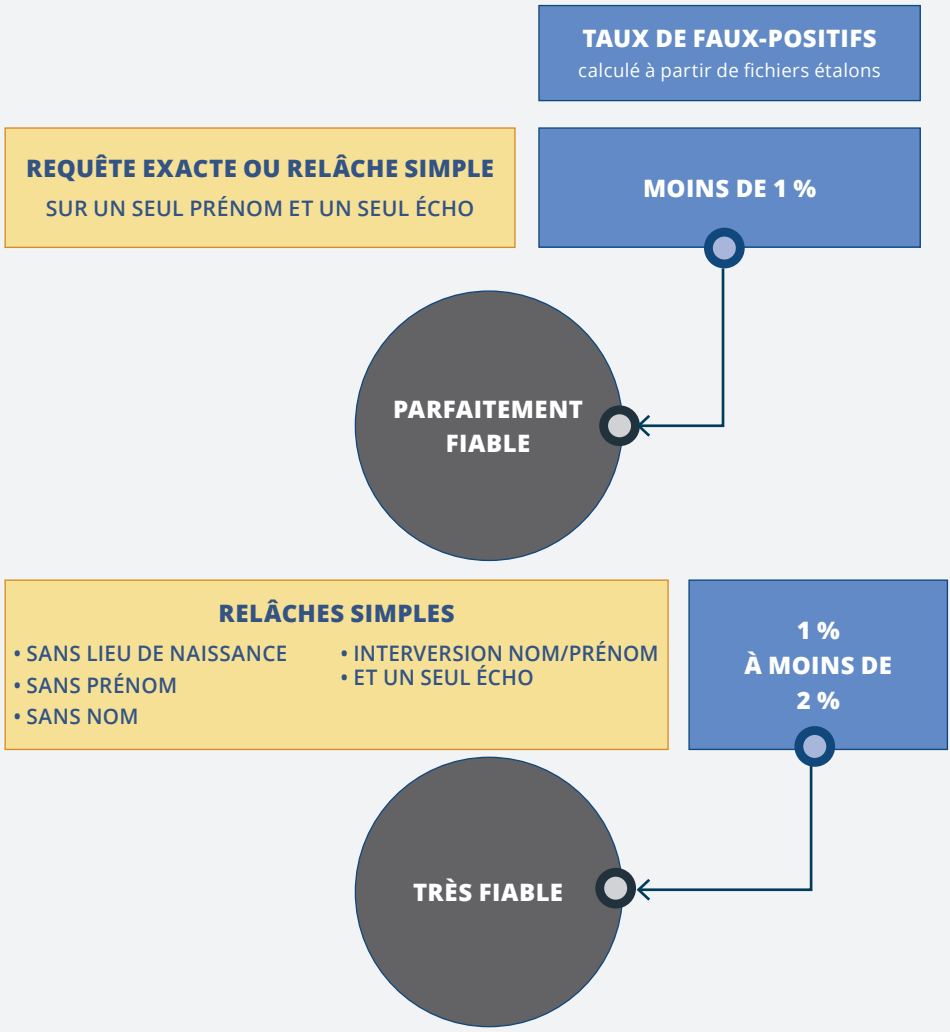
Lors de la phase de projet du CSNS, un étalonnage des taux de faux-positifs a ainsi été établi avec des fichiers comprenant le NIR pour définir des taux moyens par étape du processus. Par exemple, les calculs ont été réalisés sur les Déclarations sociales nominatives (DSN) et les NIR retrouvés après identification des traits d'identité ont été comparés aux NIR réels de ces fichiers, permettant ainsi de calculer des taux de faux-positifs sur un gros volume de données. De tels calculs empiriques ont pu aussi être menés sur l'Échantillon démographique permanent (EDP), sur les enquêtes annuelles de recensement dont le NIR avait été trouvé par un dispositif extérieur au CSNS *(Jabot et Treyens, 2018)*, et sur des fichiers

13 Parfaitement fiable (1) / Très fiable (2) / Fiable (3) / Assez fiable (4) / Peu fiable (5) / Très peu fiable (6) / Non fiable (7).

de la DREES[14] relatifs aux dispositifs d'insertion. Les valeurs moyennes de taux de faux-positifs calculées lors de cet étalonnage ont été prises comme référence pour déterminer le niveau de qualité de chaque étape.

Avec cette méthode, il ressort que les taux de faux-positifs moyens varient de 0 à 2 % lorsque la personne a été identifiée par les requêtes exactes ou simples (deux premières étapes) ***(figure 3)***.

Figure 3 - Critère principal de qualité, le taux de faux-positifs : plus il est faible, meilleure est la qualité.



14 Direction de la recherche, des études, de l'évaluation et des statistiques (DREES), service statistique ministériel dans les domaines de la santé et du social.

En revanche, pour la méthode de la requête par valeur approchée (dernière étape), la mesure de la qualité doit faire l'objet d'une approche spécifique, même si le principe de recherche des faux-positifs reste maintenu. Le score calculé pour chaque individu constitue une information intéressante, mais celle-ci ne peut pas être utilisée directement. En effet, avec la méthode des n-grams, plus un mot est long, plus il contient de tuiles de caractères et plus son score sera potentiellement élevé. Il est alors impossible de déterminer une liaison directe entre valeur du score et probabilité d'avoir un faux-positif. Toutefois, des scores très faibles correspondent souvent à des faux-positifs. Il s'agit alors de combiner cette information avec une autre.

Le rapport entre le score de l'écho retenu (le meilleur) et celui venant immédiatement en deuxième position est alors apparu comme une autre information à exploiter. En effet, les tests ont montré que plus l'écart entre les deux premiers échos est faible, plus la probabilité de se tromper est forte. Autrement dit, si les deux meilleurs échos trouvés pour une personne ont un score proche, il est probable que la différence de proximité avec les traits d'identité originaux soit trop faible pour être significative. Cette forte proximité fait prendre le risque de se tromper de personne et de retenir un faux-positif. Ainsi, les niveaux de faux-positifs calculés empiriquement sur les mêmes fichiers que pour les étapes précédentes ont été classés selon une double échelle de valeurs : celle du niveau de score, et celle du rapport « score du 2e écho sur score du 1er écho ».

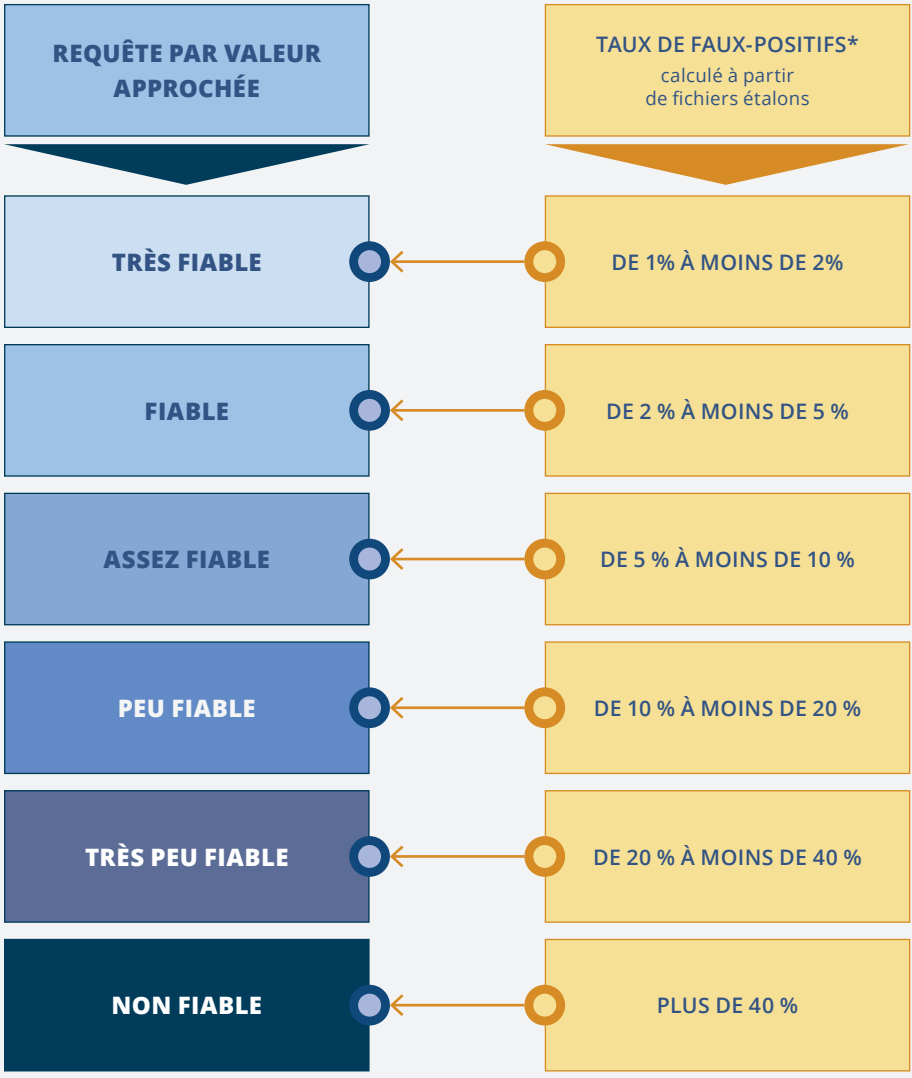
Au final, l'indicateur de qualité en 7 modalités, allant de « parfaitement fiable » à « non fiable » est calculé comme suit : les identifications trouvées à l'étape exacte et à la première requête simple de relâche sur le prénom sont classées en « parfaitement fiable » avec une probabilité de taux de faux-positifs de moins de 1 %. Celles trouvées aux autres requêtes simples sont classées en « très fiables » avec une probabilité de taux de faux-positifs de 1 à 2 %. Les identifications de l'étape par valeur approchée ont des indicateurs de « très fiable » à « non fiable » selon leur classement au regard de la valeur de leur score et de la valeur du ratio entre le deuxième et le premier score de la recherche *(Malherbe, 2022)* ***(figure 4)***.

Dernière étape de la mesure de la qualité : l'évaluation sur l'ensemble du fichier. Cette dernière phase informe l'utilisateur de la part des enregistrements pour chacun des sept niveaux de qualité : ces informations sont disponibles de façon automatique dans l'application dédiée.

En complément, d'autres outils sont fournis aux utilisateurs pour les aider à appréhender la qualité de l'identification de leur fichier. En particulier, une répartition par âge est calculée sur la population initiale du fichier en entrée du processus et sur la population identifiée en sortie. La comparaison des deux permet de voir si une sous-population particulière est sur-représentée parmi les échecs d'identification. Des informations analogues sont fournies pour comparer les populations nées en France et celles nées à l'étranger, la qualité des données d'état civil pouvant parfois être différente.

Ces indicateurs de qualité ont été testés avec des SSM volontaires et ils se sont révélés utiles et pertinents pour les appariements à venir.

▶ Figure 4 - Critère principal de qualité, le taux de faux-positifs : Davantage de modalités pour la requête par valeur approchée.



** corrélé à la valeur du score du 1er écho*
▶ plus elle est élevée, meilleure est l'identification

et

à l'écart de score entre le 1er écho et le 2e (mesuré par le ratio des scores)
▶ plus cet écart est faible, plus il est difficile de les départager et plus le risque de se tromper d'identification est élevé.

▶ Des premières utilisations prometteuses

En facilitant les appariements entre différentes sources, le CSNS contribue à l'extension des possibilités d'analyse des phénomènes économiques et sociaux. Le gisement est considérable et les premières utilisations donnent une idée du potentiel résultant de ce nouveau processus.

Lors de la phase de projet du CSNS, quatre services statistiques ministériels ont participé activement à de nombreux tests (Dares, Drees, SDES et SIES)[15], notamment pour faire des propositions, et évaluer la robustesse des choix méthodologiques et la pertinence des calculs d'indicateurs de qualité. Les exemples qui suivent montrent la diversité des sujets traités et l'intérêt en termes de connaissance de la société et d'évaluation des politiques publiques.

> ***Mieux mesurer l'insertion des diplômés de l'enseignement supérieur.***

Mieux mesurer l'insertion des diplômés de l'enseignement supérieur, c'est l'objectif du projet InserSup mené par le SIES avec la Dares. L'objectif est de produire des indicateurs d'insertion professionnelle, par établissement formateur et diplôme, sur l'ensemble des diplômes, et de mettre cette information à disposition des élèves et étudiants sur ParcourSup, MonMaster, Affelnet, Onisep[16], etc. pour les aider à choisir leur formation. Il s'agit aussi d'informer les acteurs territoriaux et les employeurs sur le lien formation-emploi, et plus généralement d'éclairer le débat public sur l'insertion professionnelle.

Le CSNS contribue à ce projet en facilitant l'appariement de sources administratives de provenances diverses sans avoir à recourir à des enquêtes spécifiques. Les sources mobilisées sont les différents répertoires étudiants du SIES, et la déclaration sociale nominative DSN (Dares) qui comprend de nombreuses informations sur les salariés. Est également envisagée l'utilisation de la base des non-salariés, des fichiers du recensement de la population et des fichiers fiscaux. Le service rendu par le CSNS permet ainsi de multiplier les appariements possibles entre les différentes sources d'observation, d'écourter les délais de mise à disposition et d'offrir une information exhaustive sur le champ couvert.

Plus généralement, l'obtention du CSNS pour les 40 millions de salariés du Système d'information statistique sur les mouvements de main d'œuvre (Sismmo) de la Dares permettra notamment d'étudier l'emploi étudiant en l'appariant avec les bases des inscrits de l'enseignement supérieur.

15 La Direction de l'animation de la recherche, des études et des statistiques (Dares) est le service statistique ministériel dans le domaine du travail ; la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) est le service statistique ministériel dans les domaines de la santé et du social ; le Service des données et études statistiques (Sdes) est le service statistique ministériel dans le domaine du logement, de la construction, des transports, de l'énergie, de l'environnement et du développement durable ; la Sous-direction des systèmes d'information et des études statistiques (Sies) est le service statistique ministériel de l'enseignement supérieur et de la recherche.

16 Parcoursup est une plateforme Web destinée à recueillir et gérer les vœux d'affectation des futurs étudiants de l'enseignement supérieur français ; MonMaster est une plateforme de consultation de l'intégralité des diplômes de Master et de dépôt des candidatures en 1re année de Master ; la procédure Affelnet permet d'affecter les élèves de troisième dans les lycées de leur académie ; l'Onisep est un opérateur de l'État qui produit et diffuse toute l'information sur les formations et les métiers.

Dans un tout autre domaine, le CSNS pourra contribuer à mieux mesurer l'impact de la transition écologique sur les différentes catégories de ménages. En rapprochant les données du Répertoire statistique des véhicules routiers (RSVERO) avec les données sur les revenus, la localisation exacte et les caractéristiques des ménages issues de la source Fidéli[17], c'est un nouveau champ d'analyse qui s'ouvre ; par exemple, le lien entre les caractéristiques des véhicules (puissance, type de carburant) et le revenu de leur propriétaire pourra être mieux appréhendé. Ces informations sont essentielles pour définir ou évaluer les politiques publiques relatives à la précarité énergétique ou à la transition écologique.

Dans le champ social, le CSNS va permettre d'enrichir les données des enquêtes sur l'autonomie et la dépendance. L'appariement de l'enquête CARE (enquête capacités, aides et ressources des seniors) avec les informations sur les prestations telles que l'APA (allocation personnalisée d'autonomie) et l'ASH (aide sociale à l'hébergement) permettra de suivre l'évolution de la dépendance des seniors deux ans après l'enquête. Parallèlement, le rapprochement des données de l'enquête VQS (vie quotidienne et santé) avec les données des régimes sociaux et celles sur l'insertion et l'emploi (notamment la déclaration sociale nominative) permettra de relier la prise en charge de la perte d'autonomie avec l'insertion professionnelle, et de traiter la question des incapacités en fin de carrière. À plus long terme, le CSNS ouvrira également des perspectives d'études et d'analyses dans de nombreux domaines tels que le devenir des enfants confiés à l'Aide sociale à l'enfance, les parcours des bénéficiaires du revenu de solidarité active (RSA) ou de l'APA… Sur ces sujets, l'exhaustivité des sources administratives appariées permettra de disposer de données territorialisées et régulièrement actualisées.

> ***Toutes ces nouvelles possibilités d'analyse offertes par le CSNS peuvent intéresser le monde de la recherche.***

Toutes ces nouvelles possibilités d'analyse offertes par le CSNS peuvent intéresser le monde de la recherche, au-delà du cercle restreint du service statistique public (SSP) *(Gadouche, 2019)*. Ainsi, même si le processus de calcul du CSNS est réservé au SSP, les résultats finaux des appariements peuvent être mis à disposition plus largement, mais sans faire figurer le CSNS lui-même.

Par ailleurs, la loi pour une République numérique de 2016 a aussi prévu un dispositif spécifique pour le monde de la recherche[18]. Une même opération de chiffrement du numéro de sécurité sociale pour aboutir à un code non signifiant utilisable comme clé d'appariement devient une nouvelle possibilité offerte aux chercheurs. La différence avec le CSNS est toutefois que ce code de recherche est attaché à un projet de recherche en particulier et ne peut pas être utilisé pour un autre projet. Ainsi un individu avec un code non signifiant dans un projet de recherche n'aura pas le même code dans un autre projet de recherche.

Quelques mois seulement après l'ouverture complète du service CSNS en octobre 2022, ce nouveau dispositif est déjà utilisé par cinq services statistiques ministériels et par plusieurs unités de l'Insee. L'intégration systématique et annuelle du CSNS est programmée pour plusieurs fichiers très utilisés par la statistique publique : Fideli

17 Fichiers démographiques sur les logements et les individus.
18 Articles 7 à 9 du décret n°2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche (voir Fondements juridiques).

(revenus fiscaux), les enquêtes annuelles de recensement (EAR), les fichiers issus de la Déclaration sociale nominative (DSN), l'échantillon démographique permanent (EDP)… L'expression « CSNSiser un fichier » se popularise auprès des statisticiens. Même si elle n'est pas très élégante, elle augure de la mise en place de bonnes habitudes qui rendront plus fluides et plus faciles les appariements à venir, et contribueront à augmenter encore et encore les sources de données nécessaires à une juste observation de notre société.

▶ Fondements juridiques

- Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données, Article 30). In : *Journal officiel des Communautés européennes*. [en ligne]. Mis à jour le 04 mai 2016. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679*.

- Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés modifiée et en particulier son article 30. In : *site de la CNIL*. [en ligne]. Mis à jour le 14 mars 2021. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.cnil.fr/fr/la-loi-informatique-et-libertes#article30*.

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique et en particulier son article 34. In : *site de Légifrance.* [en ligne]. Mis à jour le 08 octobre 2016. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

- Décret n° 82-103 du 22 janvier 1982 relatif au répertoire national d'identification des personnes physiques. In : *site de Légifrance.* [en ligne]. Mis à jour le 01 juillet 2021. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000520382*.

- Décret n° 2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche. In : *site de Légifrance*. [en ligne]. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033735139*.

- Arrêté du 28 septembre 2020 pris en application des articles 3 et 4 du décret n° 2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche. In : *site de Légifrance*. [en ligne]. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000042382996*.

▶ Bibliographie

- CHRISTEN Peter, 2012. *Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://link.springer.com/book/10.1007/978-3-642-31164-2*.

- ESPINASSE Lionel et ROUX Valérie, 2022. Le Répertoire national d'identification des personnes physiques (RNIPP) au cœur de la vie administrative française. In : *Courrier des statistiques*. [en ligne]. Novembre 2022. Insee. N°8, pp.72-92. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665188?sommaire=6665196*.

- FELLEGI, Ivan, SUNTER, Alan, JARO, 2014*. Approach to Record Linkage (Method)*. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://cros-legacy.ec.europa.eu/content/fellegi-sunter-and-jaro-approach-record-linkage-method_en*.

- GADOUCHE Kamel, 2019. Le centre d'accès sécurisé aux données (CASD), un service pour la *data science* et la recherche scientifique. In : *Courrier des statistiques*. [en ligne]. Décembre 2019. Insee. N°N3,pp.76-92. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254227?sommaire=4254170*.

- JABOT P. et TREYENS P.E. (2018). Proposition d'un nouvel appariement de l'enquête CARE par identification du plus proche écho. Actes des journées de méthodologie statistique 2018. [Consulté le 14/03/2023]. Disponible à l'adresse : *http://www.jms-insee.fr/2018/S20_1_PPT_TREYENS_JMS2018.pdf*.

- MALHERBE Lucas, 2022. Méthodologie des appariements individuels. JMS 2022. [Consulté le 14/03/2023]. Disponible à l'adresse : *http://jms-insee.fr/jms2022s07_1/*.

Quels formats pour quelles données ?

Alexis Dondon, Pierre Lamarche***

La donnée, pour être intelligible par ses utilisateurs et accomplir sa fonction de transport de l'information, doit être structurée. Une telle structure se formalise au travers d'un modèle de données, qui conditionne le choix du format. Les formats de données sont variés et répondent à des problématiques spécifiques, différentes selon le contexte de l'utilisation de la donnée par le métier. Dans l'idéal, les standards sous-jacents aux modèles de données et les formats associés sont partagés par le plus grand nombre d'utilisateurs. S'agissant de la donnée statistique, ces problématiques sont localisées dans les objets pour lesquels les données sont susceptibles de véhiculer de l'information, mais également dans la documentation de la donnée – la métadonnée – ou encore dans la volonté d'associer à la donnée des solutions logicielles spécifiques particulièrement adaptées à son traitement.

Sur ce dernier point, l'émergence ces dernières décennies de solutions open-source a permis de concilier les notions de sécurisation de l'import de la donnée, d'efficience de son traitement, de reproductibilité, etc. En particulier, des formats comme Parquet s'intègrent à des solutions logicielles accessibles à tous et adoptées par une communauté de plus en plus large, convaincue de ses avantages. Néanmoins, il n'existe pas de réponse définitive et unique pour le choix d'un format : des choix sont faits après une analyse précise des besoins relatifs à chaque étape du cycle de vie de la donnée. En cela, le choix d'un format est l'expression concrète d'un standard dicté par des impératifs propres à chacune de ces étapes.

In order to be intelligible to its users and to fulfil its function of conveying information, the data must be structured. This structure is then formalised through a data model, which determines the choice of format. Data formats are various and address specific problems, according to the context of use. Ideally, the standards behind the data models and the associated formats are shared as many users as possible. For statistical data, these problems are located in the objects for which the data are likely to convey information, but also in data documentation - i.e. metadata - or in the wish to link specific software solutions with the data particularly well suited to its processing.

On this last point, the emergence over the last few decades of open-source solutions has made it possible adress simultaneously different issues securing data import, efficiency of data processing, replicability, etc. In particular, formats such as Parquet are integrated into software solutions accessible to all and adopted by an increasingly large community, convinced of their advantages. Nevertheless, there is no clear-cut answer for the choice of a format: choices are made after a precise analysis of the needs relative to each step of the data's life cycle. In this way, the choice of a format is the concrete expression of a requirement driven standards specific to each of these phases.

* Data Engineer, DSI, Insee,
alexis.dondon@insee.fr

** Chef de la division Logement et Patrimoine, DSDS, Insee,
pierre.lamarche@insee.fr

La donnée, en tant qu'émanation concrète de l'information, est un élément-clé dans une société qui fonde la prise de décision sur l'information ; elle doit être transmissible et être codifiée selon des conventions établies et partagées (*Warnier, 1974*). La notion de transmissibilité s'entend au sens où le récepteur de la donnée doit être en mesure de la lire et de la comprendre. La donnée statistique se définit, quant à elle, comme une donnée à usage statistique (*Sundgren, 2010*), c'est-à-dire avec pour but l'énumération et la quantification de phénomènes sous forme codifiée. Dans un monde où la donnée est de plus en plus présente et massive, l'usage de celle-ci par des systèmes informatisés requiert l'adoption de standards et de conventions de stockage partagés, adaptés à son usage et aisément reproductibles. Ainsi, toute personne voulant utiliser, modifier ou enregistrer des données se trouve confrontée à la question des conventions, constantes dans le temps et l'espace, à adopter pour les lire ou les stocker sous un format adapté à l'usage qu'elle souhaite en faire. De la même manière qu'un langage véhicule de l'information qui se transmet si celui-ci est connu du locuteur et de l'auditeur, les conventions adoptées pour le stockage des données doivent être partagées entre utilisateurs.

▶ Modéliser l'information pour créer la donnée

Avec l'avènement de l'informatisation et de la numérisation de la donnée, la question de sa transmissibilité s'est très rapidement posée. Les réponses se trouvent dans la définition de standards et de conventions de stockage et de transmission.

> ***L'établissement de conventions partagées est essentiel pour une communication efficace et fluide entre différents acteurs d'un système d'information.***

La modélisation de la donnée répond à un besoin de définir un langage commun. Elle est étroitement liée à la nature de l'information : dans un cas fictif de données individuelles, il faut définir les attributs de chaque individu (nom, prénom, date de naissance, lien de parenté) ainsi que leur représentation. Une fois cette modélisation réalisée, un ensemble de formats pourra être mobilisé. Par ailleurs, l'établissement de conventions partagées est essentiel pour une communication efficace et fluide entre différents acteurs d'un système d'information. L'intégration d'un réseau qui partage une même convention peut, à ce titre, devenir un élément stratégique, voire vital. Plus cette convention est partagée et utilisée par un grand nombre d'utilisateurs, plus sa valeur *ipso facto* est importante. À titre d'exemple, les données de la Déclaration Sociale Nominative[1] font l'objet d'une modélisation et de l'adoption de standards par l'ensemble des parties prenantes de manière à permettre des échanges d'informations massifs[2].

Le standard constitue une réponse à un besoin de structuration de l'information dans un contexte donné. Il va définir une manière de présenter la donnée qui permet de répondre à des problématiques précises, essentielles pour les professionnels qui utilisent ces informations. Le format est la concrétisation de ce standard,

1 La Déclaration sociale nominative (DSN) est une déclaration obligatoire pour les entreprises du secteur privé qui sert à payer les cotisations sociales et à transmettre les données sur les salariés aux organismes sociaux.
2 Voir à ce sujet l'article sur les normes d'échanges dans le même numéro.

et à un standard donné peuvent correspondre plusieurs formats. Cependant, le standard ne fait pas tout et le champ des formats possibles peut s'avérer très large.

La représentation tabulaire, un modèle canonique pour la statistique

La sphère statistique a cherché très vite à définir des formats de données, avec des spécificités propres à ses activités. La donnée est évidemment centrale pour la statistique. Elle peut prendre une grande variété de formes tant les sources d'information utilisées sont diverses. Derrière cette question, se situe le sujet de la modélisation de la donnée, élément clé s'agissant des systèmes de bases de données utilisés par le statisticien, mais également pour le stockage de la donnée inerte[3].

> ***La donnée est évidemment centrale pour la statistique.***

La statistique a recours, dans de nombreux domaines, aux micro-données[4], que ce soient des données d'enquête ou encore des données de registres. Leur numérisation doit satisfaire les contraintes de volume et de calcul. Lorsque les capacités de mémoire des machines sont relativement limitées, les formats ouverts et peu consommateurs d'espace vont être privilégiés. La statistique s'accommode très naturellement des formats tabulaires[5], puisque cette façon de structurer la donnée correspond à l'approche théorique de la statistique. Dans la représentation tabulaire de la donnée, la ligne va correspondre de manière canonique aux observations, et les colonnes aux variables. Cette notion peut être considérée comme une application particulière d'une vision matricielle de l'information ; elle est très intimement associée dans le monde de la statistique à des outils, des modes de pensée et des méthodes de calcul largement partagés et très structurants.

Ainsi, les formats tabulaires sont privilégiés : la modélisation de la donnée va alors consister à définir des tables (ou objets) et des liens entre elles (des relations), comme pour les systèmes de gestion de base de données (ou SGBD, ***encadré 1***). Cependant, les SGBD sont particulièrement adaptés à la donnée « vivante », qui a vocation à évoluer fréquemment, et relèvent également de choix d'infrastructure informatique qui dépasse le champ de la statistique. Les notions spécifiques aux SGBD ne sont pas développées ici, contrairement aux données inertes et aux formats de stockage associés à ce type de données.

3 Qui se définit comme une donnée qui n'a pas vocation à être modifiée, mais simplement lue.
4 Au sens «données individuelles», par opposition aux données agrégées comme les statistiques.
5 Qui s'entend comme une présentation de la donnée en ligne et en colonne.

La fin définit le moyen en matière de format

L'utilisation d'un format de stockage doit assurer une qualité essentielle de la donnée statistique, son accessibilité. Le statisticien doit s'assurer que les données qu'il produit et qu'il utilise puissent être aisément réutilisées, par lui-même ou par d'autres utilisateurs, et ce en minimisant les pré-requis informatiques à la lecture et au traitement de ces données. Ce besoin est renforcé dans le contexte d'*open-data* où la donnée a de plus en plus vocation à être partagée de manière complète avec le plus grand nombre d'utilisateurs, et où la question de l'accessibilité de cette donnée est incontournable.

> ***Le format dépend avant toute chose de l'usage de la donnée et également de son volume.***

Idéalement, la donnée doit être accessible et adaptée au contexte informatique de son usage (dit autrement, on va chercher à minimiser la consommation de ressources informatiques au sens large *(Nordbotten, 1966)*). Pour résoudre cette quadrature du cercle, il n'y a pas de réponse unique ; le format dépend avant toute chose de l'usage de la donnée et également de son volume. Il convient de distinguer les micro-données, qui portent sur des observations individuelles (et qui sont souvent des données volumineuses) des données agrégées, qui résultent d'un premier exercice d'agrégation ou d'estimations.

Encadré 1 : Les bases de données, un format adapté à la donnée « vivante »

Les systèmes de gestion de base de données (SGBD) sont des solutions logicielles qui permettent de stocker, d'utiliser et de traiter des données généralement massives ; elles offrent des solutions optimisées en matière de consommation de ressources informatiques pour traiter des volumes significatifs de données. En contrepartie, elles sont constituées d'un système de gestion de fichier dont la complexité est masquée à l'utilisateur, qui a la possibilité de lancer des requêtes sur ces données à l'aide d'outils standards*. Elles vont optimiser l'utilisation de la donnée par des systèmes d'indexation des observations et de hachage des identifiants, permettant ainsi d'appréhender un large spectre de modèles de données. Le plus courant et le plus utilisé dans la statistique publique est le modèle tabulaire, incluant généralement une notion de liens entre tables (système de gestion dit relationnel)**.

Historiquement, les SGBD les plus largement utilisés sont d'abord des solutions propriétaires telles que Oracle ou MySQL. Dans la dernière décennie, les SGBD *open-source* se sont progressivement imposés dans l'univers de la donnée ; l'un des plus connus est PostgreSQL.

Les SGBD sont souvent des solutions de stockage adaptées à la donnée « vivante », c'est-à-dire à une donnée évolutive, pour laquelle ces derniers apportent des garanties en matière de préservation de l'intégrité et de la cohérence des données et de réversibilité des traitements opérés. En contrepartie, cette solution de stockage de la donnée peut être très énergivore, comparativement à d'autres. Elles doivent donc être considérées surtout pour leur usage courant, celui du stockage et du traitement de la donnée vivante. Le stockage de la donnée inerte sur un SGBD a peu d'intérêt, et peut être considéré comme un gaspillage énergétique, dans une démarche d'informatique durable***, notion appelée à devenir incontournable (Ademe, 2021).

* *Bien souvent à l'aide d'un langage standard, le SQL (Structured Query Language), qui peut varier dans ses subtilités selon le type de SGBD utilisé.*

** *Elles peuvent également gérer d'autres types de données, par exemple des données où les champs varient d'une observation à l'autre (on parle de documents flexibles).*

*** *Il existe néanmoins des SGBD adaptés au stockage de données inertes, avec une faible consommation énergétique associée.*

Naturellement, le traitement d'une information agrégée ou d'une information individuelle n'a pas les mêmes implications en matière de stockage de la donnée. Lorsqu'on veut choisir un format, il est donc essentiel de répondre aux questions suivantes :

- quel volume ?
- quels usages, pour quels utilisateurs ?
- quelle localisation de la donnée ?

Les formats textuels s'accommodent naturellement de la représentation tabulaire

Les formats textuels sont les formats les plus simples à utiliser, puisqu'ils peuvent être aussi bien interprétés par une machine que lus par un humain[6]. Les fichiers à largeur fixe ou les formats avec délimiteur (en particulier le format comma-separated values ou csv ***(figure 1b)***) permettent aisément de compacter la donnée sous un format tabulaire. Si le format à délimiteur peut se suffire à lui-même, les fichiers à largeur fixe supposent de disposer d'une information sur la position de chaque colonne ; la description des données prend ici toute son importance : sans elle, il est très difficile de retrouver la structure de la donnée, et donc de la traiter. Historiquement, le format positionnel fixe ***(figure 1a)*** a été très utilisé, car il consomme peu d'espace, en contrepartie d'une description de sa structure. Il continue à être utilisé, par exemple pour les transmissions de données entre certaines administrations et l'Insee. L'exemple de données contenant une liste d'individus et de liens de parenté entre ceux-ci nécessite cependant une description précise de la position de chaque variable.

> ***Le format avec délimiteur est utilisé fréquemment, du fait de sa simplicité d'utilisation.***

Le format avec délimiteur est utilisé fréquemment, du fait de sa simplicité d'utilisation. Il présente cependant des inconvénients, avec en particulier la contrainte de l'adoption d'un délimiteur absent des données qu'il structure[7]. Le format avec délimiteur suppose une prise de risque relative à l'intégrité de la donnée, tout comme le format à largeur fixe et de manière générale les formats textuels impliquant une opération d'import et donc d'interprétation de la structure de la donnée. Ce risque est encore plus élevé si les données ne sont pas correctement documentées ; il est question de dictionnaire des codes, de dessin de fichiers, et plus généralement d'un ensemble de documents qui permet de traiter et d'interpréter la donnée en cohérence avec les intentions des producteurs de cette donnée. Par ailleurs, les formats textuels ne concernent pas uniquement les modèles de données tabulaires ; au contraire, les modèles de données plus « élastiques » peuvent être servis par certains formats textuels de manière plus sécurisée.

6 On parlera de format « *human readable* » dans la terminologie anglo-saxonne.
7 Par exemple, si on utilise le caractère «/» comme délimiteur, il faut s'assurer qu'aucune donnée ne puisse contenir ce caractère.

▶ Figure 1 - Quelques exemples de formats

FIGURE 1A - POSITIONNEL FIXE

```
000011MARTIN      ARTHUR                0806195575014
000012MARTIN      JEANNE                1503195607010
000021MARTIN      CLAUDE    00001100001218101984750l4
...
```

FIGURE 1B - FORMAT CSV

```
ID_IND;NOM;PRENOMS;ID_PERE:ID_MERE:DATE_NAIS;COM_NAIS
000011;MARTIN;ARTHUR;;08061955;75014
000012;MARTIN;JEANNE;;15031956;07010
000021;MARTIN;CLAUDE;000011;000012;18101984;75014
...
```

FIGURE 1C - FORMAT XML

```
<OBSERVATION>
    <ID>000021>/ID
    <NOM>MARTIN</NOM>
    <PRENOMS>CLAUDE</PRENOM>
    <DATE NAIS>18101984</DATE NAIS>
    <LIEU NAIS>75014</LIEU NAIS>
    <PERE>
            <D>000011</ID>
            <NOM>MARTIN</NOM>
            <PRENOMS>ARTHUR</PRENOMS>
            <DATE NAIS>08061955</DATE NAIS>
            <LIEU NAIS>75014</LIEU NAIS>
    </PERE>
    <MERE>
            <ID>000012</ID>
            <NOM>MARTIN</NOM>
            <PRENOMS>JEANNE</PRENOMS>
            <DATE NAIS>15031956</DATE NAIS>
            <LIEU NAIS>07010</LIEU NAIS>
</MERE>
</OBSERVATION>
...
```

FIGURE 1D - FORMAT JSON

```
[
    {
    “ID”: “000021”,
    “NOM”: “MARTIN”,
    “PRENOMS”: “CLAUDE”
    “DATE NAIS”: “18101984”
    “LIEU NAIS”: “75014”,
    “PERE”:
            {
            “ID”:”000011”
            “NOM”: “MARTIN”
            “PRENOMS”:”ARTHUR”
            “DATE NAIS”:”08061955”
            “LIEU NAIS”: “75014”
            },
    “MERE”:
            {
            “ID”:”000012”,
            “NOM”: “MARTIN”
            “PRENOMS”: “JEANNE”
            “DATE NAIS”: “15031956”
            “LIEU NAIS”: “07010”
    },
...
]
```

▶ Des formats « élastiques » pour les documents flexibles —

Les données ne se présentent pas toujours sous un format tabulaire. Ainsi, certains modèles de données peuvent introduire de la flexibilité dans les champs, ou dans la notion même d'observations et de variables. Le format JSON[8] par exemple ***(figure 1d)*** est un format minimaliste qui introduit ce type de flexibilité. Il va en particulier s'accommoder de données contenant des champs largement facultatifs, comme les données du fichier permanent des occurrences de traitement des émissions (POTE)[9] *(Lamarche et Lollivier, 2021)* relatives à la déclaration des revenus, pour laquelle un grand nombre de cases reste vide dans la plupart des observations. Format très flexible mais plus consommateur d'espace, XML ***(figure 1c)*** est particulièrement adapté au stockage de données à faible volume, c'est-à-dire essentiellement les macro-données ou données agrégées. Il permet également de stocker les métadonnées qui fournissent les informations contextuelles relatives à la donnée. Caractéristique de la technologie Web, XML fonctionne sous forme de balises : il s'agit donc là encore d'un format textuel, pour lequel certains caractères spéciaux permettent d'identifier la structure. Son avantage principal est que le format s'adapte à de nombreuses représentations de la donnée, pas nécessairement tabulaires. Il permet également de mettre dans un même réceptacle la donnée et des éléments descriptifs de celle-ci ; de fait, il donne la possibilité de coupler de manière très naturelle la donnée et la métadonnée. Les apports de ces deux formats vis-à-vis des formats tabulaires traditionnels sont importants, en particulier dans leur capacité à embarquer des relations entre les tables (par exemple la filiation entre une table de parents et une table d'enfants), de manière très naturelle[10], tout en s'accommodant des contraintes des formats tabulaires.

> ***Ces formats très flexibles et auto-suffisants sont parfaitement adaptés pour mettre à disposition des données.***

De manière générale, ces formats très flexibles et auto-suffisants sont parfaitement adaptés pour mettre à disposition des données. Ainsi, un grand nombre d'API[11] offrent la possibilité aux utilisateurs de consommer soit de la donnée, soit des services de traitement de la donnée, en adoptant ces formats. La flexibilité est un avantage précieux qui permet d'accompagner l'information d'un ensemble d'éléments contextuels, assimilable à des métadonnées, en informant l'utilisateur sur la nature de la donnée récupérée ou encore sur la qualité du traitement dont elle a bénéficié. La présentation sous forme de table est moins décisive, car le requêtage d'une API se fait de ligne à ligne : le résultat de la requête est important ainsi que sa description. En revanche, le volume de données sollicitées est nécessairement limité. Même si la finalité est d'obtenir des données sous une représentation tabulaire, le traitement élémentaire se réalise dans un cadre qui ne nécessite pas cette représentation.

8 JSON : JavaScript Object Notation.
9 Le POTE est le « fichier permanent des occurrences de traitement des émissions », élaboré par les services de la DGFiP à partir des émissions des avis d'imposition sur les revenus.
10 Sous forme de données emboîtées ou encapsulées.
11 Une API (application programming interface ou « interface de programmation d'application ») est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.

Ces formats sont très adaptés pour des jeux de données de faible volume, mais dont la structure peut être évolutive, ou tout du moins difficile à établir *a priori*. Pour traiter de gros volumes, il est nécessaire d'adopter des modèles de données plus contraignants et le modèle tabulaire est bien souvent adapté ; il faut alors chercher l'efficacité dans la capacité à interpréter rapidement les caractéristiques des données.

La place cruciale des métadonnées dans la diffusion rend ces formats incontournables et les statisticiens s'y trouvent naturellement confrontés dans les processus de production.

La métadonnée, un élément de plus en plus incontournable

Les métadonnées peuvent contribuer fortement à donner plus de sens aux données qu'elles décrivent, car elles créent des liens entre les différentes sources de données. Nombre d'initiatives au niveau international visent à définir des standards et des formats associés partagés par le plus grand nombre d'acteurs ***(encadré 2)***.

De manière plus pragmatique, le code de bonnes pratiques de la statistique européenne mentionne les métadonnées comme un élément de la qualité des statistiques mises à disposition de l'usager : fournir aux utilisateurs de l'information sur la donnée elle-même est un critère très important pour juger de son accessibilité et de sa clarté. Si la communauté statistique européenne donne à ce terme de métadonnée un sens précis, celui-ci recouvre en réalité un large spectre d'informations relatives à la donnée. Le premier élément est la description de la donnée en termes informatiques :

Encadré 2 : le standard DDI et le format SDMX

Les formats incorporant des métadonnées sont très souvent des formats « élastiques », qui permettent de relier la métadonnée avec la donnée de manière naturelle. Ils sont spécifiques, car ils comportent une couche d'abstraction supplémentaire et embarquent la sémantique propre à la métadonnée. Acteur important de définition d'une sémantique partagée pour la métadonnée, la Data Documentation Initiative (DDI) est un consortium international d'instituts de recherche et de producteurs de statistiques qui vise à définir des standards pour la documentation des données statistiques, avec un focus particulier sur les données d'enquêtes, des méthodes de collecte et des référentiels (nomenclatures, codifications, etc.) utilisés pour la collecte. Ce consortium existe depuis 1995 : il travaille sur la définition d'un format de stockage de ces métadonnées visant à créer un éco-système d'outils, à la fois en amont de la collecte et en aval pour la diffusion des données produites. Cette vision intégrée et l'adoption des outils produits par cet éco-système doivent bénéficier à la qualité des données produites et à l'utilisateur final tout autant qu'au producteur. La pertinence de ces outils, leur simplicité et leur exhaustivité sont déterminants pour qu'un format s'impose.

SDMX* est une initiative impliquant des instituts de statistiques et des organismes internationaux, visant en particulier à renforcer le caractère transmissible de la donnée. Le format SDMX s'appuie sur la métadonnée, car la description standardisée de la donnée facilite sa mobilisation. Là aussi, un éco-système d'outils accompagne l'adoption du standard (Salou et Sosnovsky, 2010) : les institutions internationales comme Eurostat**, la BCE***, la BIS**** ou encore l'OCDE***** poussent pour l'adoption de ces standards, car leur adoption rend possible la résolution de nombreux sujets relatifs à l'agrégation de données et aux questions d'entrepôts de données, sujets très importants pour ces instituts. Par ailleurs, l'adoption la plus massive possible de standards communs permet de consolider les connaissances et d'accroître les potentialités de relier les données entre elles.

** SDMX : Statistical Data and Metadata eXchange.*

*** Eurostat est l'institut statistique communautaire, associé à la Commission européenne.*

**** La BCE est la Banque centrale européenne.*

***** La BIS est la Bank for International Settlements ou Banque des règlements internationaux.*

****** L'OCDE est l'Organisation de coopération et de développement économiques.*

> ***Il faut distinguer la métadonnée qualifiée de technique de la métadonnée documentaire.***

quel type des variables présentes dans les tables, quelles unités de mesure, quelle largeur de colonnes pour les variables en format caractère, etc. Mais le champ couvert par les métadonnées va beaucoup plus loin, et doit permettre de documenter le processus de génération des données de manière générale. Il faut distinguer la métadonnée qualifiée de technique, qui donne de l'information sur le type des variables par exemple, de la métadonnée documentaire, qui informe l'utilisateur sur le contexte de production de la donnée, les choix méthodologiques, les nomenclatures adoptées, etc. Cas particulier de cette dernière, la notion de ligne de données permet de suivre le cycle de vie de la donnée, en traçant les sources à l'origine des données et les transformations subies par celles-ci pour aboutir au résultat. L'ensemble de ces métadonnées constitue un enjeu essentiel dans un contexte où les statistiques produites sont de plus en plus utilisées dans le cadre de comparaisons internationales, et pour lesquelles il devient crucial pour l'utilisateur de comprendre les raisons de divergences potentielles liées à des modes de collecte radicalement différents. Ainsi, le sujet de la métadonnée est porté par les institutions internationales traitant de la production statistique, au premier plan desquelles Eurostat.

S'agissant des données statistiques, ce mouvement d'internationalisation des conventions s'inscrit dans un contexte où les données, de par leur nature numérique, ont de plus en plus vocation à s'appuyer sur des technologies *web*. La flexibilité de ces technologies pour le stockage des données, et en particulier leur caractère adaptatif à la structure que celles-ci peuvent prendre, favorise leur adoption dans les standards ; en particulier, le consortium *Data Documentation Initiative* (DDI) repose sur le format XML pour la définition de ses conventions de stockage.

Tabulaires ou flexibles, les formats textuels ont pour grand avantage leur lisibilité et une forme de stabilité dans le temps. Sauf modification très profonde du système informatique, les données stockées sous ces formats seront lisibles et mobilisables longtemps après, et ce avec d'autant plus de facilité si les données contiennent des métadonnées. Cette disponibilité dans le futur a un coût dans le présent : le stockage de données toujours plus nombreuses s'accommode assez mal des contraintes qu'un format textuel implique. Il est donc essentiel d'introduire des formats qui, s'ils apportent moins de garantie de stabilité, permettent en revanche de répondre aux problématiques présentes.

La nouvelle donne de la donnée : volume, vélocité, variété, virtualisation, transversalité

Les statisticiens sont concernés par plusieurs sujets, qui ont directement un impact sur les choix de formats. L'adoption par le plus grand nombre de conventions et de standards est une source d'efficience qui bénéficie directement à tous les acteurs qui font le choix de les adopter. De ce point de vue, il est nécessaire d'intégrer les problématiques qui traversent l'ensemble du monde de la donnée et les choix qui en découlent de la part de la plupart des acteurs.

Tout d'abord, l'expansion continue de la donnée impose de développer des solutions logicielles permettant de traiter des volumes toujours plus importants de la manière la plus efficace possible. Le *Big data*, ou donnée massive, a mis sur le devant de la scène les préoccupations de stockage et de performance auxquelles des logiciels propriétaires tels que SAS® ou Oracle avaient apporté de premières réponses. En particulier, le fait que cette donnée massive s'accompagne de plus en plus d'une structuration flexible de l'information impose des formats alternatifs au format tabulaire historique.

Cette expansion s'accompagne d'un recours toujours plus important à la virtualisation des traitements et du stockage, avec l'avènement du *cloud computing*. Ce principe conditionne les solutions techniques et les formats associés. Une solution de stockage de type S3[12] est ainsi souvent associée à des outils de requêtage optimisés pour certaines catégories de format. Son adoption va donc de pair avec ces outils.

Par ailleurs, l'efficience budgétaire est un élément important qui préside souvent au choix de solutions *open source* et de formats ouverts que celles-ci impliquent, considérées comme étant moins coûteuses pour les processus de production. Toutefois, si l'*open source* est un vecteur très intéressant de mutualisation des investissements, il n'est pas une ressource disponible gratuitement, pour deux raisons essentiellement. D'une part, il faut disposer des compétences permettant de maîtriser ces outils ; d'autre part, le développement de ces outils a un coût, et la sécurisation des processus fondés sur ces outils nécessite une identification de ces développements et une maîtrise de l'assurance que ceux-ci se poursuivent dans le temps.

> ***Le renforcement nécessaire de la confiance dans la statistique officielle passe nécessairement par une attention plus forte aux notions de transparence et de reproductibilité.***

Plus important encore, dans une société concernée par les sujets de *fact-checking* et de *fake news*, le renforcement nécessaire de la confiance dans la statistique officielle passe nécessairement par une attention plus forte aux notions de transparence et de reproductibilité. Ces notions sont le moteur d'un recours toujours plus fort à la donnée ouverte ou *open data*, et à la possibilité donnée à des utilisateurs de reproduire de la manière la plus complète possible les processus aboutissant à la production de la donnée sous son état final. L'*open data* implique par essence le recours à des formats ouverts, et la reproductibilité à des solutions logicielles *open source*, par définition transparentes vis-à-vis des algorithmes qu'elles mettent en œuvre.

Enfin, l'approche historique, avec un unique logiciel statistique comme « couteau suisse » permettant de réaliser l'intégralité du traitement de la donnée, est progressivement remplacée par une vision plus fragmentée ; chaque étape du processus (que l'on peut modéliser dans le cadre du GSBPM[13]) est mise en œuvre grâce à une solution logicielle dédiée et spécifique, pour laquelle l'outil le plus approprié sera choisi. De ce point de vue,

12 Une solution de stockage développée par Amazon qui tend à s'imposer comme un standard dans le monde de la donnée.

13 Le modèle générique de description des processus de production statistique (GSBPM pour Generic Statistical Business Process Model) décrit les différentes étapes à suivre pour produire des statistiques publiques.

> ***Chaque étape du processus est mise en œuvre grâce à une solution logicielle dédiée et spécifique, pour laquelle l'outil le plus approprié sera choisi.***

la complémentarité entre les langages de programmation R et Python est souvent mise en avant ; elle suppose l'adoption d'un format de données transversal, qui permet de basculer aisément d'un outil à l'autre sans les phases d'import-export de données fastidieuses et périlleuses. Ce type de format est un élément de sécurisation du processus dans son ensemble. Les formats de stockage ont ainsi évolué, nécessairement vers des solutions fondamentalement nouvelles.

▶ Le format binaire, une efficience sous contrainte

L'import de données à partir de fichiers texte est une tâche risquée, compliquée ou coûteuse en calculs : la sécurisation de ce processus d'import passe par sa minimisation (on le réalise le moins de fois possible). À ces formats ouverts, lisibles par un œil humain, on oppose les formats de stockage de données binaires, qui contiennent la donnée sous un format autre que textuel. Cet impératif de stocker la donnée sous un format illisible pour un simple éditeur de texte permet de répondre à un autre besoin, celui de présenter la donnée de la manière la plus rapidement interprétable par la machine. La conversion de l'information entraîne la transformation des données en une séquence non directement déchiffrable de morceaux élémentaires de donnée informatique, les bits. Ces bits ne prennent que deux valeurs possibles, 0 ou 1 : on parlera alors de "format binaire". On distingue ainsi deux manières de stocker de l'information en informatique, selon qu'on puisse le déchiffrer par un éditeur de texte (format ouvert) ou pas (format binaire).

Lorsque le format binaire est un format propriétaire, il ne peut généralement être déchiffré que par les outils associés à ce format. Les formats propriétaires présentent des avantages considérables en matière d'utilisation de la donnée, propres au format binaire : accessibilité quasi immédiate à celle-ci, et bien souvent à un grand nombre de métadonnées stockées de manière native, permettant de décrire la donnée.

Les formats propriétaires (au premier rang desquels SAS®) ont rapidement émergé dans les activités où la donnée était un élément central. Les solutions logicielles associées à ces formats se sont imposées par leur capacité à traiter des volumes importants de données de manière optimisée, en fournissant à l'utilisateur une sorte de « couteau suisse » permettant à la fois de traiter la donnée et de lui appliquer des procédures statistiques proches de l'état de l'art. On a ainsi une forte cohérence entre le format de stockage et la solution logicielle adoptée par la statistique publique pour la production et l'analyse de la donnée.

Dans un contexte d'augmentation du volume de données, le choix d'un outil comme SAS®, qui les traite comme un système de fichiers, a permis d'apporter une première réponse au défi de calcul que de telles données ont pu poser aux machines. Cette vision a longtemps présenté des avantages dépassant largement les inconvénients que peuvent constituer l'adoption d'un format par essence fermé, imposant à l'ensemble des utilisateurs l'usage de la solution logicielle associée, ou encore la modélisation purement tabulaire,

et donc par essence frustre, de la donnée. Elle a néanmoins été battue en brèche ces dernières décennies par une demande d'ouverture de la donnée, par une exigence de reproductibilité, et par l'apparition dans le paysage de véritables écosystèmes de logiciels libres, associant ouverture et performance.

Combiner les avantages : les formats binaires et *open-source*

Si les formats propriétaires ont permis d'associer un format à une solution logicielle, il existe des formats binaires développés en *open-source*, qui ont pour but de répondre au seul besoin d'optimiser leur utilisation par la machine, tout en relâchant la contrainte de la solution logicielle.

De tels formats sont souvent inter-opérables par différentes solutions logicielles, souvent elles aussi *open-source*. C'est la première contrainte à laquelle doit s'adapter le format de stockage convenant au mode actuel d'utilisation des données : sa capacité à s'abstraire d'une solution logicielle définie *a priori* (le « couteau suisse » mentionné précédemment), pour que l'utilisateur puisse choisir l'outil le plus adapté au traitement qu'il veut réaliser, dans un contexte où celui-ci s'insère dans une chaîne incorporant des traitements de natures et de finalités variées. En résumé, l'utilisateur doit pouvoir recourir à une palette d'outils divers sans pour autant modifier à chaque changement d'outil la nature de l'objet qu'il traite.

> ***Dans le monde de la statistique, la donnée est très souvent une donnée que l'utilisateur va lire plus fréquemment qu'il ne la modifie.***

Par ailleurs, une statistique est une donnée qui sera bien plus fréquemment lue que modifiée. De ce point de vue, le format de stockage peut s'apparenter au principe WORM (*Write Once, Read Many*) : le format doit chercher à optimiser le processus de lecture qu'en fait l'utilisateur, moins que son processus de transformation et d'écriture.

Ce principe est d'autant plus pertinent dans un contexte où l'infrastructure correspond à l'état de l'art, avec une virtualisation du traitement de la donnée et de son stockage, accompagnée d'un découplage de ces deux notions : le stockage et le calcul se déroulent sur des infrastructures différentes, tout en maintenant une grande capacité à mobiliser très rapidement les données pour les traiter ensuite.

Enfin, s'agissant des données massives, dans un contexte contraint où il devient impératif de consommer l'énergie avec parcimonie et discernement, la donnée stockée sous forme de bases de données énergivores doit l'être pour de bonnes raisons : il s'agit de pouvoir transformer cette donnée de manière sécurisée et à grande vitesse. En règle générale, la donnée inerte suffit à couvrir de manière très satisfaisante la plupart des usages analytiques ou de production des statisticiens.

Néanmoins, d'autres problématiques ont vu le jour dans le paysage de la statistique : la nécessité, en matière d'efficience cette fois, d'adopter des standards et des conventions les plus partagées possibles et des solutions de calcul toujours plus complexes, en raison de l'expansion considérable des données disponibles.

▶ Parquet, un format compact et décomposable

Les formats répondant à ces impératifs ont émergé ces dernières années avec l'avènement de solutions intégrées de traitement de la donnée massive, telles Hadoop[14]. En particulier, le format Parquet ***(figure 2)*** permet de solliciter de manière très naturelle la donnée de façon parallélisée, c'est-à-dire en la scindant, en la distribuant très rapidement à plusieurs unités de traitement et en la traitant de cette manière en parallèle. Le nom de ce format résume à lui seul ses propriétés : schématiquement, la donnée va être stockée sous forme de « lames » denses, plus ou moins fortement compressées et mobilisables chacune de manière indépendante. Ce format permet de stocker la donnée grâce à différents algorithmes de compression, qui réduisent de manière très significative la taille des données sans dégrader la vitesse à laquelle cette donnée peut être mobilisée *(Uber, 2022)*.

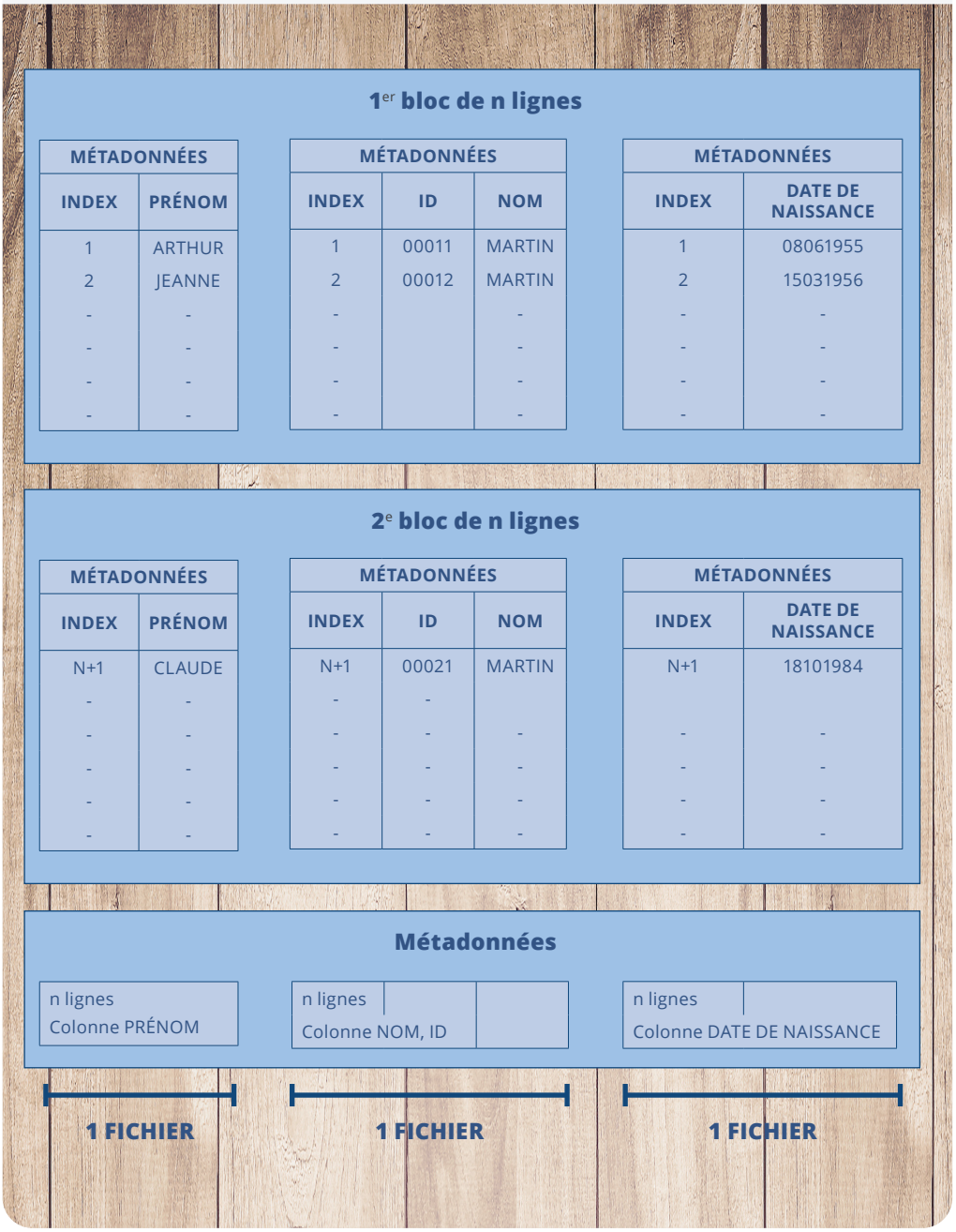
> ***Le format Parquet permet de solliciter de manière très naturelle la donnée de façon parallélisée.***

Parquet se base sur un principe algorithmique de stockage décrit et mis en œuvre par les équipes de Google dans leur processus de requêtage *(Melnik et al, 2010)*. Les données sont représentées sous format tabulaire, et les « lames » vont regrouper plusieurs colonnes et un groupe d'observations. Selon l'usage que l'on veut faire des données, il s'agira donc de définir des regroupements de colonnes ainsi qu'une stratification des observations la plus efficiente possible. Chaque « lame » contient également un ensemble de métadonnées décrivant les colonnes présentes, ainsi que la partie d'observations contenues dans cette « lame ». Les métadonnées vont être lues de manière indépendante des données *via* les « connecteurs » des solutions logicielles, de manière à permettre une navigation optimisée dans la table. Une des spécificités de cette conception de format est de permettre également de décomposer les données en plusieurs fichiers, de manière à rendre plus concrète et visible à l'utilisateur la notion de « lame » ; les connecteurs vont alors récupérer l'information associée aux métadonnées pour l'ensemble des fichiers Parquet contenus dans un même dossier, en considérant ces fichiers comme décrivant une seule et unique table.

Le format Parquet est donc particulièrement adapté pour gérer de la donnée volumineuse et distante[15], sans la dupliquer *(Mauvière, 2022)*. En s'imposant comme un standard *open source*, il donne la possibilité à l'utilisateur de travailler avec différents outils logiciels dont la complémentarité est précieuse. Il ne saurait néanmoins être vu comme une réponse unique à l'ensemble des problématiques du monde moderne de la donnée.

14 Système de calcul distribué développé en *open-source* et qui s'est imposé comme un outil largement utilisé pour traiter des données volumineuses.
15 C'est-à-dire stockée dans un endroit différent de celui où se réalise le calcul.

▶ Figure 2 - Le format Parquet



▶ Un format pour chaque étape du cycle de vie de la donnée

Quelles sont les questions à se poser pour choisir un format ? Tout d'abord, quel usage ? Le stockage des données et le formatage qu'il implique doit répondre à trois objectifs distincts : le traitement de la donnée, son analyse et son stockage pour un usage ultérieur. À chacune de ces étapes correspondent des usages très différents, et donc des besoins très spécifiques.

Ensuite, le volume est naturellement déterminant, mais les caractéristiques attendues du formatage vont également dépendre des utilisateurs et plus précisément de leur degré de maîtrise des outils de traitement de données. La micro-donnée, plus volumineuse et plus exigeante, sera plutôt réservée à un public averti, quand le grand public et de manière générale l'utilisateur *lambda* se référeront plus souvent à des tableaux agrégés.

> ***Le stockage des données et le formatage qu'il implique doit répondre à trois objectifs distincts : le traitement de la donnée, son analyse et son stockage pour un usage ultérieur.***

Enfin la localisation des données est un critère à prendre en compte pour la détermination du format le plus adapté : ce choix est en effet très dépendant de l'infrastructure informatique et des solutions logicielles disponibles pour traiter la donnée. En particulier, l'avènement de la virtualisation[16] du traitement est déterminant pour la définition des formats et de leurs aspects purement techniques. La virtualisation de la donnée amène à privilégier des formats offrant de bonnes performances en matière de lecture [17], comme le format Parquet, et incite à bien distinguer la donnée qui a vocation à être transformée de celle qui va être uniquement lue. À ces questions sont donc associés des critères relatifs au caractère temporaire ou durable de la donnée, ou dit autrement, à son degré de maturité.

Les différents formats sont déclinés selon l'usage des données et leur degré de maturité, en se plaçant dans différents contextes informatiques, comme décrit précédemment, et où l'utilisateur a la possibilité de virtualiser la chaîne de traitement qu'il applique aux données ***(voir Tableau)***. En particulier, dans ce contexte, la mise en place d'un SGBD peut se faire de manière transitoire, celui-ci étant construit de manière à être à « usage unique » pour procéder au traitement des données, puis être immédiatement supprimé une fois le traitement accompli, et la donnée en sortie stockée de manière plus permanente.

16 La notion de virtualisation décrit un principe selon lequel les données sont stockées sur des espaces physiques, bien souvent distants, pour lesquels l'utilisateur n'a pas besoin de connaître le détail de l'infrastructure de stockage, et peut les utiliser de façon transparente.
17 On retrouve ici la notion de WORM de la donnée.

Tableau : Propositions de formats associés à chaque usage de la statistique publique

	TRAITEMENT	DIFFUSION ET ANALYSE	ARCHIVAGE
DONNÉES AGRÉGÉES		XML, JSON	CSV, XML, JSON
MICRO-DONNÉES PEU VOLUMINEUSES	Format ouvert type CSV ou binaire type Parquet	Binaire type Parquet	CSV, XML, JSON
MICRO-DONNÉES VOLUMINEUSES, STOCKÉES LOCALEMENT	SGBD, binaire ad hoc lié à la solution logicielle ou Parquet	Binaire ad hoc lié à la solution logicielle ou Parquet	CSV
MICRO-DONNÉES VOLUMINEUSES ET VIRTUALISÉES	SGBD type Postgres ou binaire type Parquet	Binaire type Parquet	CSV

Dans son cycle de vie, la donnée doit initialement présenter une grande flexibilité car elle est sujette à de nombreuses modifications. Le format utilisé pour son stockage doit tenir compte de ces caractéristiques. Mais, plus la donnée gagne en maturité, plus il est souhaitable de recourir à des formats adaptés à des données figées, permettant une plus grande facilité de lecture de la donnée et optimisant son utilisation.

En guise de conclusion

Le choix du format est fonction de la finalité recherchée, du volume manipulé, de la localisation de la donnée. S'agissant de la production statistique, il est lié à la question de l'infrastructure et des solutions logicielles adoptées, qui appellent à tirer parti du potentiel offert par différents types d'infrastructure *(Comte et al., 2022)*. Les choix de stockage effectués par les producteurs ne sauraient ignorer ces nouveautés, visant à stocker la donnée de la manière la plus disponible possible, tout en donnant à l'utilisateur le maximum de liberté quant aux outils à mettre en œuvre pour l'utilisation souhaitée.

La notion de format adapté est appelée à évoluer dans le temps avec l'avènement de nouvelles solutions techniques et de nouveaux standards. Se pose également la question de la gestion du patrimoine des instituts statistiques. Le modèle de données a peu évolué au cours du temps, et la plupart de l'information produite et analysée par les statisticiens se présente sous format tabulaire. La conversion purement technique d'un format vers un autre est par essence relativement triviale. En revanche, les standards de description de la donnée et la notion de métadonnée posent avec une acuité nouvelle la question de la gestion de ce patrimoine. La façon d'y répondre — et en particulier avec une documentation la plus proche des standards actuels — aura un impact sur la capacité des chargés d'études à utiliser ce patrimoine de données ; cela constitue un fort enjeu pour la statistique publique lorsqu'il s'agit de décrire des phénomènes sur longue période.

Par ailleurs, les travaux récents relatifs à la question des *linked open-data (Zimmermann et al. 2011)* a renforcé l'importance des métadonnées, en particulier sur la question critique du référencement de la donnée et de son accessibilité. Ces travaux, expérimentaux à ce stade, vont nécessairement conditionner les standards, en particulier du point de vue de la diffusion, et de ce fait, avoir un impact sur les choix de format associés.

Choisir efficacement le format appelle aussi à réviser les principes de gouvernance de la donnée au sein du service statistique public, en décrivant de manière plus formalisée son cycle de vie et en cherchant à caractériser de façon normalisée et systématique la position de la donnée traitée au sein de celui-ci. De ce point de vue, les enjeux de parcimonie de stockage, de jeux de données de référence ou encore le principe de minimisation issu du Règlement Général à la Protection des Données conditionnent pour beaucoup la manière dont la donnée doit être stockée, traitée et gérée. Les formats les plus adaptés de ce point de vue sont ceux qui s'accommodent le mieux des solutions techniques propres à ces principes de gouvernance.

▶ Bibliographie

- ADEME, 2021. Prospective - Transitions 2050 - Rapport : Agir maintenant pour le climat. In : *Rapport de l'Ademe*. [en ligne]. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://librairie.ademe.fr/recherche-et-innovation/5072-prospective-transitions-2050-rapport.html*.

- COMTE, Frédéric, DEGORRE, Arnaud et LESUR, Romain, 2022. Le SSPCloud : une fabrique créative pour accompagner les expérimentations des statisticiens publics. In : *Courrier des Statistiques*. [en ligne]. 20 janvier 2022. Insee, N° N7, pp. 68-85. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035940?sommaire=6035950*.

- LAMARCHE, Pierre et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des Statistiques*. [en ligne]. 8 juillet 2021. Insee, N° N6, pp. 28-46. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- MAUVIÈRE, Éric, 2022. Parquet devrait remplacer le format CSV. Post de blog. [en ligne]. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.icem7.fr/cartographie/parquet-devrait-remplacer-le-format-csv/*.

- MELNIK, Sergey *et al.*, 2010. *Dremel : interactive analysis of web-scale datasets*. In : *Proceedings of the VLDB Endowment*, 3 (1-2), pp. 330-339. [en ligne]. Septembre 2010. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://dl.acm.org/doi/abs/10.14778/1920841.1920886*.

- NORDBOTTEN, Svein, 1996. *A statistical file system*. In : *site researchgate.net*. [en ligne]. Septembre 2010. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.researchgate.net/publication/283934373_A_Statistical_File_System.*

- SALOU, Gérard et SOSNOVSKY, Xavier, 2010. *SDMX as the logical foundation of the data and metadata model at the ECB*. [en ligne]. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://ideas.repec.org/h/bis/bisifc/33-09.html*.

- SUNDGREN, Bo, 2010. *Statistical databases – an introduction*. [en ligne]. 20 octobre 2010. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.diva-portal.org/smash/get/diva2:386476/FULLTEXT01.pdf*.

- UBER, Blog, 2022. *Cost Efficiency At Scale in Big Data File Format*. Post de blog. [en ligne]. 25 janvier 2022. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.uber.com/en-TZ/blog/cost-efficiency-big-data/*.

- WARNIER, Jean-Dominique, 1974. L'organisation des données d'un système, In : *Précis de logique informatique*. Édité par Honeywell Bull, CII Honeywell Bull. ISBN 978-2-7081-0229-3.

- ZIMMERMANN, Antoine., LOPES, Nuno, POLLERES, Axel, et STRACCIA, Umberto. (2011) *A general framework for representing, reasoning and querying with annotated semantic web data.* In : *Journal of Web Semantics, 11, 72-95*. [en ligne]. 7 mars 2011. [Consulté le 22 juin 2023]. Disponible à l'adresse : *https://arxiv.org/pdf/1103.1255*.

L'intégration des données administratives dans un processus statistique

Industrialiser une phase essentielle

F. Cotton, O. Haag***

La statistique publique a de plus en plus recours à des sources externes, en particulier à des données administratives, pour produire des statistiques. Ceci nécessite d'industrialiser davantage les processus de production et notamment le processus d'intégration de ces données, afin de sécuriser, d'assurer une meilleure traçabilité et de rendre cette intégration la plus reproductible possible.

L'objectif du statisticien public est de mettre en place un cadre général d'intégration de données permettant une démarche automatisée sur des données structurées, livrées par un producteur externe. Plus précisément, il s'agit d'implémenter un pipeline jalonné de points de contrôle qui permettent de s'assurer que la succession des tâches (renommer, restructurer les données, recoder, pseudonymiser, etc.) se déroule correctement et d'arrêter le processus dès qu'un éventuel problème est rencontré. En outre, l'utilisation de standards et de métadonnées actives le long de ce pipeline permettent au concepteur d'être le plus autonome possible et ainsi de pouvoir l'adapter plus facilement aux évolutions des sources externes.

Official statistics increasingly rely on external sources, particularly administrative data, to produce statistics. This requires greater industrialisation of the production and integration processes for these data, in order to make them more secure, more traceable and as reproducible as possible.

The aim is to implement a general data integration framework based on an automated approach to structured data delivered by an external producer. This involves setting up a pipeline with checkpoints to ensure that the succession of tasks (renaming, data restructuring, recoding, pseudonymisation, etc.) is carried out correctly and to stop the process as soon as any problems are encountered. In addition, the use of standards and active metadata upstream of this pipeline allows the designer to be as autonomous as possible, making it easier to adapt to changes in external sources.

* Expert, DSI, Insee,
franck.cotton@insee.fr
** Directeur du programme « Répertoire statistiques d'individus et de logements », DSDS, Insee,
olivier.haag@insee.fr

La statistique publique produit des chiffres et des études à partir d'enquêtes qu'elle gère de façon totalement autonome, de la conception à la diffusion après les étapes de collecte et de traitement de la non-réponse. Toutefois, avec la profusion de données produites et mises à disposition par d'autres organismes, le recours massif à des sources externes, et notamment à des données administratives, pour produire des statistiques augmente fortement *(Lamarche et Lollivier, 2021, Hand 2018)*. Cette pratique se développe également au niveau international (*UNECE Integration, statswiki, Cros 2014*). De ce fait, il y a aujourd'hui une forte volonté d'industrialiser davantage l'intégration de sources administratives afin de sécuriser, d'assurer une meilleure traçabilité et de rendre le plus reproductible possible cette intégration de données.

> ***Une forte volonté d'industrialiser davantage l'intégration de sources administratives afin de sécuriser, d'assurer une meilleure traçabilité et de rendre le plus reproductible possible cette intégration de données.***

Bien souvent, la donnée administrative n'est pas utilisable en tant que telle *(Courmont, 2021)*. L'intégration consiste alors à détacher les données et les métadonnées de leur univers de gestion d'origine pour les attacher au monde statistique (unité statistique, concepts, nomenclature, etc.). Cette opération nécessite une appropriation de la source par ses futurs utilisateurs.

Lorsque la qualité de la donnée administrative est jugée suffisante pour produire des statistiques fiables (« qualification de la source »), la première étape du processus consiste à intégrer les informations externes dans le système d'information statistique. À l'instar du processus de collecte par enquête, l'intégration est la première étape des traitements statistiques (contrôle, validation, redressement) qui permettent de passer du statut de données brutes à celui de données diffusables.

Les objectifs, le périmètre exact et les étapes de ce processus d'intégration des données sont présentés dans cet article. La phase de qualification n'est abordée ici que brièvement bien qu'elle soit essentielle. En effet les sources administratives peuvent présenter des défauts de couverture (exemple de données absentes en Alsace-Lorraine car définies après le régime concordataire) ou utiliser des concepts éloignés de ceux de la statistique publique (par exemple : données collectées au niveau de compteurs électriques pour facturer la consommation d'un logement) : il est nécessaire de bien identifier ces problèmes en amont afin de les traiter au plus tôt.

La qualification de la source : un pré-requis indispensable avant l'intégration

Avant d'intégrer une source, il convient de s'assurer qu'elle dispose des qualités suffisantes pour pouvoir être utilisée à des fins statistiques. Ainsi, il est nécessaire d'échanger avec le producteur de la donnée afin de vérifier que la source est :

- exploitable (les données contenues peuvent être restructurées pour mesurer des concepts statistiques) ;

- complète (aucune sous-couverture évidente qui empêcherait son exploitation) ;
- disponible dans un délai raisonnable ;
- documentée (présence de métadonnées).

À noter que l'office national statistique anglais envoie des statisticiens directement dans les administrations pour réaliser cette phase de qualification *(Fermor-Dunman et Parsons 2022).*

▶ De quoi parle-t-on ?

Les données administratives sont recueillies et structurées par des organismes pour leurs besoins propres. Par exemple, la déclaration sur le revenu est collectée au niveau du foyer fiscal (unité gérée par la DGFiP[1]). Cette notion désigne l'ensemble des personnes inscrites sur une même déclaration de revenus et ne correspond pas directement aux concepts d'individu et de ménage utilisés pour la statistique. En effet un foyer fiscal peut regrouper un ou plusieurs individus et il peut y avoir plusieurs foyers fiscaux dans un seul ménage : par exemple, un couple non marié compte pour deux foyers fiscaux si chacun remplit sa propre déclaration de revenus. Il faut donc restructurer cette information administrative de base pour l'exploiter à des fins statistiques.

> ***La phase d'intégration transforme la donnée administrative individuelle en donnée individuelle brute statistiquement exploitable.***

La phase d'intégration transforme la donnée administrative individuelle en donnée individuelle brute statistiquement exploitable. Elle sélectionne uniquement les informations indispensables des sources externes, respectant ainsi les principes de minimisation et de nécessité auxquels les statisticiens sont tenus. On change de gouvernance en passant du monde administratif au monde statistique. Les données pourront être modifiées par le statisticien sans en informer le producteur initial. Par ailleurs, ces données deviennent alors des données statistiques et sont donc soumises au secret statistique.

Cette phase permet de constituer un ensemble cohérent à partir de plusieurs fichiers différents, voire de plusieurs sources différentes. Un exemple très intéressant à cet égard est celui de Fidéli[2]. Ce fichier intègre plusieurs sources fiscales : une source décrivant les membres des foyers imposables, une source contenant les informations sur l'impôt sur le revenu de ces foyers, une source contenant les informations relatives à la taxe d'habitation et une autre à la taxe foncière, chacune pouvant être constituée de plusieurs fichiers, plus d'une centaine par source pour obtenir une information sur la France entière.

La phase d'intégration produit la version de base de la donnée qui alimentera la suite du processus. À ce stade, il s'agit seulement de réorganiser l'information contenue dans les fichiers administratifs mais en aucun cas de les contrôler et encore moins de les

1 Direction générale des finances publiques : direction de l'administration publique centrale française qui dépend du ministère chargé de l'économie.
2 Fichier démographique sur les logements et les individus.

corriger. Néanmoins, un effort de traçabilité est nécessaire pour garantir la reproductibilité de cette base de départ.

L'intégration, point de départ du processus statistique, voire point de reprise en cas de problème, est également un point de référence. En comparant la donnée brute à la donnée modifiée par des gestionnaires dans la phase de « contrôle et corrections manuelles et automatiques », il est possible de juger de l'efficience des contrôles mis en place. De même, en comparant les données brutes aux données issues de redressements automatiques, on peut mesurer la variance de ces traitements et le gain en qualité, pour différents indicateurs, notamment en matière de réduction de biais.

L'intégration, une étape fondamentale à réaliser pour produire des statistiques

Si l'on compare les cycles de vie de la donnée d'une production statistique réalisée à partir d'une enquête et celle réalisée à partir d'une source externe, on constate qu'ils diffèrent sur la phase amont mais qu'ils sont identiques à partir des traitements de transformation des données brutes statistiques en données diffusables.

En référence au GSBPM[3] *(UNECE, GSPBM)*, l'intégration de données précède les phases aval de traitement des données et d'analyse conduisant aux données diffusables (contrôle, correction manuelles et automatiques, validation des données individuelles, agrégation et validation des données agrégées, diffusion des données et archivage des données). Ces phases ne sont pas évoquées dans cet article.

> ***L'intégration est la première étape du processus.***

Ainsi, si le fichier administratif contient des valeurs manifestement erronées (30 février par exemple), ces valeurs ne seront pas modifiées durant la phase d'intégration des données mais corrigées par une valeur possible lors de la phase suivante de « contrôle, correction manuelles et automatiques et validation des données individuelles ».

L'intégration est la première étape du processus. C'est aussi le moment d'initialiser les métadonnées statistiques qui seront utiles tout au long du cycle de vie de la donnée (concepts, description des variables, listes de codes et nomenclatures). Ce point s'inscrit dans une logique de pilotage des processus par les métadonnées dites « actives[4] ». ***(cf. infra)***

3 GSBPM pour *Generic Statistical Business Process Model*, modèle générique de description des processus de production statistique.

4 Les métadonnées peuvent aller au-delà de leur rôle de description, *via* une interface dédiée ou des applications clientes. Ces outils permettent de tirer parti du caractère exhaustif et normalisé des métadonnées pour produire automatiquement des composants du processus statistique. Les métadonnées acquièrent ainsi un statut nouveau, passant du stade d'« informations facilitant la compréhension des statistiques », au stade de « données participant au processus de production » ; d'où l'idée de métadonnées actives (*Bonnans, 2019*).

L'intégration : phase qui distingue les enquêtes des statistiques produites à partir de données administratives

L'enquête statistique est prise en charge par les statisticiens dès le début et permet de s'assurer, lors de la conception du questionnaire, que les données collectées répondent à des concepts et unités statistiques. Elles n'ont donc pas besoin d'être transformées après leur collecte et sont immédiatement des données statistiques brutes exploitables.

A contrario, les données externes doivent être transformées (constituer des unités statistiques, renommer, etc.) avant de pouvoir être considérées comme des données brutes statistiques. Certaines métadonnées sont initialisées durant cette phase d'intégration, telles que la définition des variables, leurs règles de calcul, leurs liens avec les concepts statistiques, etc.

À l'Insee, il existe des processus statistiques qui associent des données d'enquêtes et des données administratives afin de produire des statistiques robustes tout en limitant la charge statistique, à savoir le temps que les répondants doivent consacrer à l'enquête. On peut citer l'enquête « Revenus Fiscaux et Sociaux » qui fusionne des données de l'enquête « Emploi » avec des données fiscales et sociales, ou encore le système d'élaboration des statistiques annuelles d'entreprises (ESANE) fondé sur les résultats des enquêtes sectorielles annuelles et sur les liasses fiscales et les données de la déclaration sociale nominative (DSN).

Description des différentes étapes du processus d'intégration

Les données administratives sont souvent liées à des politiques publiques, dont les contours juridiques peuvent évoluer rapidement au cours du temps (ajout ou suppression d'un impôt ou d'une prestation sociale, etc.). Il est donc indispensable que les outils et méthodes développés dans le cadre de cette phase d'accueil des données soient les plus transparents et les plus adaptables possibles.

> ***La transformation de la donnée administrative en donnée statistique brute est composée d'une succession de tâches élémentaires.***

L'intégration est une succession d'opérations élémentaires écrites de façon déclarative, c'est-à-dire en spécifiant le quoi et non le comment, qu'il est aisé de « rejouer » permettant d'obtenir une chaîne de traitement pérenne et réutilisable. Pour mettre en œuvre cette reproductibilité et assurer son adaptabilité, il est indispensable de s'appuyer sur des standards et d'outiller le statisticien pour qu'il soit le plus autonome possible. Le langage VTL est une solution permettant de répondre à ces deux besoins majeurs. La transformation de la donnée administrative en donnée statistique brute est composée d'une succession de tâches

élémentaires (de façon analogue à une composition de fonctions en mathématiques) qui peuvent se décomposer selon les six catégories suivantes :

- renommer les variables sélectionnées ;
- restructurer les données par unité statistique ;
- recoder ;
- calculer les variables dérivées ;
- filtrer les enregistrements utiles ;
- pseudonymiser.

Cet ordre est indicatif et peut différer d'une source à l'autre. En particulier, les phases de pseudonymisation et de filtrage peuvent intervenir à différents moments de la séquence.

Cette catégorisation s'inspire du service d'accueil des sources mis en place par le programme Résil (Répertoire statistique des individus et des logements) *(Durr et alii, 2022)* qui va jouer ce rôle d'intégration de données administratives dans le système d'information démographique et social en utilisant l'outil Accueil Réception Contrôle ARC ***(encadré 1)***.

▶ Encadré 1 : Le module Accueil Réception Contrôle (ARC)

Le module Accueil Réception Contrôle (ARC) a été conçu comme un bloc d'infrastructures mutualisées pour l'ensemble des déclarations, mobilisé par le Système d'Information sur l'Emploi et les Revenus d'Activités. Il a été développé en *open source*, ce qui facilite sa mutualisation Il est aujourd'hui utilisé par d'autres systèmes à l'Insee et même en dehors, par l'Istat (Institut de statistique italien) notamment.

Ce module fonctionnel permet l'accueil, le contrôle et la transformation des données administratives en données statistiques élémentaires, avec une possibilité de filtrage en amont.

Le paramétrage du chargement (fichier plat ou fichier XML*), la reconnaissance des normes associées à des déclarations, les contrôles appliqués aux données et la transformation des données administratives en données statistiques (le *mapping*) peuvent être modifiés et améliorés au cours du temps de façon interactive par l'utilisateur. Étant donné les volumes de données très importants à traiter, la possibilité de modifier les traitements de façon interactive ainsi que d'éventuelles erreurs de spécification pourraient entraîner des risques sur la production.

**XML : Extensible Markup Language.*

Pour limiter ce risque, les utilisateurs disposent d'espaces de tests similaires mais distincts de l'espace de production, afin de tester et qualifier les nouvelles règles spécifiées et de mesurer leur impact sur les données chargées avant leur mise en production.

Pour que le module ARC puisse réceptionner et contrôler ces différents types de fichiers, il faut lui spécifier au préalable trois informations indispensables sur les caractéristiques des fichiers à charger et contrôler :

1. la famille spécifique de rattachement du fichier administratif : (exemple : DSN, Particuliers employeurs, etc.) ; ces familles distinguent les grands types de fichiers à contrôler.
2. la norme de rattachement du fichier administratif. La norme est un ensemble de caractéristiques décrivant le fichier (sous-ensemble de la famille), le plus souvent valable durant une période précise et qui permet d'étalonner les tests.
3. la périodicité du fichier reçu ; elle renseigne l'application sur la temporalité des fichiers devant faire l'objet du contrôle.

La suite de l'article sera illustrée à partir d'un exemple de fichier administratif fictif utilisé par l'Insee. Cet exemple est simpliste, mais permet de se rendre compte des différentes transformations possibles et d'illustrer la mise en place d'un « pipeline » pour l'intégration d'une source.

Imaginons que le contenu des données envoyées à l'Insee par fichier au format CSV (avec séparateur « ; ») soit le suivant :

▶ Exemple

- l'identifiant du foyer fiscal (**dirindik**) ;
- le taux d'imposition du foyer fiscal (**8OT**) ;
- l'adresse associée au foyer fiscal (**ZAFR1**) ;
- les nom (**LNCON**), prénom (**LNCOPF**), sexe (**LCCOT**), date de naissance (**DNCO**) et salaire (**1AJ**) du déclarant ;
- les nom (**LNCJN**), prénom (**LNEPP**), sexe (**LCPPT**), date de naissance (**DAEPAD**) et salaire (1BJ) du conjoint ;
- le nombre d'enfants du foyer fiscal (**7EA**).

Le début du fichier pourrait ressembler à ce qui suit :

```
dirindik;8OT;ZAFR1;LNCON;LNCOPF;LCCOT;DNCO;1AJ;LNCJN;LNEPP;LCPPT;DAE-
PAD;1BJ;7EA
570001;11;18 rue des bleuets 57 000
Metz;DURANT;Jonathan;1;01/12/55;10500;DURANT;Germaine;2;27/05/58;7500;3
570002;30;11 rue des Lilas 57000
Metz;DUPUIS;Sophie;2;30/02/74;30000;;;;;;2
570003;41;523, rue du Général De
Gaulle;WENDLING;Zoé;2;23/07/74;50000;HENRY;Jérôme;;;25000;
570004;42;47 rue des Plantes 57 000
Metz;THIERY;Robert;1;17/08/80;1000000000;;;;;;52
540005;0;23 rue Jordan 54 000 Nancy;DURANT;Jonathan;1;01/12/55;;;;;;;0
```

Dans notre exemple, le salaire d'un individu n'est affecté qu'à un seul foyer fiscal, mais un individu peut être affecté à plusieurs foyers fiscaux, s'il possède par exemple une résidence secondaire dans un département différent de celui dans lequel il a déclaré son salaire : c'est ici le cas de Durant Jonathan.

Renommer les variables : une nécessité pour mieux les comprendre

> ***Pour exploiter les variables provenant d'une source administrative, il est plus facile qu'elles aient un nom compréhensible.***

Pour exploiter les variables provenant d'une source administrative, il est plus facile qu'elles aient un nom compréhensible (par exemple : `salaire_decl`) plutôt que de conserver le nom du fichier d'origine qui correspond par exemple au nom de la case de la déclaration fiscale (`1AJ`) dans l'exemple.

Les noms des variables des fichiers externes sont souvent liés au contexte dans lequel elles ont été collectées et il est souhaitable de les renommer dès cette phase afin de faciliter leur utilisation par des statisticiens (qui n'ont pas forcément participé directement à la phase d'intégration des données). Renommer les données pour en faciliter l'appropriation par le statisticien n'est toutefois pas suffisant et s'accompagne de la saisie de métadonnées descriptives pour décrire plus précisément le concept, le format des variables, etc. Les nouveaux noms des variables se rapprochent alors des concepts statistiques qu'elles recouvrent. Ainsi dans notre exemple :

Exemple

• `dirindik` devient `id_foyer_fisc`	• `8OT` devient `tx_imposition`
• `ZAFR1` devient `adresse`	• `LNCON` devient `nom_decl`
• `LNCOPF` devient `prenom_decl`	• `LCCOT` devient `sexe_decl`
• `DNCO` devient `d_nais_decl`	• `1AJ` devient `salaire_decl`
• `LNCJN` devient `nom_conj`	• `LNEPP` devient `prenom_conj`
• `LCPPT` devient `sexe_conj`	• `DAEPAD` devient `d_nais_conj`
• `1BJ` devient `salaire_conj`	• `7EA` devient `nb_enfants`

Restructurer les données par unité statistique et les relier

Les unités de gestion administratives (compteur électrique, foyer fiscal) sont en général différentes des unités statistiques diffusées (logement, ménage, individu, entreprise, etc.). Il faut donc dériver ces unités statistiques si elles ne sont pas directement présentes dans la source administrative. Cette étape se traduit par l'agrégation de plusieurs enregistrements ou la suppression d'enregistrements pour créer une unité statistique. Par exemple, on peut regrouper des données à un niveau « ménage » à partir de données d'individu ou encore regrouper des données au niveau « individus » à partir de données de contrat de travail, etc. Ceci est possible si les données contenues dans la source externe le permettent grâce à la présence de l'identifiant du ménage dans les enregistrements d'individus, par exemple. Si un tel traitement nécessite des données externes à la source, cette opération se fera dans une phase ultérieure du processus.

Inversement, on peut vouloir éclater un enregistrement du fichier en entrée en plusieurs enregistrements dans le modèle de données statistiques (cas des individus dans les fichiers fiscaux structurés en foyers fiscaux). Il est alors nécessaire d'expliciter les liens entre différentes unités statistiques. Dans le cadre des données fiscales, faire le lien entre individus et foyers fiscaux d'une part, et foyers fiscaux et adresses d'autre part, permet *in fine* de constituer des ménages.

Dans l'exemple, le fichier envoyé « mélange » deux types d'unités statistiques au sein d'un même enregistrement.

Dans une première étape, on créerait deux tables différentes, individus et foyers fiscaux, dans le modèle de données « statistique ».

Les données sont intégrées dans les deux tables suivantes :

Table 1 : La table des foyers fiscaux

ID	TAUX IMPOSITION	ADRESSE	NB ENFANTS
570001	11	18 rue des bleuets 57000 Metz	3
570002	30	11 rue des Lilas 57000 Metz	2
570003	41	523, rue du Général De Gaulle	
570004	42	47 rue des Plantes 57000 Metz	52
540005	0	23 rue Jordan 54000 Nancy	0

Table 2 : La table des individus

NOM	PRÉNOM	SEXE	DATE DE NAISSANCE	SALAIRE	STATUT	ID FOYER
DURANT	Jonathan	1	01/12/55	10500	déclarant	570001
DURANT	Germaine	2	27/05/58	7500	conjoint	570001
DUPUIS	Sophie	2	30/02/74	30000	déclarant	570002
WENDLING	Zoé	2	23/07/74	50000	déclarant	570003
HENRY	Jérôme			25000	conjoint	570003
THIERY	Robert	1	17/08/80	1000000000	déclarant	570004
DURANT	Jonathan	1	01/12/55		déclarant	540005

Une jointure de ces deux tables sur la variable id_foyer présente dans les deux tables permet de reconstituer les données initiales. On ne change pas l'information, on la restructure.

Pseudonymiser les données

Les sources administratives peuvent contenir des informations nominatives (NIR[5], nom, prénoms etc.) inutiles à la production statistique et qu'il faut supprimer dès cette phase afin de respecter la confidentialité des données au plus tôt dans le processus de production. On pseudonymise. Pour séparer au plus tôt les données d'état civil (nom, prénom, date et lieu de naissance et adresse) des données « métier » utiles pour produire des statistiques (information sur l'emploi, le revenu, etc.), on apparie les données d'état civil de la source avec un référentiel d'identité pour remplacer, dans les données intégrées, les données d'état civil par l'identifiant leur correspondant dans ce référentiel[6].

La mise en œuvre de cette pseudonymisation dans notre exemple conduit à scinder la table des individus en deux : une table nommée « individu_etat_civil » contenant les données d'état civil et une deuxième (« individu_salaire ») contenant les autres variables. Le lien entre les deux tables se fait par le biais de la variable id_ind qui identifie un individu de façon univoque selon son nom, son prénom et sa date de naissance. On constate que la table individu_etat_civil ne contient que 6 lignes car DURANT Jonathan est en double dans le fichier.

Table 1 : La table individus_etat_civil

ID_IND	NOM	PRÉNOM	SEXE	DATE DE NAISSANCE
1	DURANT	Jonathan	1	01/12/55
2	DURANT	Germaine	2	27/05/58
3	DUPUIS	Sophie	2	30/02/74
4	WENDLING	Zoé	2	23/07/74
5	HENRY	Jérôme		
6	THIERY	Robert	1	17/08/80

Table 2 : La table individu_salaire (pseudonymisée)

ID_IND	SALAIRE	STATUT	ID FOYER
1	10500	déclarant	570001
2	7500	conjoint	570001
3	30000	déclarant	570002
4	50000	déclarant	570003
5	2500	conjoint	570003
6	1000000000	déclarant	570004
1		déclarant	540005

5 Numéro d'identification au répertoire national des personnes physiques (RNIPP), plus connu comme le « numéro de sécurité sociale ».
6 Voir l'article de Yves-Laurent Bénichou, Lionel Espinasse et Séverine Gilles sur le Code statistique non signifiant (CSNS) dans ce même numéro.

Recoder les valeurs des variables

> ***Les sources administratives peuvent utiliser des nomenclatures différentes de celles utilisées pour la statistique.***

Les sources administratives peuvent utiliser des nomenclatures différentes de celles utilisées pour la statistique. Cette étape permet de se conformer à ces dernières à condition que cette transformation soit totalement automatisée (passage de « H » à 1 pour le sexe par exemple). Pour les enregistrements où une intervention humaine complémentaire est nécessaire (codification de la PCS[7] par exemple), un indicateur de reprise peut être posé durant l'intégration et le recodage « manuel » sera reporté à une phase ultérieure du processus de production.

Ce traitement concerne directement les valeurs d'une ou plusieurs variables du fichier en entrée. Il peut s'agir :

- d'une harmonisation des valeurs manquantes ou « refuge »[8] dans le cas où les valeurs manquantes sont traitées de façon différente selon les variables. Ceci facilitera leur repérage dans la phase de redressement[9] ;
- d'une correction des valeurs aberrantes. Ce processus peut être reporté dans les phases aval du processus statistique mais s'il est mis en œuvre dès l'intégration, il doit pouvoir être fait sans intervention humaine. Dans l'exemple, il s'agirait de traitement simple comme, par exemple, la mise à valeur manquante d'une date erronée (30 février) ;
- de la normalisation d'une variable. Il peut s'agir de supprimer des caractères spéciaux, du passage en lettres majuscules, etc. Ce traitement peut concerner des variables utiles pour l'identification afin de s'assurer que les règles de normalisation sont les mêmes dans le fichier à identifier et le référentiel auquel il doit être confronté. De tels traitements sont mis en œuvre pour l'attribution du Code Statistique Non Signifiant par exemple.

Calculer des variables dérivées

Pour ses études, le statisticien utilise souvent des nomenclatures (tranche d'âge, catégorie d'entreprises etc.). Le calcul de variables dérivées permet, par exemple, de passer d'une année de naissance contenue dans la source administrative à une tranche d'âge.

> ***Créer des nouvelles variables qui vont être utiles dans les phases ultérieures.***

Il s'agit de créer des nouvelles variables qui vont être utiles dans les phases ultérieures. Par exemple, créer une variable statistique par agrégation ou éclatement de variables en entrée. Le revenu peut être défini comme la somme de différents postes de la liasse fiscale, l'adresse peut être décomposée en plusieurs champs, etc.

7 PCS : nomenclature des professions et catégories socioprofessionnelles.
8 Le jour de naissance est par exemple mis à 00 plutôt que laissé à blanc.
9 Il faut bien distinguer cette étape, qui reste au niveau de la représentation des variables, des traitements d'imputation qui viendront plus tard dans le processus et qui eux s'appuient sur un modèle statistique. Imaginons que pour une variable donnée, il y ait plusieurs façons de représenter la valeur manquante (un blanc, un 00 etc.). Lors de l'intégration, ces valeurs seront harmonisées (toutes mises à blanc par exemple). Leur imputation n'interviendra qu'ultérieurement.

Dans l'exemple, des recodages et des créations de variables dérivées automatiques ont été faits. Ils ne demandent aucune intervention humaine :

- codage du sexe en : 1 = « M » et 2 = « F », autre = « blanc » ;
- calcul d'une nouvelle variable (salaire_foyer) dans la table foyer_fiscaux et qui correspond à la somme des salaires des déclarants et conjoints ;
- calcul d'une variable permettant de juger de la vraisemblance du salaire du foyer. Cette variable vaut 1 si le salaire du foyer est inférieur à 10 millions d'€ et vaut 0 sinon. À noter qu'aucune correction n'est effectuée lors de cette phase du processus. Les redressements manuels ou automatiques seront réalisés dans les phases ultérieures du traitement. Ce calcul n'est pas forcément appliqué à toutes les variables. Dans l'exemple, on laisse la valeur aberrante des 52 enfants. Cette valeur sera corrigée dans les traitements postérieurs à la phase d'intégration ;
- calcul du nombre de membres du foyer, qui est égal au nombre d'enfants + 2 si un conjoint est présent et au nombre d'enfants + 1 sinon ;
- calcul d'une variable « unique » pour les individus. Cette variable vaut 1 si l'identifiant de l'individu n'est présent qu'une fois dans la table individus_salaire et 0 sinon.

Les tables obtenues à l'issue de cette phase d'intégration sont les suivantes :

Table 1 : La table foyers_fiscaux

ID	TAUX IMPOSITION	ADRESSE	NB ENFANTS	NB FOYER	SALAIRE	SALAIRE COHÉRENT
570001	11	18 rue des bleuets 57000 Metz	3	5	18000	1
570002	30	11 rue des Lilas 57000 Metz	2	3	30000	1
570003	41	523, rue du Général De Gaulle	0	2	75000	1
570004	42	47 rue des Plantes 57000 Metz	52	53	1000000000	0
570005	0	23 rue Jordan 54000 Nancy	0	1		1

Table 2 : La table individus_etat_civil

ID IND	NOM	PRÉNOM	SEXE	DATE DE NAISSANCE
1	DURANT	Jonathan	M	01/12/55
2	DURANT	Germaine	F	27/05/58
3	DUPUIS	Sophie	F	30/02/74
4	WENDLING	Zoé	F	23/07/74
5	HENRY	Jérôme		
6	THIERY	Robert	M	17/08/80

Table 3 : La table individu_salaire (pseudonymisée) :

ID IND	SALAIRE	STATUT	ID FOYER	UNIQUE
1	10500	déclarant	570001	non
2	7500	conjoint	570001	oui
3	30000	déclarant	570002	oui
4	50000	déclarant	570003	oui
5	2500	conjoint	570003	oui
6	1000000000	déclarant	570004	oui
1		déclarant	540005	non

Mettre en place des filtres

> ***Le champ de la source en entrée peut être beaucoup plus vaste que le champ d'intérêt statistique.***

Le champ de la source en entrée peut être beaucoup plus vaste que le champ d'intérêt statistique. Dans ce cas, si la source contient des variables permettant de filtrer la population d'intérêt, il est possible de le faire dès l'étape d'intégration afin d'alléger les bases statistiques. Ceci permet également de respecter les principes de minimisation et de nécessité qui obligent tout statisticien à se limiter autant que possible aux données dont il a un réel besoin. Par exemple, supprimer les locaux commerciaux d'un fichier administratif lorsqu'on ne s'intéresse qu'aux locaux d'habitation.

Dans l'exemple, les lignes des individus en double ayant un montant de salaire égal à valeur manquante ne sont pas intéressantes. La variable unique est recalculée sur cette nouvelle table.

Au final la table individu_salaire est donc la suivante :

Table : La table individu_salaire

ID IND	SALAIRE	STATUT	ID FOYER	UNIQUE
1	10500	déclarant	570001	non
2	7500	conjoint	570001	oui
3	30000	déclarant	570002	oui
4	50000	déclarant	570003	oui
5	2500	conjoint	570003	oui
6	1000000000	déclarant	570004	oui

▶ Construire le *pipeline* de données

La phase d'intégration des sources doit être totalement automatisée et reproductible. Elle consiste à mettre en place un cadre général pour une démarche systématique sur des données structurées par un producteur externe. Elle doit se traduire par la mise en place d'un *pipeline* jalonné de points de contrôle pour s'assurer que la succession des tâches se déroule correctement et d'arrêter le processus dès qu'un éventuel problème est rencontré.

L'objectif de cette industrialisation est de mettre en place l'enchaînement de traitements génériques paramétrables les plus indépendants possibles de la source en entrée. L'intégration d'une nouvelle source (ou la mise à jour d'une source existante) revient alors à décrire la source en entrée, le modèle de données souhaité en sortie et les transformations permettant de passer de l'un à l'autre. L'enchaînement des étapes d'intégration des données se fait ensuite de façon automatique à partir de ces métadonnées qui deviennent actives. Ainsi, la spécification de la représentation attendue pour une variable (M, F pour la variable sexe par exemple) permet de générer automatiquement des traitements de contrôle des valeurs.

Dans ces conditions, certaines métadonnées plus précises ont un intérêt particulier dans le cadre des transformations de données. En amont, la documentation des traitements peut aller jusqu'à leur spécification dans un langage plus ou moins formel, par exemple BPMN[10] ou graphe acyclique direct (DAG)[11] pour le processus d'ensemble, SQL ou VTL ***(encadré 2)*** pour les transformations elles-mêmes. Cela présente l'avantage de permettre l'automatisation complète ou partielle du processus dans une démarche de métadonnées actives. La spécification en VTL est une métadonnée (elle décrit la transformation dans un langage compréhensible par statisticien) qui est ensuite implémentée automatiquement à l'aide d'outils informatiques ce qui la rend donc directement active. En aval, le traçage précis des opérations (quelles variables ont servi pour le calcul de telle autre, par exemple), permet une meilleure maîtrise du processus et favorise la reproductibilité et la transparence. On parle de métadonnées de provenance ou de lignage *(data lineage)*.

> ***Ces métadonnées actives permettent au concepteur d'être le plus autonome possible et d'adapter plus facilement son pipeline.***

Ces métadonnées actives permettent au concepteur d'être le plus autonome possible et d'adapter plus facilement son *pipeline* aux évolutions des sources externes. Par exemple, la DGFiP a récemment modifié sa façon d'enregistrer les dépendances des maisons dans son système d'information ; par le passé, elles étaient liées à un local principal, maintenant elles sont considérées comme des logements à part entière. Une telle mise à jour conduit donc à revoir le filtre de définition du champ, voire à ajouter de nouvelles variables. Par le biais des métadonnées actives, le statisticien peut créer une nouvelle variable dans le modèle de données d'accueil de RMéS[12] (*Bonnans, 2019*) qui permet de repérer les dépendances, d'établir le mode de calcul de cette nouvelle variable dans ARC à partir des variables du fichier de la DGFiP et de modifier son script

10 Le Business Process Model and Notation (BPMN) est la norme standard pour la modélisation de processus métier.
11 Un graphe acyclique direct ou orienté (directed acyclic graph ou DAG) est une façon standard de décrire des processus.
12 Référentiel de métadonnées statistiques.

VTL de définition du champ pour prendre en compte cette nouvelle variable. Le statisticien est alors totalement autonome pour modifier son *pipeline*, prendre en compte la modification de structure du fichier en entrée et les tracer.

Dans le *pipeline* relatif aux différentes étapes de l'intégration des données ***(figure)***, le statisticien saisit en amont les métadonnées (définition des variables, attachement à un concept statistique, règle de transformation, etc.). Les étapes s'enchaînent ensuite automatiquement à partir de ces métadonnées, conduisant à une standardisation et une séparation claire entre spécification logique et réalisation technique et permettent une plus grande autonomie du statisticien.

En sortie de ce processus, les données administratives sont donc intégrées dans le système d'information sous forme de données brutes. En outre, une évaluation de la qualité est également produite en parallèle afin de disposer de premiers indicateurs qualité (nombre d'enregistrements, taux de non-réponse partielle, etc.) qui permettent de donner des premiers éléments sur la qualité des données brutes.

▶ Encadré 2 : VTL, langage de validation et de transformation de données

VTL (*Validation and transformation language*) est un langage permettant de spécifier des traitements de validation ou de transformation de données développé dans le cadre du standard SDMX*, norme d'échange de données et métadonnées statistiques. Destiné aux statisticiens, il fournit une vue neutre (indépendante de l'implémentation technique) du processus de données au niveau métier. En tant que langage de spécification, VTL est suffisamment riche et expressif pour définir des traitements relativement compliqués.

VTL possède des caractéristiques qui le rendent particulièrement intéressant dans un contexte d'industrialisation et d'automatisation des processus statistiques. VTL est donc adéquat pour l'intégration de sources externes.

Tout d'abord, comme VTL se positionne au niveau logique, intermédiaire entre le concept et l'implémentation, il n'est pas directement exécutable comme peuvent l'être Java, R ou Python. Les expressions VTL doivent être transmises à un moteur qui l'exécutera sur une plate-forme de plus bas niveau, par exemple Java, Python ou C#. Ceci permet une claire séparation des préoccupations (ou **séparation des responsabilités****) entre le statisticien qui se concentre sur la spécification des traitements et l'informaticien qui se charge de l'implémentation. Dans les langages directement exécutables, la formulation logique du traitement est souvent noyée dans les détails d'implémentation et difficile à reconstituer. Avec VTL, la spécification est traitée comme un objet en soi, qui peut donc être géré, versionné, tracé, partagé, documenté, etc.

Une autre propriété intéressante de VTL est d'être basée sur un modèle de données qui dérive des standards internationaux (GSIM, SDMX, DDI***) et qui est adapté à la statistique et à différents types de données (détaillées, agrégées, qualitatives, quantitatives, etc.). Au cœur du modèle se trouve le *Data Set*, composé de *components* (les colonnes dans un fichier tabulaire) jouant différents rôles (identifiants, mesures et attributs) et de lignes (*Data Points*). Ce modèle permet de simplifier les expressions : ainsi, si l'on effectue une somme sur un jeu de données, il n'est pas utile de préciser que l'opération ne s'applique qu'aux mesures et non aux identifiants ou aux attributs.

Enfin, VTL est décrit par une grammaire formelle, qui assure l'assise logique du langage et permet de l'exploiter de façon automatisée, notamment par la construction d'outils comme des éditeurs ou des moteurs d'exécution pour différentes plates-formes techniques. Ceci assure qu'une même expression sera exécutée de façon cohérente dans différents langages de plus bas niveau.

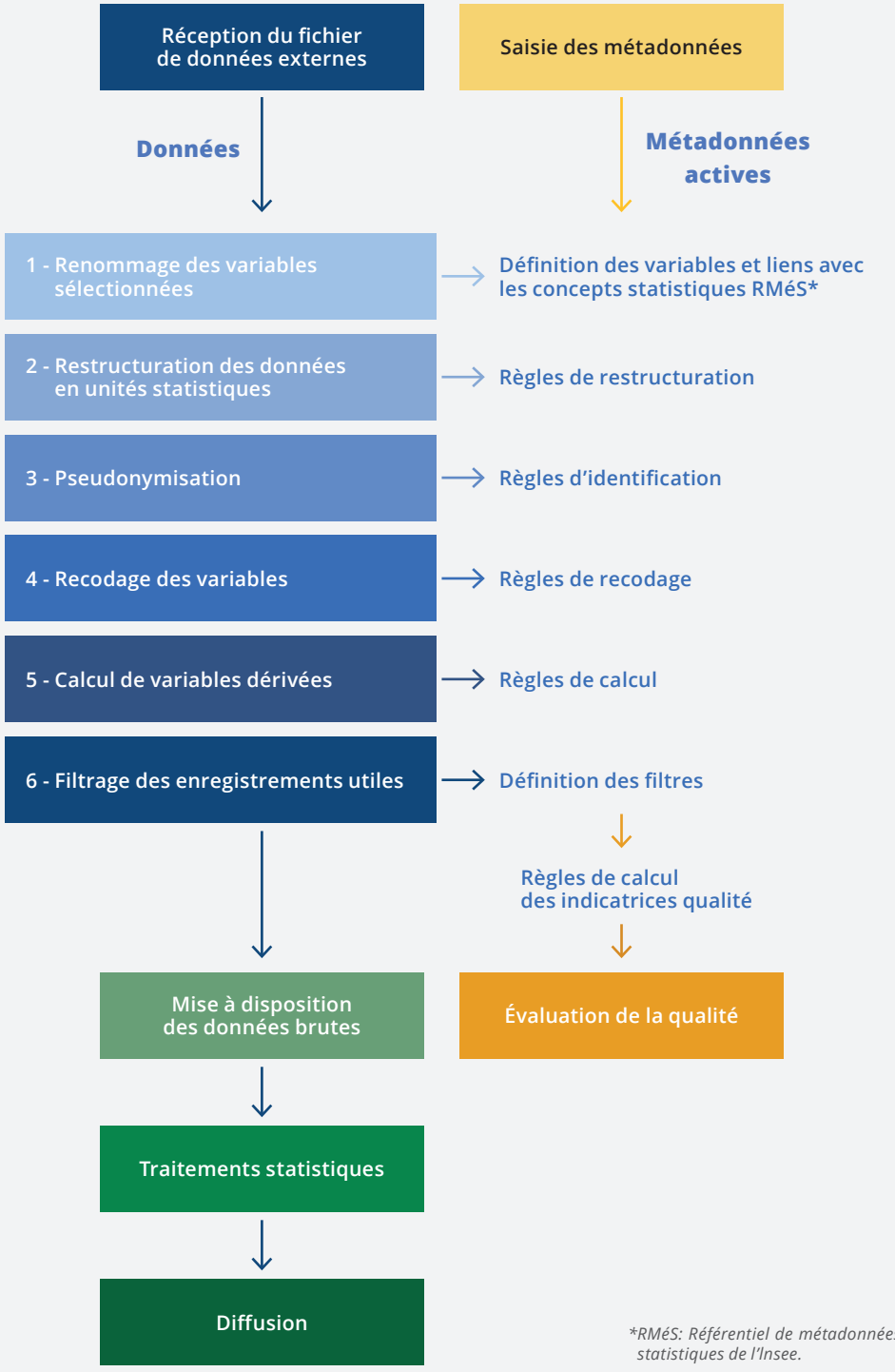
De nombreux outils ont été développés autour de VTL par les communautés statistiques et bancaires. Parmi les intervenants les plus actifs, on peut citer la Banque d'Italie, l'Insee ainsi que des sociétés privées.

* *SDMX pour Statistical Data and Metadata eXchange.*

** *https://fr.wikipedia.org/wiki/S%C3%A9paration_des_pr%C3%A9occupations.*

*** *GSIM pour Generic Statistical Information Model et DDI pour Data Documentation Initiative.*

Figure - Le pipeline de données et de métadonnées



RMéS: Référentiel de métadonnées statistiques de l'Insee.

▶ Intégrer les données c'est bien mais avec les métadonnées c'est mieux…

> ***L'existence de métadonnées complètes et exactes est fondamentale pour le bon déroulement des opérations.***

Pour les transformations de données comme pour toute étape du processus statistique, l'existence de métadonnées complètes et exactes est fondamentale pour le bon déroulement des opérations. Plus encore, comme les traitements de transformation se situent souvent au tout début du processus, il faut capturer ou créer les métadonnées qui seront réutilisées dans les étapes ultérieures.

En règle générale, les données administratives sont décrites avec leurs propres métadonnées. Certaines pourront être reprises mais pas forcément toutes. Et les métadonnées statistiques issues du processus d'intégration sont différentes de celles d'origine (les nomenclatures peuvent être différentes, les concepts statistiques associés sont spécifiques aux données statistiques, etc.).

Différents types de métadonnées revêtent une importance particulière dans le cadre de la transformation de données.

Les métadonnées peuvent, elles-mêmes, être l'objet d'une transformation pour accompagner la transformation des données. Par exemple, le renommage des variables peut être vu comme une transformation des métadonnées.

Les métadonnées de structure, à savoir la définition des variables utilisées, les concepts statistiques auxquels elles se raccrochent, le type et le domaine de valeurs, parfois contraints par des listes de codes, doivent être définies au plus tôt. Il est important qu'elles soient disponibles dans des formats qui permettent aisément d'en automatiser l'utilisation pour valider les traitements effectués. Par ailleurs, les métadonnées descriptives constituent une autre catégorie importante de métadonnées. On range ici tout ce qui permet de faciliter la découverte et la compréhension des données, d'évaluer leur qualité ou leur adéquation à une utilisation spécifique, etc. La documentation des traitements et des méthodes utilisées et les conditions de fourniture des données peuvent également être considérées comme des métadonnées descriptives.

▶ … avec une mesure de la qualité au plus tôt

> ***Il est important d'avoir une première mesure de la qualité des données dès leur intégration.***

Même si la source a été qualifiée en amont et que sa qualité a été jugée suffisante pour produire des statistiques, il est important d'avoir une première mesure de la qualité des données dès leur intégration (*Six et Kowarick, 2022, UNECE Statswiki Quality*). Ceci permet d'avoir une première idée de la qualité intrinsèque de la source et de pouvoir enclencher au plus tôt un processus d'amélioration continue de la qualité. Cette phase permet également de se familiariser avec les nouvelles sources en comprenant mieux leur structure et leur contenu.

Au niveau de l'intégration, cette mesure de la qualité (communément appelée *data profiling* dans la littérature internationale) consiste à définir des indicateurs qui permettront d'avoir une première évaluation de la qualité de la source et de suivre son évolution au cours du temps, en comparant les livraisons successives de la source (*monitoring*). Ces indicateurs sont par exemple :

- le nombre d'enregistrements reçus par type d'unité statistique ;
- les totaux de variables d'intérêt ;
- le taux de non-réponse partielle par variable ;
- la fréquence des modalités dans le cas de variables qualitatives, permettant d'identifier des modalités peu ou pas utilisées ;
- la distribution et mise en évidence de valeurs aberrantes ;
- l'identification des codes non utilisés pour les variables associées à une nomenclature ;
- l'identification de doublons, etc.

L'analyse de ces indicateurs nécessitera souvent, au moins lors des premières réceptions de fichiers, de demander au producteur des précisions. Ceci peut également conduire à ce que le producteur ajoute des contrôles dans son propre processus afin d'améliorer la qualité de la collecte. Par exemple les travaux menés à l'Insee lors de l'accueil de la déclaration sociale nominative (DSN) ont conduit le GIP-MDS[13] à ajouter un contrôle du NIR dans son interface de saisie des données de la DSN, ce qui a amélioré notablement la qualité du NIR dans la DSN. De tels indicateurs permettent de repérer au plus tôt si les données fournies sont incomplètes, lorsque le nombre d'enregistrements ou les totaux de certaines variables sont plus faibles que lors des livraisons précédentes. Ces indicateurs servent également à identifier des problèmes qui pourront être corrigés par les processus ultérieurs (exemple de l'indicateur qualité du salaire du foyer qui permet d'identifier le salaire d'un milliard d'euros qui est suspect). Ces indicateurs « qualité » seront donc utiles à l'utilisateur des données brutes pour piloter ses propres traitements.

Cette phase étant primordiale, des outils spécifiques ont été développés.

▶ Perspectives

> ***… changement de nature et de «propriétaire» de la donnée, qui devient donc statistique, avec le cadre légal, méthodologique et organisationnel correspondant.***

L'intégration des données administratives dans un processus statistique est une phase essentielle dont l'importance est reconnue au niveau international ***(encadré 3)***. Elle doit être soigneusement identifiée, spécifiée et outillée. Elle correspond à un changement de nature et de « propriétaire » de la donnée, qui devient donc statistique, avec le cadre légal, méthodologique et organisationnel correspondant.

13 Groupement d'intérêt public Modernisation des déclarations sociales.

Elle doit se traduire par l'implémentation d'un *pipeline* documenté et reproductible de transformations élémentaires ; il est souhaitable de le définir dans un formalisme indépendant d'une technologie particulière. Lors de la spécification de ce *pipeline*, il est important de prendre en compte les métadonnées, que ce soient celles de la source, qui peuvent elles-mêmes faire l'objet de transformations, ou celles qui découlent du processus d'intégration comme les métadonnées de provenance. Il faut également, comme pour tout processus, produire des indicateurs de qualité qui permettront de contrôler et d'améliorer les traitements.

Au-delà des données administratives, l'utilisation de données externes pour la statistique est appelée à se développer, et la mise en place d'un cadre méthodologique et d'outils communs pour l'acquisition de données permettra d'industrialiser cette fonction, à l'instar de ce qui a pu être fait pour la filière d'enquêtes à l'Insee (*Cotton et Dubois, 2019 ; Koumarianos et Sigaud, 2019*). Comme pour cette dernière, le pilotage des processus par l'activation des métadonnées, conduisant à une standardisation et une séparation claire entre spécification logique et réalisation technique, permettra une plus grande autonomie du statisticien et une meilleure réactivité des systèmes, importante pour s'adapter aux changements externes.

▶ Encadré 3 : Transformation de données, le cadre international

La collaboration internationale pour la **modernisation de la statistique officielle*** pilotée par l'Unece s'est initialement attachée à une vision orientée processus avec en particulier le modèle GSBPM, puis le moins connu GSDEM (2015). Les aspects relatifs aux données sont arrivés sur le devant de la scène dans le cadre de différentes initiatives, d'abord sous l'égide de l'Unece :

- projet «*Data Integration*» (2016) : il met l'accent sur l'intégration de données (après des réflexions sur le *Big data*), définie comme l'activité consistant à combiner au moins deux sources de données différentes dans un ensemble de données. La transformation des données n'y est pas définie comme une étape en soi, l'accent étant plutôt mis sur les méthodes d'appariement ;
- conception d'une architecture de données statistique commune (**CSDA**** 2017 – 2018) : son ambition est de « soutenir les organismes statistiques dans la conception, la collecte, l'intégration, la production et la diffusion de statistiques officielles basées à la fois sur les types de sources de données traditionnelles et nouvelles. » Elle s'appuie sur des **principes** courants dans les démarches modernes orientées données (donnée en tant qu'actif, accessibilité, ré utilisabilité, emploi de modèles standard), et définit les «capacités» (*capabilities*) nécessaires pour utiliser et gérer données et métadonnées statistiques. La transformation des données est une des capacités de haut niveau ;
- projet «*Data Governance Framework*» (2022, piloté par le Mexique) : il couvre certains domaines abordés ici, tels que gouvernance des données, qualité, métadonnées, etc.

L'évolution est similaire dans le Système statistique européen (SSE) :

- projet ESSnet ISAD (*Integration of Survey and Administrative Data*) terminé fin 2008 : il mentionne la transformation de données dans la phase de préparation des données avant l'intégration (par exemple, par appariement de fichiers) ;
- puis ESSnet «*Data Integration*» ou projet stratégique «ADMIN» (*ESS2020 ADMIN*) : les traitements de transformation des sources sont inclus dans une phase de préparation des données, peu détaillée, bien que connue pour mobiliser une part importante du travail statistique ;
- ESSnets *Big Data* I et II (2015 à 2021) : l'architecture métier élaborée identifie une fonction de *data wrangling*, c'est-à-dire « la possibilité de transformer les données du format source d'origine en un format cible souhaité, mieux adapté à une analyse et à un traitement ultérieur », et une fonction de représentation des données, c'est-à-dire l'ajout d'éléments de contexte et de structure (données dérivées, codes, catégories) aux données brutes. Ces éléments sont positionnés dans une couche de convergence des données, entre la couche des données brutes et celle des données statistiques ;
- démarche stratégique «*Trusted Smart Statistics*» : les récents développements lancés dans ce cadre réutilisent l'architecture définie par les projets *Big Data*.

À travers ces différents exemples, la problématique de transformation des données est clairement identifiée en tant que telle dans les instances internationales.

* *https://unece.org/statistics/modernization-official-statistics*.

***https://statswiki.unece.org/display/DA/CSDA+2.0*.

La spécification au niveau logique des traitements d'intégration, qui restent encore largement organisés en silos, permettra aussi de les rendre plus modulaires et mieux partageables. Dans cette optique, on peut envisager le déport en amont de certaines étapes chez le producteur de données (par exemple des traitements de pseudonymisation ou de recodification sur des smartphones avant le transfert de l'information). Cette problématique est notamment rencontrée dans le cadre des *smart statistics*[14]. Enfin, cette phase d'intégration automatisée des données prend encore plus son sens aujourd'hui avec la disponibilité massive de données protéiformes provenant de capteurs, d'outils numériques voire de réseaux électroniques/sociaux qui caractérisent la réalité sociale, environnementale et économique. Reste à s'assurer qu'on peut les utiliser pour produire des statistiques de qualité.

14 *https://www.cambridge.org/core/journals/data-and-policy/article/trusted-smart-statistics-how-new-data-will-change-official-statistics/380C6B6408D84C16164F33A1F4BF2F07.*

▶ Bibliographie

- BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 46-57. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168396?sommaire=416841*.

- COTTON, Franck et DUBOIS, Thomas, 2019. Pogues, un outil de conception de questionnaires. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N° N3, pp. 17-28. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254216?sommaire=4254170*.

- COURMONT, Antoine, 2021. *Quand la donnée arrive en ville. Open data et gouvernance* urbaine. Presses universitaires de Grenoble. ISBN 2706147350.

- CROS, 2014. *Handbook on Methodology of Modern Business Statistics. In : site de Collaboration in Research and Methodology for Official Statistics*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://cros-legacy.ec.europa.eu/content/handbook-methodology-modern-business-statistics_en*.

- DURR, Jean-Michel, DUPONT Françoise, HAAG, Olivier et LEFEBVRE, Olivier, 2022. *Setting up statistical registers of individuals and dwellings in France: Approach and first steps.* In : Statistical Journal of the IAOS (SJIAOS). Volume 38, n°1, pp. 215-223. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210916*.

- ESS Vision 2020 ADMIN (Administrative data sources). In : *site de Collaboration in Research and Methodology for Official Statistics*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://cros-legacy.ec.europa.eu/content/ess-vision-2020-admin-administrative-data-sources*.

- FERMOR-DUNMAN, Verena, and PARSONS Laura, 2022. *Data Acquisition processes improving quality of microdata at the Office for National Statistics. Q2022 Vilnius*.

- HAND, David, 2018. *Statistical challenges of administrative and transaction data. In : Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 9 février 2018. Volume 181, N°3, pp. 555-605. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.jstor.org/stable/48547504*.

- KOUMARIANOS, Heïdi et SIGAUD, Éric, 2019. Eno, un générateur d'instruments de collecte. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N° N3, pp. 29-44. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254218?sommaire=4254170*.

- LAMARCHE, Pierre et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 28-46. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- SIX, Magdalena et KOWARIK, Alexander, 2022. *Quality Guidelines for the Acquisition and Usage of Big Data*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://lsvdv-my.sharepoint.com/:w:/g/personal/ingabal_stat_gov_lt/Evmr-Cle195BrkRRDtPr2OMBkJsZqEI7Arzy8H2auuaTPw?rtime=TaOXDVpp2kg*.

- UNECE Integration, statswiki, *A Guide to Data Integration for Official Statistics. In : site statswiki de l'UNECE*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://statswiki.unece.org/display/DI/Guide+to+Data+Integration+for+Official+Statistics*.

- UNECE quality, statswiki, Quality. In : *site statswiki de l'UNECE*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://statswiki.unece.org/display/DI/Quality*.

- UNECE GSBPM. In : *site statswiki de l'UNECE*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://statswiki.unece.org/display/GSBPM*.

Une norme d'échange pour alimenter des référentiels et en assurer la qualité

Bertrand Dubrulle, Olivier Rosec**, Christian Sureau****

Améliorer la qualité du service pour les assurés de la sécurité sociale, c'est-à-dire disposer d'une information fiable, faciliter les interactions ou servir les prestations à bon droit dans les meilleurs délais, a nécessité la création de grands référentiels et un échange accru de flux de données. Pouvoir garantir la qualité de ces flux et alimenter correctement les référentiels est essentiel. Ainsi, la Cnav a défini et appliqué les principes des "normes d'échange" pour à la fois fournir une description claire et détaillée de l'échange de données et contrôler automatiquement les flux, afin de garantir la qualité de l'alimentation des référentiels. La norme représente un apport essentiel pour le statisticien qui maîtrise ainsi les objets grâce à la documentation structurée et peut alors utiliser des données de meilleure qualité. Les principes d'une norme d'échange et sa structure sont présentés. Un outillage lui est associé pour disposer d'un système générique, cohérent et réactif dans un contexte de forte évolutivité des réglementations. Ces principes et le logiciel associé ont permis de créer un outil de conversion de flux, d'une structure à une autre, en bénéficiant des avantages susmentionnés. La normalisation des échanges ainsi outillée est aujourd'hui effective dans deux cas : la Déclaration sociale nominative et l'alimentation du Répertoire de gestion de carrières unique (RGCU). Avec un recul d'une douzaine d'années la question se pose d'un déploiement à plus grande échelle et de son positionnement dans la stratégie de management de la donnée.

Improving the quality of service for those insured by social security, i.e. having access to reliable information, facilitating interactions or providing benefits as quickly as possible, has required the creation of major benchmarks and increased data exchange. It is essential to ensure the quality of these data flows and correctly feed the repositories. Thus, the Cnav has defined and applied the principles of "exchange standards" to both provide a clear and detailed description of the exchange and automatically control the flows in order to guarantee the quality of the input to the repositories. The exchange standard represents a main information for the statistician who has a better control over data thanks to the structured documentation. The principles of an exchange standard and its structure are presented. Cnav has developed tools to have a generic, coherent and reactive system in a context of frequent changes of regulations. These principles and this tool have made it possible to create another generic tool for converting a flow, from one structure to another, while benefiting from the advantages of the previous one. This standardization of exchanges is now effective in two cases: the Nominative Social Declaration and the feed of the Répertoire de Gestion des Carrières Unique (RGCU). With a decade of hindsight, the question arises of a larger-scale deployment and its positioning in the data management strategy.

* Adjoint au directeur de la Gestion de la Donnée (DGD), DSI Caisse nationale de l'assurance vieillesse (Cnav), *bertrand.dubrulle@cnav.fr*

** Responsable du pôle Expertise Technique Saturne et EBX, DGD, DSI, Cnav, *olivier.rosec@cnav.fr*

*** Directeur de la Gestion de la Donnée, DSI Cnav, *christian.sureau@cnav.fr*

Le domaine de la protection sociale se caractérise par des systèmes d'information complexes nécessitant de fréquents échanges d'information. Dès février 2016, les nombreux échanges de données volumineuses entre les régimes et organismes de la protection sociale étaient cités dans le rapport IGAS-IGF[1] *(Gratieux et Le Gall, 2016)*. Par exemple, en 2015, la Caisse nationale d'assurance vieillesse (Cnav) a comptabilisé, sur les 11 premiers mois, 1881 échanges avec 381 partenaires, correspondant à un total de 1,152 million de fichiers reçus ou émis. Assurer la qualité de ces échanges et la cohérence des informations devenait alors nécessaire pour alimenter correctement les référentiels.

> ***1881 échanges avec 381 partenaires, correspondant à un total de 1,152 million de fichiers reçus ou émis.***

Pour la branche retraite, plus de 20 millions d'actifs et 15 millions de retraités sont gérés, avec des flux de 50 millions d'éléments de carrière en provenance des entreprises, 30 millions en provenance de Pôle emploi, 10 millions en provenance de la branche maladie mais aussi plusieurs dizaines de millions en provenance ou à destination des autres organismes de retraite.

Toutes ces informations sont utilisées pour calculer les droits et servir aux assurés des prestations exactes. Pour cela, des informations complètes et de qualité sont essentielles afin de disposer d'éléments certains pour exécuter des calculs corrects. Pour stocker et centraliser les données et en garantir la qualité, la création de référentiels s'est imposée.

Compte tenu du nombre d'organismes, du nombre d'échanges et du volume de données, définir les principes de normalisation des échanges et les appliquer est essentiel pour garantir la qualité d'alimentation des référentiels, en extraire toute la plus-value et offrir des services à valeur ajoutée à toute la sphère sociale.

Les aspects juridiques, d'infrastructure, la notion de norme d'échanges et ses principes sont présentés. Les différents composants de la structure en blocs de la norme sont décrits ainsi que les dimensions fonctionnelles et techniques de ces blocs et des données qui les composent (les « rubriques »). On explicite ensuite les différents types de contrôles automatiques sur les flux. Un ensemble d'outils, nommé Saturne, a été développé par la Cnav pour générer automatiquement ces contrôles et d'autres livrables. On explique cet outillage et son rôle majeur pour reproduire la démarche et disposer de la réactivité dans la prise en compte des évolutions. Pour conclure, les apports après 12 ans d'expérience sont détaillés ainsi que les perspectives.

▶ Des aspects juridiques jusqu'à la norme d'échanges : des conventions entre les acteurs…

Les échanges se présentent sous différentes dimensions. Pour identifier les acteurs et le niveau de service lié à l'échange (fréquence, volume, délai de traitement…), ils sont encadrés par une convention juridique qui précise l'infrastructure technique permettant de garantir la robustesse et la traçabilité et définit les éléments fonctionnels liés à la structure des données et à la qualité attendue.

1 Inspection générale des affaires sociales et Inspection générale des finances.

Les données doivent être accessibles, échangées uniquement entre acteurs autorisés et seulement sur le périmètre correspondant à leur besoin d'utilisation. Les échanges sont encadrés par une convention de service fixant les objectifs et modalités (fondements juridiques, nature, modalités de transmission et de conservation des données, sécurité et traçabilité des échanges, échéances à respecter, règles de protection des données, conditions financières et modalités de suivi de la convention).

> ***Les données doivent être accessibles, échangées uniquement entre acteurs autorisés et seulement sur le périmètre correspondant à leur besoin d'utilisation.***

Par exemple, les données des déclarations employeurs issues de la déclaration sociale nominative (DSN[2]) *(Humbert-Bottin E., 2018)* et les autres ressources (exemple : retraite) sont transmises et partagées mensuellement pour le prélèvement à la source ou encore pour les prestations sociales quand elles sont soumises à conditions de ressources (allocation logement, pension de réversion…). Pour ces échanges, une convention régit les rapports entre le promoteur du service, le Groupement d'Intérêt Public de Modernisation des Déclarations Sociales (Gip-MDS), et les maîtrises d'œuvre qui y contribuent :

- dans le cas de la DSN d'une part la Cnav et d'autre part l'Urssaf[3] Caisse nationale ;
- dans le cas de Pasrau[4], la DGFiP[5].

▶ … une infrastructure technique…

Historiquement, les régimes et organismes de la sphère sociale ont instauré des liens informatiques spécifiques pour chaque échange. Les difficultés sont d'en assurer la robustesse, la sécurité et la traçabilité.

Des infrastructures ont été mises en place pour véhiculer, piloter et superviser chaque grand projet d'échange. Pour mettre fin à cette prolifération opportuniste, l'article R. 114-31 du code de la sécurité sociale crée le DGE[6] devenu un vecteur incontournable de communication et d'échange de données au sein de la sphère sociale.

Ce dispositif de gestion des échanges est un outil permettant les échanges entre Organismes de protection sociale (OPS), d'informations relatives aux assurés/bénéficiaires tout en garantissant la robustesse, la sécurité et la traçabilité. Il permet le routage de données concernant un unique assuré ou des millions d'assurés en flux de masse de la part d'un ou plusieurs émetteurs vers un ou plusieurs destinataires.

2 DSN : Déclaration Sociale Nominative.
3 Union de recouvrement des cotisations de sécurité sociale et d'allocations familiales.
4 Le dispositif PASRAU (Passage des revenus autres) résulte de travaux de simplification et de rationalisation des déclarations sociales. Il est le prolongement logique de la DSN (Déclaration Sociale Nominative) et a constitué ces dernières années une simplification majeure des procédures déclaratives concernant les salaires et les revenus versés par un employeur. Le dispositif PASRAU (fondé sur la norme NEORAU) complète donc la DSN (fondée sur la norme NEODeS) pour les « revenus de remplacement ».
5 Direction générale des Finances publiques.
6 Dispositif de Gestion des Échanges adossé au RNCPS, Répertoire National Commun de la Protection Sociale fournissant pour chaque assuré social l'identité des organismes lui servant des prestations.

Il répond ainsi à des besoins d'échanges journaliers entre organismes. De plus, il expose un catalogue d'échanges très utile dans une démarche « Dites-le nous une fois[7] ».

▶ … une normalisation du contenu : le concept de norme d'échange

> ***Définir la structure des données et les dispositifs de partage de la connaissance, garantir la qualité et la cohérence des données des différents flux.***

L'étape suivante consiste à définir la structure des données et les dispositifs de partage de la connaissance et à garantir la qualité et la cohérence des données des différents flux.

Pour y répondre, le principe de norme d'échange pour caractériser toutes les dimensions des flux de données est formalisé comme suit :

- une structure fonctionnelle et technique du message avec :
  - une arborescence de blocs de données ;
  - le typage de chaque donnée ;
  - la liste des contrôles entre rubriques ;
- une cinématique de l'échange, c'est-à-dire l'enchaînement des différentes étapes de l'échange.

Au-delà de la définition théorique, une norme se matérialise par des livrables :

- la documentation de la norme ;
- des services de contrôle automatique ;
- un outil d' « auto-contrôle » automatique pour l'émetteur ;
- une gestion des versions des différents livrables : ce point est essentiel dans un contexte où la norme bouge beaucoup, notamment, en raison de contraintes réglementaires.

Le besoin n'est pas récent. Différentes normes ont déjà été appliquées dans la sphère sociale :

- la Norme de Dématérialisation Des Déclarations de Données Sociales (N4DS) porte la Déclaration Automatisée des Données Sociales Unifiées (DADS-U) émise par les employeurs ou les concentrateurs de paie en direction du portail *www.net-entreprises.fr* géré par le Gip-MDS ;
- la norme A identifie les personnes physiques pour le bénéfice de ses partenaires par le Système National de Gestion des Identifiants (SNGI) *(Préveraud de Vaumas, 2022)*.

7 Le principe « Dites-le-nous une fois (DLNUF) », consiste à éviter aux usagers de fournir, lors de leurs démarches en ligne, des informations ou pièces justificatives déjà détenues par d'autres administrations, en s'appuyant sur le partage automatique de données entre organismes.

En dehors de la sphère sociale, d'autres normes d'échange de données existent. Par exemple, l'arrêté du 19 octobre 2018 approuvant le schéma national des données sur l'eau, les milieux aquatiques et les services publics d'eau et d'assainissement et l'article R. 131-34 du Code de l'environnement, ont mis en place le Sandre (Service National d'Administration des Données et Référentiels sur l'Eau).

De la structure fonctionnelle des messages : arborescence en blocs de données…

> ***La logique de structuration des données est le premier élément à prendre en compte dans la norme.***

La logique de structuration des données est le premier élément à prendre en compte dans la norme pour clarifier la présentation et formaliser la signification des objets utilisés et des liens entre eux. Elle s'appuie sur un format standard pour les données élémentaires.

Pour décrire ces notions, deux exemples :

- la DSN (norme NEODeS) et Pasrau (norme NEORAU), mises au point par le Gip-MDS avec la collaboration de la Cnav et des organismes de la sphère sociale ;
- la norme R pour les flux d'alimentation du Répertoire de gestion de carrières unique (RGCU) *(Sureau & Merlen, 2021)*, mise au point par la Cnav, avec les partenaires de la sphère sociale.

La description fonctionnelle est structurée avec des rubriques élémentaires (nom d'une personne, etc.) regroupées en blocs de données cohérents portant sur les mêmes concepts fonctionnels (bloc identité dans le cas présent avec le nom, le prénom, le numéro d'identification, etc.) et les liens entre les différents blocs (un « bloc » individu possède plusieurs « blocs » adresse par exemple). La représentation est cohérente avec le modèle conceptuel de l'ensemble des objets et de leurs relations, préalablement défini.

Cette structure est présentée sous la forme d'un arbre pour présenter les blocs et leurs rubriques ainsi que les dépendances entre les blocs. Les normes R, NEODeS ou NEORAu ont donc une structure hiérarchique.

Cette structuration est ensuite déclinée en modèles de message reprenant tout ou une partie des objets précédents pour répondre aux différents besoins d'échanges spécifiques (alimentations ou restitutions par exemple). Dans chaque modèle de message, un bloc se caractérise par une cardinalité : présence une fois et une seule, 1 à n fois, 0 à n fois ***(figure 1)***.

La norme comprend un en-tête avec les informations concernant l'émetteur et la nature et plusieurs messages contenant les données fonctionnelles.

Par exemple, pour la DSN, quelques blocs et les rubriques identifiantes dans chacun d'eux :

- Bloc « Entreprise » ;

  - SIREN (Système d'identification du répertoire des entreprises)

- Bloc « Établissement » ;

  - NIC (Numéro interne de classement : 5 derniers chiffres du code SIRET)

- Bloc « Individu » ;

  - Numéro d'inscription au répertoire (NIR)
    et/ou Numéro technique temporaire (NTT)

  - Nom de famille

  - Prénoms

  - Date de naissance

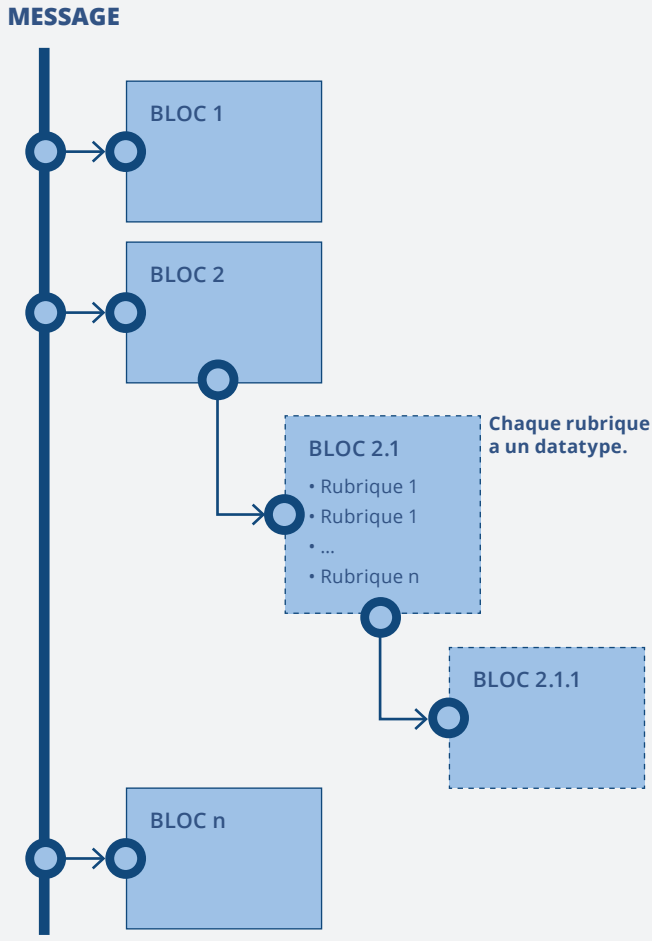
  - Bloc « Contrat (contrat de travail, convention, mandat) »

  - Numéro de contrat

  - Date de début de contrat

  - Identifiant du contrat d'engagement
    (pour rattacher les lignes de service au contrat de travail).

Figure 1 - Structure de la norme



▶ … à la structure technique des messages

La structure technique des normes d'échange évolue dans le temps ; les plus anciennes étaient généralement des lignes à champ de longueur fixe ou du texte avec séparateur[8] :

- Format « Texte longueur fixe »

  Le format "Texte longueur fixe" a pour caractéristique principale d'utiliser des champs de taille fixe. Dans chaque colonne de données, toutes les valeurs ont le même nombre de caractères. Comme il est impossible que des noms propres ou des montants aient tous la même longueur, des caractères de remplissage sont utilisés pour combler les "trous" ;

> ***Un changement de modèle des données du référentiel en lien avec la norme peut avoir un fort impact.***

- Format CSV

  Le format CSV est le plus universel. Il sépare les champs d'un enregistrement par une virgule ou un point-virgule et délimite les enregistrements par des *quotes*.

- Format XML

  Ces problèmes sont résolus avec les nouvelles syntaxes issues des protocoles développés avec le web. Le XML *(eXtensible Markup Language)* est le standard le plus communément utilisé. Il permet de créer ses propres balises (adaptation au contexte de l'information représentée) en séparant les données (contenu et structure) de la forme (présentation). Le XML apporte les notions d'extensibilité (nouvelles balises), de lisibilité (structuration de balises) et de portabilité (interopérabilité entre les environnements d'exploitation des fichiers XML).

▶ De la structure fonctionnelle des rubriques…

Une démarche identique à la précédente est ensuite appliquée pour les rubriques. On descend d'un cran en passant de la formalisation du message à celle des rubriques qui la composent, tout d'abord dans une démarche fonctionnelle qui décrit précisément chaque rubrique, puis dans une approche plus « technique » avec la liste des caractères autorisés. Une définition précise de chacune est fournie ainsi que son format et la plage de valeurs qu'elle peut prendre. Le format précise la typologie du champ (numérique, chaîne de caractères) et sa longueur.
Outre le format, la liste des valeurs que peut prendre la rubrique est précisée. Selon la nature de celle-ci, la typologie des valeurs est expliquée :

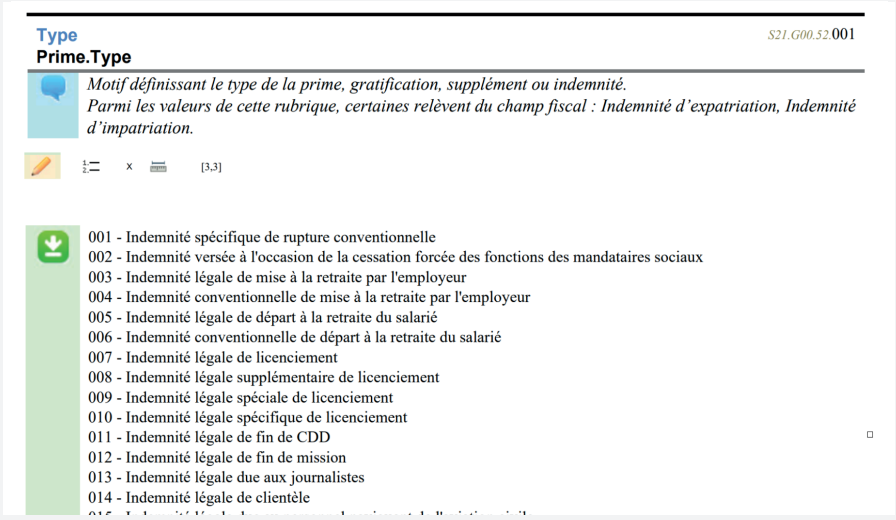
- intervalle de valeurs pour les rubriques numériques (la rubrique peut par exemple prendre une valeur de 0 à 100) ;
- gamme de valeurs basées sur une nomenclature (exemple : la rubrique doit prendre la valeur 1, 2 ou 3) ;

8 Voir l'article d'Alexis Dondon et Pierre Lamarche sur les formats de données dans ce même numéro.

- la rubrique respecte un format défini par une expression régulière[9], c'est-à-dire, avec une description formelle de la composition de la chaîne de caractères ; par exemple, le NIR – Numéro d'Inscription au Répertoire – a une structure du type SAAMMDDCCCOOO – Sexe, année de naissance, mois de naissance, département et commune de naissance, numéro d'ordre de naissance dans le lieu de naissance ;
- La rubrique appartient à une liste de valeurs elle-même gérée dans des référentiels externes ; par exemple, le répertoire SIRENE de l'Insee *(Alviset, 2020)*, ou la nomenclature des catégories socio-professionnelles *(Amossé, 2020)*.

De cette façon, dans la norme, est associé à chaque rubrique ce que l'on appelle un « type » ou « *datatype* », caractérisé de façon précise par la nature de la rubrique (numérique, alphanumérique…), la liste de valeurs possibles (s'il y a lieu), l'expression régulière associée (s'il y a lieu), et les longueurs minimum et maximum ***(figure 2)***.

Figure 2 - Exemple de rubrique (type de la prime) dans le cahier des charges de la norme NEODeS



… et structure technique : liste des caractères autorisés

Compte-tenu de la diversité des types de caractère, de la quantité et du volume des flux, il convient d'être très rigoureux sur la structure technique de chaque rubrique, jusqu'au format des caractères. La gamme des caractères autorisés est aussi normalisée. À savoir : on autorise de 0 à 9, de a à z (majuscule et minuscule) ainsi que quelques caractères spéciaux comme & ' ( ) … et on interdit les caractères spéciaux comme ! # $ % < > … ***(figure 3)***.

9 Décrire la notion d'expression régulière sort du cadre de cet article. Disons simplement, que la « grammaire » des expressions régulières permet de décrire des situations complexes, comme l'ensemble des possibilités pour une adresse mail, pour une date, … Voir *(Fourmond 2005)*.

Contrôle du contenu des messages : les contrôles syntaxiques induits par le modèle…

La description de la structure du message et des rubriques conduit mécaniquement à une série de contrôles à effectuer. Ils consistent à vérifier que le message respecte le « modèle de message » : par exemple, les blocs de données apparaissent dans l'ordre prévu, avec la bonne cardinalité (nombre minimum et maximum), les rubriques au sein de chaque bloc s'enchaînent dans le bon ordre, elles respectent le « type » prévu (par exemple, l'appartenance à une liste de valeurs prédéterminée ou la conformité à une expression régulière). Ainsi, on récupère les données attendues « dans les bonnes cases ».

Les contrôles fondés sur la structure sont appelés *contrôles syntaxiques*, puisqu'ils s'appuient sur la description d'une syntaxe, à la fois pour le message et pour chaque rubrique prise isolément.

… les contrôles sémantiques : cohérence interrubriques

> ***Il faut assurer la cohérence interne du message.***

Cependant, les contrôles syntaxiques ne suffisent pas. Il faut assurer la cohérence interne du message. Ainsi, il existe au sein de chaque norme[10] d'autres contrôles, qui matérialisent ce besoin de cohérence entre rubriques au sein d'un même message.

Figure 3 - Table des caractères autorisés ou non pour une rubrique

La table des caractères autorisés pour la valorisation des rubriques est un sous-ensemble de la table référencée ISO/IEC 8859-1. Les caractères interdits apparaissent sur fond grisé.

ISO/CEI 8859-1															
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	caractères de contrôle et divers non imprimables															
1x																
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	

10 Ou du moins chaque norme telle que ce concept est utilisé à la Cnav.

Il s'agit souvent de contrôles conditionnant la présence / absence d'une rubrique à la présence / absence d'une autre. Par exemple, dans le bloc Établissement de la DSN, il existe un contrôle exprimé en français : « Si le code postal est présent alors le code pays et le code de distribution à l'étranger sont absents et réciproquement. ». Ou, dans le cas de la norme R : « Si un élément de carrière est de type « Période assimilée maladie », on ne va pas trouver de revenu pour le régime général. ».

Certains contrôles inter-rubriques sont de simples inégalités : « La date de fin prévisionnelle de contrat doit être supérieure ou égale à la date de début du contrat. »

Il existe aussi des contrôles plus complexes sur le plan logique. Par exemple :

« Pour une déclaration donnée, le numéro de contrat doit être unique pour un établissement et un individu ».

Ces contrôles inter-rubriques, appelés contrôles sémantiques, sont écrits en langage naturel, afin qu'ils soient lisibles, et dans le même temps ils sont décrits en langage mathématique. Chaque contrôle sémantique est ainsi exprimé sous forme de formule logique, respectant une grammaire formelle prédéfinie.

La principale différence entre les contrôles syntaxiques et sémantiques ***(encadré)*** est la suivante : les seconds sont explicites (on écrit la règle de contrôle inter-rubriques), alors que les premiers sont implicites[11] (on décrit le type de la rubrique, ou la structure du message, et implicitement on dit que ce type ou cette structure doivent être respectés).

Quelques chiffres : en 2022, la Déclaration Sociale Nominative (DSN, norme NEODeS) comporte 7 modèles de messages, 61 blocs, 599 rubriques, 284 types. La norme R en 2023 comporte 26 modèles de messages, 163 blocs, 1174 rubriques, 282 types, 1163 contrôles sémantiques.

La mise en œuvre *via* une cinématique d'ensemble des contrôles

> ***L'ensemble de ces règles de contrôle, syntaxiques et sémantiques, caractérise la norme d'échange.***

L'ensemble de ces règles de contrôle, syntaxiques et sémantiques, caractérise la norme d'échange. Chaque règle décrit « ce qui est interdit ». Il faut souligner que ces contrôles sont intrinsèques au message : Il s'agit de vérifier si dans le message, la structure d'ensemble est correcte et si les données sont bien celles attendues (le bon type, le bon ordre…) et sont conformes à des règles de cohérence.

À ce stade, on n'a fait que décrire des règles. Il faut maintenant les mettre en œuvre, et donc vérifier le message dans toutes ses composantes. C'est nécessaire sur le plan métier car cela permet à toutes les parties prenantes, utilisatrices du message,

11 Il existe quelques rares cas particuliers où les contrôles syntaxiques sont décrits de façon explicite.

d'être assurées de la qualité du message, et en corollaire de la qualité d'alimentation des référentiels. Cette vérification est implémentée dans des interfaces applicatives (API[12]) qui exécutent des contrôles automatiques pour accepter ou rejeter les éléments contenus dans le flux.

Ces contrôles sont des contrôles bloquants. Le flux, dans le cas d'un traitement unitaire, fait l'objet d'un retour KO (le traitement s'arrête) et dans le cas d'un traitement de masse, il fait l'objet en retour d'un fichier de rejet (le traitement continue quant à lui pour les autres données du flux). Ces contrôles peuvent être complétés par des contrôles non bloquants qui permettent de continuer le traitement mais positionnent une alerte à l'utilisateur.

La succession des étapes, entre émetteur et récepteur du message, lorsque le message passe les contrôles, ou lorsqu'il ne les passe pas, est décrit avec précision dans la documentation de la norme.

▶ Encadré : un langage pour l'écriture des contrôles sémantiques

Le principe général est de décrire formellement une norme et de ne pas écrire de traitements : ceux-ci sont automatiquement générés. Ce principe s'applique aux contrôles sémantiques : on écrit simplement une expression logique (et non un traitement), qu'on associe à une rubrique.

Le langage utilisé n'est donc pas un pseudo-code. Il s'appuie sur une grammaire générale*, et requiert le vocabulaire nécessaire à l'écriture d'une formule logique dans le contexte d'une norme :

Noms de constantes : nombres, codes, noms de rubriques, nom de la rubrique courante ($rub),
Connecteurs logiques (or, and, not, =>, , ...)
Opérateurs de comparaison (=, !=, <, >=, ...), d'appartenance ensembliste (in), de présence (is_present)
Fonctions simples : opérateurs arithmétiques, sur les chaînes de caractères, sur les dates, ...
Quantificateurs universels et existentiels (every, some, ...), et noms de variables (x, y, z)

Exemple : on veut s'assurer que si la rubrique **aa** vaut 08 et que **bb** est présente, alors **cc** appartient à {1,2} . Ceci s'écrirait : `((aa='08') and isPresent(bb)) => cc in {'01','02'}`

Autre exemple : dans la mesure où un même bloc aa peut apparaître plusieurs fois (par exemple un bloc contrat), on peut retrouver plusieurs fois la même rubrique. Si on veut qu'elles soient toutes différentes de la rubrique courante, on écrira :

```
isPresent($rub) => (every x:aa satisfies ($x != $rub))
```

La génération automatique des traitements est une opération complexe, car il faut « aller chercher » les différentes rubriques dans le message, qui a une structure arborescente, avec de multiples blocs. Cette complexité est cachée à l'utilisateur, à qui il suffit d'écrire les formules logiques.

L'existence de ce langage s'est révélée très utile au-delà des contrôles, pour le développement des transformations : elles sont là aussi exprimées comme fonction mathématique, à partir de laquelle on génère les traitements de transformation de norme proprement dite.

***** *Techniquement, une grammaire compatible avec le framework ANTLR.*

12 Une API *(application programming interface* ou « interface de programmation d'application ») est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.

Les livrables associés à une norme

La norme représente le format d'échanges pour l'envoi et la réception de flux entre les différents acteurs et en assurer la standardisation et la qualité. Les normes d'échange concernent de nombreux acteurs et sont susceptibles d'évoluer fréquemment (changements réglementaires, nouvelles demandes…). Un ensemble de livrables cohérents entre eux et systématiquement associés à un numéro de version doit être impérativement transmis à tous les acteurs.

Ces livrables sont les suivants :

- les documentations décrivant la norme ;
- le fichier informatique de description de la structure utilisable pour les développements ;
- les interfaces applicatives (APIs) des contrôles ;
- la brique de contrôle destinée à la validation de gros volumes de flux (dans la DSN, une performance de 1 million de lignes en 20 minutes en moyenne) ; elle contient la « base de connaissance » avec le modèle et les logiques de validation spécifique à une norme et elle est configurable pour les besoins d'industrialisation de la production ;
- un outil d'auto-contrôle pour que l'émetteur du flux puisse faire les contrôles avant envoi pour tests au destinataire.

La documentation

> ***Le premier livrable, documentaire, est le cahier technique de la norme.***

Le premier livrable, documentaire, est le cahier technique de la norme. Il contient les objectifs, le périmètre, la description de la structure, ainsi que la définition de chaque bloc, chaque rubrique avec son format *(datatype)* et la gamme de valeurs associée. Le cahier technique est proposé au format pdf et au format html pour faciliter la mise à disposition ainsi que la recherche d'information à partir du poste de travail.

Ce document est riche et très détaillé : dans sa version 2022, le cahier technique de la norme NEODeS fait 354 pages[13]. Il commence par une centaine de pages d'explications générales sur la DSN, ses usages, et sur toutes les notions liées à une norme d'échange. Le document décrit ensuite les différents modèles de message, en l'occurrence les modèles de déclaration (DSN mensuelle, DSN signalement fin de contrat, DSN signalement arrêt de travail), et pour chaque modèle, l'arborescence de blocs correspondante. Il explicite ensuite tous les blocs, toutes les rubriques dans les blocs avec pour chaque rubrique son type et les contrôles associés ***(figure 4)***.

13 Voir *https://www.net-entreprises.fr/media/documentation/dsn-cahier-technique-2022.1.pdf*.

La documentation produite, au-delà du besoin de standardisation, comprend l'ensemble des informations et leur précision associée (blocs, rubriques, types de données, gammes de valeurs, contrôles appliqués…), tout en étant accessible grâce à une composition, un choix de maquette, des polices et des graphiques qui en facilitent la lecture.

La norme fait l'objet d'un affichage précis des versions successives ainsi que de la traçabilité des évolutions d'une version à l'autre, et elle est « à jour » et en cohérence avec les autres livrables techniques.

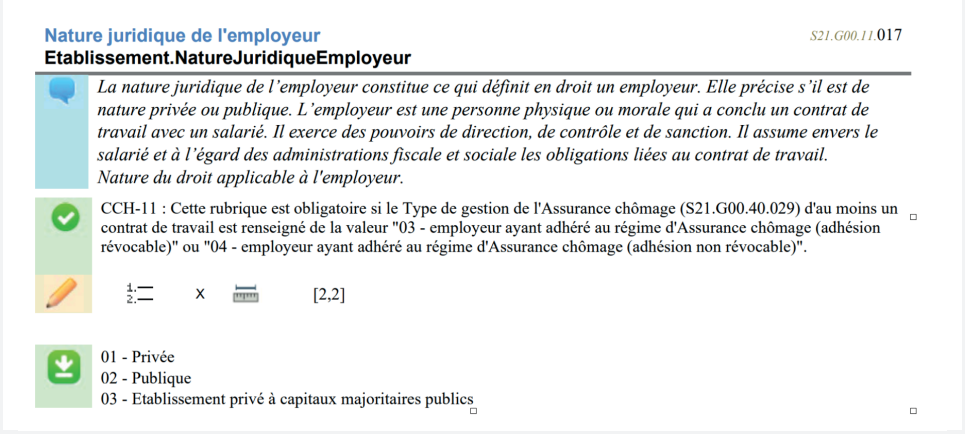
D'une version à l'autre, la liste de toutes les évolutions apportées est tracée dans un document spécifique livré avec chaque nouvelle version de norme.

Le cahier technique est complété par les **fichiers techniques de description du flux** (format xsd[14]) exposant la norme sous forme de schémas XML utilisables par les équipes de développement pour en dériver des objets.

Interface applicative pour les contrôles automatiques

L'API ou brique de contrôle est utilisée par les applications qui garantissent la qualité du flux (et, dans le cas de la norme R, la bonne alimentation du référentiel de carrières). Cette API ou brique de contrôle applique l'ensemble des vérifications et contrôles de la norme pour autoriser ou rejeter les enregistrements en fonction du résultat du contrôle. Il existe une API ou une brique par norme.

Figure 4 - Extrait du cahier technique de la norme Neodes pour la rubrique « Nature juridique de l'employeur »



Pour chaque rubrique, sont indiqués le nom, le libellé, l'explication détaillée, le ou les contrôles sémantiques (ici le « CCH-11 »), les caractéristiques de domaine (type « liste de valeurs), la longueur minimum et maximum et la liste de valeurs effective (01,02,03).

14 Xsd : xml schema definition : fichier de description du flux XML.

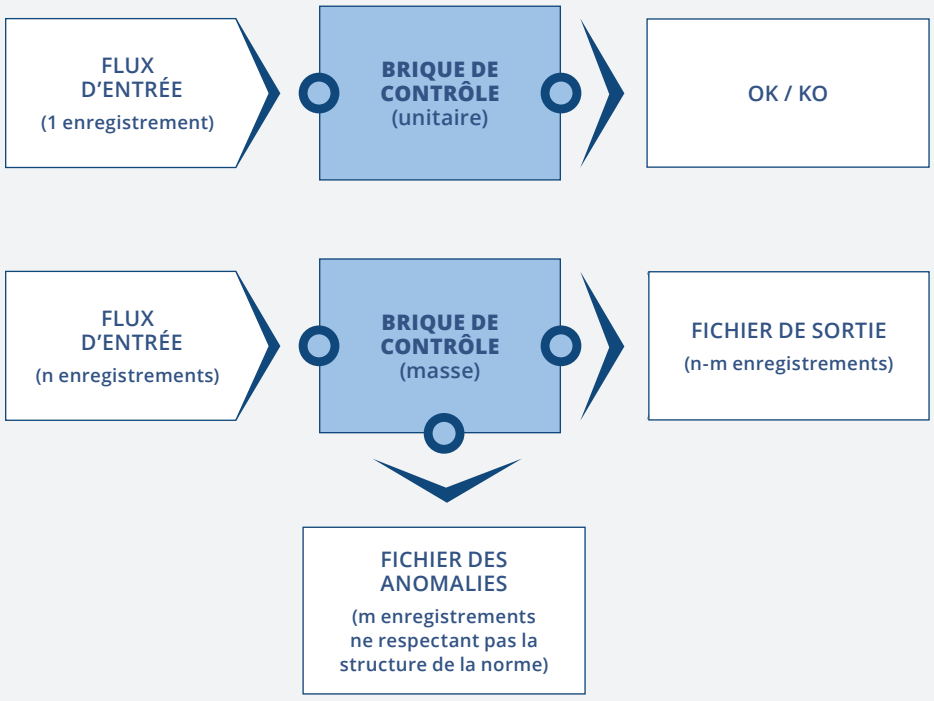
> ***Cette brique de contrôle applique l'ensemble des vérifications et contrôles de la norme pour autoriser ou rejeter les enregistrements en fonction du résultat du contrôle.***

L'API ou brique de contrôle traite deux types de flux, les flux unitaires et les flux de masse ***(figure 5)***.

Dans le cadre d'une saisie par un technicien sur une application ou par un assuré sur le web (cas du flux unitaire), si un seul contrôle s'avère négatif, le flux est rejeté : dans le cas du référentiel de carrières, cela implique que le flux n'est pas intégré dans le référentiel.

Dans le cadre d'échange en masse entre systèmes d'information, les enregistrements contrôlés « incorrects » sont rejetés. Les enregistrements qui satisfont les contrôles constituent un flux valide à intégrer. Un bilan de traitement du flux est envoyé à l'émetteur pour information et redressement des enregistrements incorrects.

Figure 5 - La brique de contrôle



Application d'autocontrôle

L'outil d'autocontrôle (ou application d'autocontrôle) est destiné aux émetteurs de données afin qu'ils vérifient la validité de leur flux avant de le transmettre à la plateforme de collecte, pendant les phases de test principalement.

Il s'agit d'une application qui comporte une interface graphique pour présenter le message à valider. Il génère le même bilan de traitement que la brique de contrôle implantée sur le serveur de collecte en production.

L'outil d'autocontrôle décrit la norme et identifie, dans le flux, toutes les données en anomalie. Il comporte aussi une aide en ligne et un mécanisme qui appelle le site de référence. En comparant sa version de la norme avec celle de référence, il propose sa mise à jour le cas échéant sans avoir à télécharger un nouvel outil complet.

Dans le cadre de la DSN, cet outil est très utile pour les entreprises, qui peuvent tester les déclarations sociales « chez elles » avant l'envoi dans les référentiels de la sphère sociale.

Un besoin d'outillage et d'automatisation

Les structures d'échanges peuvent comporter un grand nombre de blocs, de rubriques et de contrôles (pour la norme Neodes ou la norme R, plusieurs centaines de contrôles) et avec des évolutions fréquentes. Le risque d'incohérence entre les livrables est important et peut devenir très coûteux pour les opérations de vérification voire de remise en cohérence de la norme et des développements qui l'utilisent.

Il y a une dizaine d'années, à la Cnav (par exemple sur la N4DS et la DSN), le principe était d'écrire le cahier technique et de le faire évoluer régulièrement en fonction des nouveautés réglementaires ou des nouvelles demandes, puis de développer les contrôles automatiques et l'outil d'autocontrôle. Cela générait un coût élevé et le risque d'incohérence était permanent entre les différents livrables.

> ***Pour gagner en efficacité et cohérence, il fallait disposer d'une description formelle en un seul endroit pour y décliner tous les livrables.***

Pour gagner en efficacité et cohérence, il fallait disposer d'une description formelle en un seul endroit pour y décliner tous les livrables… sans besoin de développer les contrôles ou réécrire la documentation à chaque fois. C'était un gage de réactivité et de cohérence compte-tenu du nombre d' évolutions à gérer :

- plus de mille par an dans le cas de la N4DS à l'origine du projet ;
- plus de cinq cents par an par version de norme pour NEODeS, sachant que trois versions de NEODeS sont administrées de front : la norme en production de l'année N, celle de l'année suivante N+1 et celle de l'année d'après N+2.

Le principe appliqué est celui de la séparation des aspects métier et technique en construisant les outils les plus génériques possibles applicables à une norme d'échange. Le code relatif aux données et le code relatif aux traitements doivent être séparés ***(figure 6)***.

L'outil doit être basé sur la description d'un **modèle** et la déclinaison des livrables à partir de celui-ci.

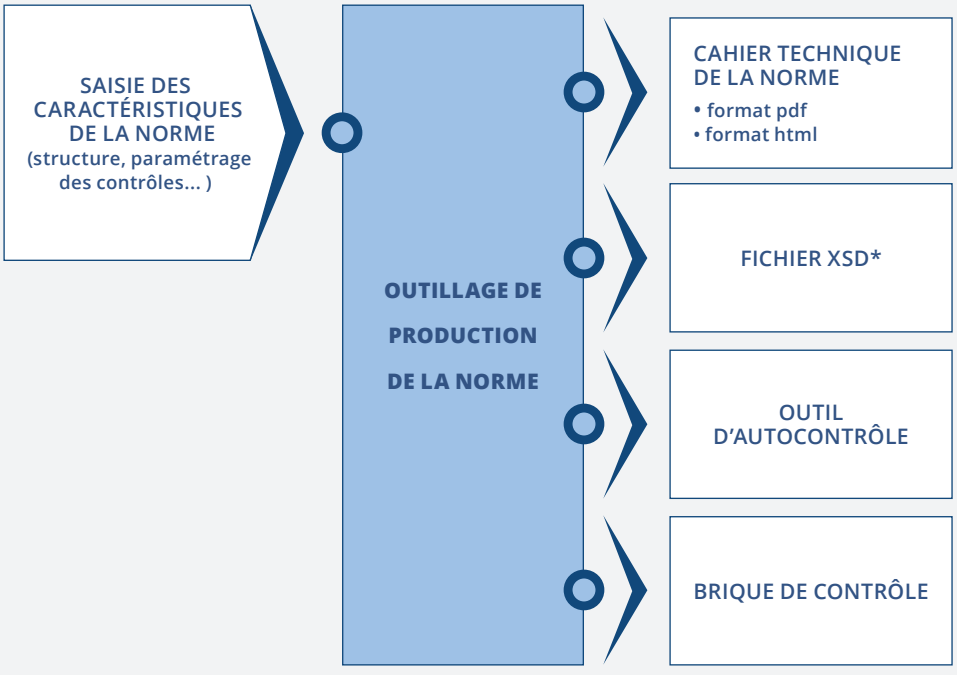
La déclinaison automatique d'un modèle unique permet de maintenir tout au long des années **un modèle intègre, valide et tracé**. Toute opération sur le modèle est attachée à une série de spécifications s'inscrivant dans un cycle de publication versionné. À son tour, le pilotage projet reprend ce versionnage pour construire les plans de déploiement.

Cet automatisme permet d'associer une équipe « maîtrise d'ouvrage » qui spécifie la norme et une équipe « maîtrise d'œuvre » qui réalise la chaîne de validation des flux. Ainsi, cette équipe mixte avec ces compétences couvre la modélisation de la norme et la génération de la base de connaissance associée qui sera lue par un moteur de validation unique.

▶ Saturne : un outil basé sur la modélisation

Cet outil Saturne pour « Suite Applicative pour des Traitements Unifiés et Rationalisés fondés sur une Norme d'Échange » *(Rivière P. et Rosec O., 2013)*, possède une interface, **l'Outil de gestion de la norme (OGN)**, pour décrire la norme. Il constitue alors un « référentiel de norme » fondé sur la modélisation. Cet outil permet de générer automatiquement la documentation, les formats XML et les outils de contrôle automatique ou d'autocontrôle.

▶ Figure 6 - La chaîne des livrables



** Les fichiers XML Schema Definition (XSD) permettent de décrire la structure d'un document XML. Le grand intérêt de ce fichier est de servir à la validation du document XML en définissant des règles.*

Techniquement, l'implémentation fondée sur un modèle a été réalisée avec un *Domain Specific Language* (DSL). Il s'agit d'un langage concis et adapté à un domaine métier *(Spinellis D., 2001 ; Fowler M., 2010)*. La mise en œuvre de ces principes a permis de faciliter l'implémentation, la généricité et la réactivité.

L'outil Saturne s'appuie sur un *framework*[15] de modélisation des normes. En première approche, la personne qui crée la norme n'a pas besoin de faire de développement. Elle modélise son projet par la saisie de la structure de la norme, des paramètres de contrôles dans Saturne, qui génère l'ensemble des livrables indiqués précédemment.

Extension et industrialisation

La richesse et la complexité d'une norme nécessitent une approche industrielle pour la développer et pour en garantir la reproductibilité. Un processus et un outillage sont créés en ce sens et intégrés à une chaîne continue comprenant la validation de son niveau de qualité.

Le modèle d'une norme comporte des milliers de propriétés. La qualification de la norme passe par la rédaction d'un cahier de tests très complet et la réalisation de milliers de tests. La charge de réalisation d'une norme dense peut être de l'ordre d'une quinzaine de jours mais la couverture de test peut, elle, prendre 2 à 4 fois plus de temps.

L'approche industrielle de Saturne est fondée sur une démarche et un outillage d'intégration continue, complétés, compte-tenu du nombre de cas de tests liés au nombre de blocs et rubriques de la norme, par un outillage spécifique pour tester la norme avec un **outil de qualification de norme (OQN)**. Il permet de dupliquer à partir de stéréotypes d'instances de norme valides les jeux de tests pour gagner en délai et en qualité de ceux-ci. Deux modèles auxiliaires représentent d'une part les jeux de tests et d'autre part le bilan d'une campagne de test.

Un outil complémentaire : l'outil de transformation de norme

> ***Un nouveau type d'outil a été créé : la transformation d'un flux de données d'une norme à une autre.***

Compte tenu des caractéristiques génériques de Saturne basées sur des modèles, un nouveau type d'outil a été créé : la transformation d'un flux de données d'une norme à une autre. La grammaire utilisée pour représenter les règles de contrôle dans la norme ***(encadré)*** est utilisée pour définir les règles de transformation d'une norme à une autre.

Selon le même principe, des programmes sont générés automatiquement à partir de la description mathématique des règles de transformation : c'est l'**outil de transformation de norme (OTN)**.

15 Un *framework* est un cadre de travail, sorte de boîte à outils qui facilite le travail de développement informatique.

Les OTNs[16], par l'utilisation de Saturne, bénéficient donc des mêmes avantages que la norme pour la cohérence des livrables, la standardisation de la documentation, la reproductibilité et la réactivité pour des demandes d'évolutions.

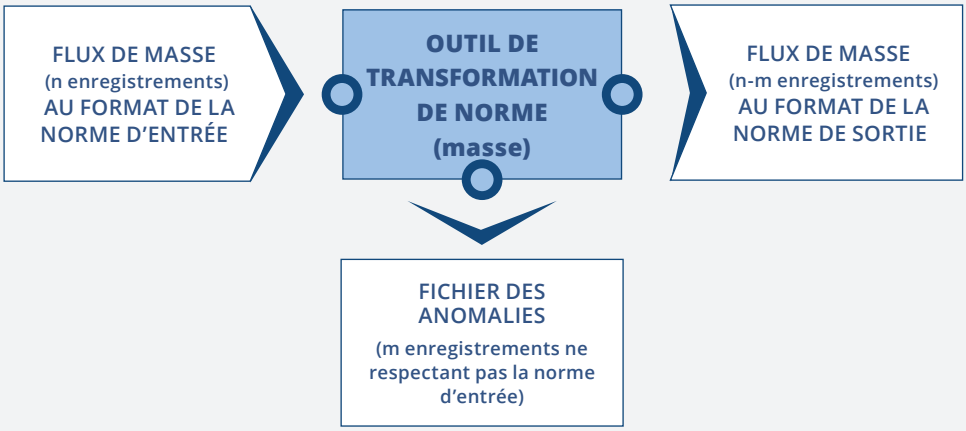
L'outil de transformation de norme (OTN) peut être conservé alors que le nouveau référentiel est généralisé, si des applications métier conservent les interfaces de l'ancien référentiel ***(figure 7)***.

La difficulté principale est de réunir les partenaires pour normaliser la transformation, écrire les règles de passage d'une norme à une autre, mettre en place la couverture de tests et le processus de qualification qui étend au processus de transformation la garantie de qualité que l'on veut faire peser sur le flux d'alimentation nativement à la norme d'entrée *(Miotto E., 2011)*.

Dans le cadre du remplacement du SNGC (Système National de Gestion des Carrières) par le RGCU, les OTNs sont des éléments déterminants. En effet le SNGC était alimenté par de nombreux organismes de la sphère sociale (Cnam[17], Pôle emploi, etc.) utilisant pour la plupart leur propre norme d'échange pour l'alimentation et la restitution des éléments de carrière. La bascule vers le RGCU de ces flux d'alimentation et de restitution a nécessité de les adapter à la norme R. Depuis, une vingtaine de normes ont été ainsi transformées vers la norme R.

De plus, dans le cas de l'alimentation du RGCU par la DSN, sachant que les deux normes sont partenariales, il faut écrire une spécification inter-régimes qui s'étend au fur et à mesure que ceux-ci basculent dans le RGCU. La spécification est publiée et éventuellement amendée au regard des remarques des représentants des régimes avant de devenir opposable pour la modélisation et le développement.

▶ Figure 7 - Les outils de transformation de norme



16 Par abus de langage, «les» OTNs sont les différentes transformations de norme réalisées à partir de l'OTN, outil générique.
17 Caisse nationale de l'Assurance Maladie.

Une mise en œuvre de plusieurs années sur plusieurs normes

À l'origine, la suite Saturne devait supporter la N4DS (Norme pour la dématérialisation des déclarations de données sociales) qui, en 2012, remplaçait la norme DADS-U. Mais à la suite d'un prototype et à la relance du projet DSN (déclaration sociale nominative, norme NEODeS), fin 2011, un rapport Igas – IGF a posé comme prérequis la conception d'un format d'échange plus simple et concis que la N4DS.

La déclaration sociale nominative (DSN, norme NEODeS) a été déployée par phases portant des scénarios métier de plus en plus ambitieux. En phase 1, elle transmettait mensuellement à un bloc applicatif de stockage les données qui servaient à générer à la place de l'employeur les attestations maladie ou chômage dont a besoin le salarié. En phase 2, elle collectait les cotisations précomptées par les employeurs et l'instrument de leur paiement. En phase 3, elle collecte les données carrière qui serviront à calculer les retraites des salariés.

Fin 2016, le prototype de la norme NEORAu a été utilisé par les employeurs hors DSN (les fonctions publiques jusqu'en 2022) puis par tous les collecteurs d'impôt sur le revenu dans le cadre du prélèvement à la source entré en production en 2019.

En septembre 2014, la conception de la norme R (comme retraite) a porté toutes les interactions entre le répertoire de gestion des carrières unique (RGCU) et les systèmes d'information (SI) métier des plus de trente régimes de retraite français qui peu à peu vont y basculer. L'ancien système national de gestion des carrières (SNGC) était alimenté par une multitude de flux hétérogènes. Dans le cas du RGCU toutes les interactions avec le répertoire sont unifiées dans la même norme, la norme R.

> ***La réactivité grâce au nommage des versions est vérifiée.***

Pour l'ensemble des projets pour laquelle la norme a été implémentée, le gain est indéniable. La compréhension et le partage des concepts liés à la qualité de la documentation sont facilités et la cohérence des livrables est toujours présente. La réactivité grâce au nommage des versions est vérifiée et est indispensable pendant la montée en charge des projets où le nombre d'évolutions reste très important. La mise en œuvre des contrôles automatiques permet d'assurer la qualité d'alimentation du référentiel dans le cadre d'arrivée régulière de nouveaux fournisseurs de flux. La mise à disposition d'outils d'autocontrôle autorise les émetteurs de flux à tester par anticipation l'adéquation de ceux-ci avec la structure de la norme et ainsi faciliter leur intégration lors des phases de tests, limitant ainsi les coûts et les délais.

D'un *Proof of Concept*[18] en 2012, Saturne est devenu un outil irremplaçable au service de deux programmes nationaux majeurs, la DSN et le RGCU.

18 Le PoC, *Proof of Concept*, ou preuve de concept en français, est une méthode qui permet d'évaluer la faisabilité d'un projet.

▶ Perspectives : étendre cette offre

La mise en œuvre de grands référentiels avec l'accroissement du volume, de la variété des données et des échanges associés génère un besoin de gouvernance de la donnée. La qualité des données est une condition indispensable à l'adoption de ces référentiels. La gouvernance de la donnée s'inscrit dans la stratégie *data* destinée à fournir des services à valeur ajoutée .

Cette stratégie est aujourd'hui portée par une direction dédiée à la Cnav, la direction de la gestion de la donnée (DGD), chargée de délivrer des services liés à la donnée grâce à une plateforme technologique, et en amont d'une démarche de mise en qualité de la donnée indispensable à la fourniture de services à haute valeur ajoutée.

> ***Cette mise en qualité est une orientation forte de la stratégie* data.**

Cette mise en qualité est une orientation forte de la stratégie *data*, que ce soit pour les SI internes de la Cnav ou pour les référentiels majeurs utilisés par toute la sphère sociale.

Les processus et les outils mis en œuvre autour de Saturne s'inscrivent pleinement dans cette orientation. Ils constituent aujourd'hui une suite industrielle qui tire parti des leviers de productivité offerts en matière de production logicielle par la génération de code par les modèles, la réutilisation de composants et la notion de ligne de produits.

C'est un gage de qualité pour l'intégration des flux dans les référentiels mais aussi un moyen de partage et de compréhension des données avec les différents partenaires. C'est aussi un gage de réactivité pour la prise en compte de modifications dans les phases de spécifications ou dans les phases de tests.

La documentation détaillée des données et l'apport de qualité sur les référentiels facilitent le travail de l'ensemble des acteurs utilisant les données, en particulier, les statisticiens pour qui la phase de prise de connaissance des données et de préparation est une composante lourde de leur processus.

La démarche aujourd'hui est de constituer une offre autour des principes de la norme et de la suite Saturne pour l'étendre à d'autres référentiels et bénéficier de ses atouts.

À la suite de ce retour d'expérience et dans le cadre de la démarche sur la stratégie *data*, la Cnav s'interroge pour étendre cette offre au-delà, à travers un cadre de mise en qualité général des données, dans un contexte d'échanges de plus en plus massifs entre les organismes.

Bibliographie

- ALVISET, Christophe, 2020. La troisième refonte du répertoire Sirene : trop ambitieuse ou pas assez. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee, N° N4, pp. 101-121. [Consulté le 6 juin 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497083?sommaire=4497095*.

- AMOSSÉ, Thomas, 2020. La nomenclature socioprofessionnelle 2020 : Continuité et innovation, pour des usages renforcés. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee. N° N4, pp. 62-80. [Consulté le 6 juin 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497076?sommaire=4497095*.

- FOURMOND Vincent, 2005. *Les expressions régulières par l'exemple*. H & K éditeurs. Collection Technique et pratique. ISBN 978-2-914010-65-8.

- FOWLER, Martin et PARSONS, Rebecca, 2010. *Domain-Specific Languages*. 23 septembre 2010. Addison-Wesley Signature Series. ISBN 978-0321712943.

- GRATIEUX, Laurent et LE GALL Olivier, 2016. *L'optimisation des échanges de données entre organismes de protection sociale.* IGAS, RAPPORT N°2015-090R1 / IGF N°2015-N-055.

- HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative. Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 25-34. [Consulté le 6 juin 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3647025?sommaire=3647035*.

- MIOTTO, Éric et VARDANEGA, Tullio, 2011. *On the integration of domain-specific and scientific bodies of knowledge.* In : *Model Driven Engineering.* [en ligne]. [Consulté le 6 juin 2023]. Disponible à l'adresse : *https://web.archive.org/web/20110724223732/http:/adams-project.org/standrts09/proceedings/miotto_vardanega_standrts09_final.pdf*.

- PRÉVERAUD DE VAUMAS, Joseph, 2022. Un référentiel des identités pour les besoins de la sphère sociale. Le système national de gestion des identifiants (SNGI). In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 93-114. [Consulté le 6 juin 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665190?sommaire=6665196*.

- RIVIÈRE, Pascal et ROSEC, Olivier, 2013. *Model-Based Interchange Formats : a Generic set of Tools for Validating Structured Data against a Knowledge Base. Poster Workshop of the Complex Systems Design & Management Conference CSD&M 2013*. [en ligne]. pp. 127-138. [Consulté le 6 juin 2023]. Disponible à l'adresse : *https://ceur-ws.org/Vol-1085/CSDM2013_PW.pdf*.

- SUREAU, Christian et MERLEN, Richard, 2021. Le Répertoire de gestion des carrières unique (RGCU). Un nouveau référentiel ouvrant des perspectives pour l'analyse sociale. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 64-81. [Consulté le 6 juin 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398687?sommaire=5398695*.

Présentation du numéro N9

Cette neuvième édition du Courrier des statistiques est caractérisée par plusieurs articles empreints de technicité, et par des sujets inhabituels pour la revue.
Tout commence par une histoire : celle de la statistique publique, prise ici sous l'angle du débat démocratique, dans les 40 années qui ont suivi la création de l'Insee.
Pour nourrir le débat public, l'Insee a récemment innové, avec la mise en place de « comptes nationaux distribués », qui permettent de mieux analyser la distribution de la croissance et son impact sur les revenus des ménages. Le second article en explique les principes, la mécanique et les perspectives.
Changement de thème avec deux articles sur la confidentialité des données. L'un donne le cadre législatif, les risques afférents à la rupture de confidentialité et les subtilités de l'application du secret statistique dans un contexte évolutif. L'autre, plus opérationnel, explique la logique du « code statistique non signifiant » (CSNS), et en quoi il facilite l'appariement de différentes sources tout en assurant la protection des données individuelles.
Les trois derniers papiers portent sur des sujets liés, importants dans un « monde de *data* ». On commence par les formats de données, sujet peu abordé mais que la statistique ne peut négliger. Bien choisir, bien gérer les formats est incontournable quand les statisticiens utilisent des sources de données externes. Avec l'article sur l'intégration des données administratives, on découvre un pipeline de traitement automatisé, piloté par les métadonnées, étape préalable avant une production statistique plus classique. Enfin, la Cnav explique l'importance de normes d'échange formalisées et documentées, générant automatiquement des outils de contrôle, pour mieux maîtriser la qualité des données au sein de la protection sociale.

ISSN 2107-0903
ISBN 978-2-11-162394-1

9 782111 623941

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE