Juin 2024

Courrier des statistiques

11

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE

Rédaction en chef
Catherine Fresson-Martinez

Contribution
Insee : Françoise Dupont, Josy Dussart, François Guillaumat-Tailliet, Heidi Koumarianos, Olivier Lefebvre, Lucas Malherbe, Jocelyne Mauguin, Nicolas Sagnes, Manuel Soulier, Thomas Tortosa
DGAFP : Gaël de Peretti
Ensai : Ronan Le Saout
France Stratégie : Nicolas Riedinger
SDES : Bérengère Mesqui
Université de Lille : Béatrice Touchelay

Directeur de la publication
Jean-Luc Tavernier

Directeur de la collection
Pascal Rivière

Rédaction
Catherine Fresson-Martinez, Solenn Ily, Marine Le Roux, Pascal Rivière

Composition
Agence Efil
90 boulevard Heurteloup
37 000 Tours
02 47 47 03 20
www.efil.fr

Photo de couverture
Getty Images

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

www.insee.fr

Courrier des statistiques N11

SOMMAIRE

PRÉSENTATION DU NUMÉRO

Dans ce nouveau numéro : trois suites, et un dossier.

Vous aviez lu avec curiosité « Statistiques publiques et débat démocratique : de la création à la consolidation (1946-1987) », dans le numéro N9, l'an dernier ? Vous attendiez la deuxième saison ? La voici, avec les mêmes auteurs. Il restait cependant une incertitude : l'année de fin de ce deuxième épisode. Allait-on couvrir une période allant de 1988 à aujourd'hui ? La réponse est non : l'article de Gaël de Peretti et Béatrice Touchelay se termine au début des années 2010, quand l'influence du numérique se met à changer la donne. À la fin des années 1980, tout est en place pour favoriser les échanges, s'interroger sur les usages des statistiques et faire apparaître de nouveaux thèmes. Le Cnis, enceinte qui permet et favorise ces discussions, succède au CNS. Mais le contexte change : l'Europe prend une place croissante et les statistiques européennes imposent plusieurs contraintes aux statistiques nationales. Les indicateurs clés relatifs au déficit et à la dette jouent un rôle structurant. De nouvelles problématiques sont abordées, par exemple celle des sans-abri, ou celle concernant la recherche d'indicateurs alternatifs au PIB. En parallèle, l'émergence de demandes d' « ouverture » prend différentes formes : ouverture des données individuelles aux chercheurs, visibilité accrue sur les conventions, mais aussi mise à disposition gratuite de l'ensemble des statistiques de l'Insee, réalisant la promesse de l'open data avant l'heure.

Vous aviez été surpris par l'article sur la datavisualisation, dans le numéro N10, par son histoire, ses méthodes, ses innovations et de multiples exemples ? L'article de Jocelyne Mauguin et Nicolas Sagnes s'inscrit dans la continuité, puisqu'il reste sur le thème des nouvelles méthodes de diffusion des statistiques publiques. Les auteurs partent d'un constat simple : l'Insee met à disposition une quantité absolument considérable de données, qu'elles soient statistiques ou individuelles. Face à cet océan de data, comment faire pour faciliter la vie des utilisateurs, pour qu'ils accèdent aisément et dans de bonnes conditions aux informations dont ils ont besoin ? La réponse à cette question… dépend des types d'utilisateurs. Pour des utilisateurs qui découvrent le site, ou qui en ont un usage plus épisodique, le catalogage, la simplicité de présentation et la datavisualisation seront prioritaires. Pour des utilisateurs chevronnés, voulant réaliser leurs propres statistiques, la question des formats et plus généralement des métadonnées sera essentielle, de même que la possibilité d'accès à des hypercubes faisant l'objet d'une forte standardisation. Dans le cas d'un usage plus industriel des données, avec notamment le moissonnage de données par des machines, la mise à disposition d'APIs est incontournable. Tout cela implique des évolutions du site, quant à son organisation, son alimentation et les standards qu'il utilise : c'est tout l'objet du projet Melodi (Mon espace de livraison des données en open data de l'Insee).

Vous êtes intéressés par les papiers présentant les statistiques relatives à un secteur économique, avec par exemple le logement, en 2020 (N4), ou bien le sport ou la défense en 2023 (N10) ? L'article de Ronan Le Saout, Nicolas Riedinger et Bérengère Mesqui s'attaque à un autre domaine spécifique, l'énergie. Celui-ci se prête naturellement à des mesures, à des quantifications, et ce depuis des siècles. On pourrait penser que cela facilite considérablement la tâche de la statistique publique… mais les auteurs nous expliquent que

ce n'est pas si simple. Il faut sans cesse (re)définir des conventions, sur la provenance de l'énergie, ou sa comptabilisation, et rien ne va de soi. Chaque source d'énergie a ses propres spécificités, ce qui vaut autant pour la production que pour le lien entre consommation et prix. L'appareil d'observation statistique doit aussi faire face à de nouveaux usages, dans le contexte de la transition écologique et climatique, avec par exemple l'amélioration de l'efficacité énergétique des logements… et à de nouvelles sources de données, avec par exemple les compteurs communicants.

Les quatre autres articles de ce numéro du Courrier constituent un dossier organisé autour d'un seul et même sujet : le projet de Répertoire statistique des individus et des logements (Résil), très structurant pour la statistique publique, car il fournit la colonne vertébrale nécessaire à l'obtention et l'appariement de sources diverses, qu'elles soient administratives ou d'enquêtes.

Dans le premier article du dossier, le maître d'ouvrage du programme, Olivier Lefebvre, présente le projet Résil. Cet article est d'intérêt en soi mais aussi en tant que « chapeau » pour les trois articles qui suivent. Le répertoire permettra, au même titre que Sirus dans le monde des entreprises (voir numéro N8), de construire des bases de sondage ou encore de vérifier la couverture des données administratives. Il s'agit plus généralement de produire de façon maîtrisée, efficace, réactive, des fichiers enrichis par appariement de sources diverses. Pour réaliser cela, Résil s'appuie sur plusieurs piliers, de natures très diverses : assurer la qualité du répertoire, se doter de traitements statistiques performants et innovants, bénéficier d'un fondement juridique clair et solide, et enfin disposer d'un mandat social à conforter en permanence, c'est-à-dire d'une légitimité, au-delà des dimensions technique ou juridique.

Le second article du dossier, écrit par Françoise Dupont, Josy Dussart et François Guillaumat-Tailliet est en lien avec ce dernier point. Pour assurer la légitimité de Résil, il a été indispensable de mettre en place tout un processus de concertation avec les parties prenantes. Il s'agit là, de façon générale, d'un sujet essentiel pour la statistique publique, et pourtant peu abordé dans la littérature habituelle de la profession. La concertation engagée par l'Insee dans le cadre du Cnis a permis de partager et d'ouvrir la réflexion sur ce que devait être Résil et sous quelles conditions il pouvait être utilisé, mais aussi comment l'expliquer au plus grand nombre. Elle a mobilisé des compétences diverses, pour la plupart éloignées du monde de la statistique, pour prendre en compte les considérations éthiques (protection des libertés publiques, transparence, etc.) et parvenir à une évaluation partagée des principes de nécessité, minimisation et proportionnalité, plus solide et pertinente qu'appréciée au départ. L'Insee a traduit les recommandations du groupe de concertation en dispositions juridiques, techniques (dans la conception et le contenu du répertoire), organisationnelles et de communication. La concertation, dont l'article donne une définition générale, ne se limite pas à ces dispositions, le travail de communication et d'écoute devant se poursuivre.

Les deux derniers papiers portent sur deux techniques tout à fait centrales dans le dispositif Résil : l'appariement de fichiers et l'accueil-réception-contrôle de sources administratives. Ils doivent être utilisés fréquemment, de façon quasi industrialisée, et s'adapter tous deux à des situations variées, ce qui requiert une certaine généricité.

La question de l'appariement a déjà été abordée à plusieurs reprises dans le Courrier, avec par exemple le système d'information sur l'insertion des jeunes (N6) ou les données de patrimoine (N7). Ici, Heidi Koumarianos, Lucas Malherbe et Olivier Lefebvre se penchent sur le sujet dans toute sa généralité, sans être spécifique à tel ou tel domaine. Le papier peut donc être lu aussi bien dans la perspective de Résil que comme une introduction à la problématique, à portée très large. Les auteurs constatent d'abord que les usages sont nombreux : enrichir des fichiers certes, mais aussi alléger des questionnaires d'enquête, compléter un champ d'analyse, vérifier la couverture d'une source, etc. L'article clarifie le vocabulaire (appariement, interconnexion, couplage), décrit le cadre juridique, puis présente la méthodologie, dans la situation où on ne dispose pas d'identifiant commun entre les deux fichiers à apparier. Il apparaît que même s'il existe un cadre théorique, on ne peut faire fi d'une bonne connaissance des sources pour paramétrer efficacement les algorithmes. Enfin, pour une bonne maîtrise statistique des fichiers appariés, les auteurs insistent sur l'importance de mesures de qualité des appariements.

Mais avant d'apparier, il faut au préalable avoir récupéré les fameuses sources administratives. On pourrait considérer qu'il s'agit d'une question purement technique, d'un pur chargement de fichier, avec simplement des questions de format à régler. Dans un numéro précédent (N9), à travers l'intégration des données administratives, force était de constater que le sujet était bien plus vaste. Ce que décrivent Thomas Tortosa, Manuel Soulier et Olivier Lefebvre, c'est un véritable service, permettant l'accueil des fichiers administratifs dans un univers statistique. En tant que service à utiliser de façon fréquente et dans des contextes variés, à travers Résil ou pas, il doit posséder un certain nombre de propriétés, en particulier adaptabilité, performance, traçabilité et sécurité. L'accueil des sources, que l'on fait en sorte de mutualiser, doit être clairement découplé des phases ultérieures, à savoir les traitements statistiques. Les auteurs replacent le sujet dans le cadre plus général du General Statistical Business Process Model (GSBPM) - modèle générique décrivant les différentes étapes à suivre pour produire des statistiques publiques - pour mettre en évidence les étapes élémentaires qui relèvent de l'accueil-réception-contrôle (ARC). L'outil ARC a été appliqué en premier lieu à la déclaration sociale nominative (DSN), et dans un deuxième temps on a procédé à sa généralisation, essentielle au projet Résil et porteuse d'enjeux pour d'autres processus de l'Insee.

Pascal Rivière
Directeur de la collection, Insee

IN MEMORIAM

Michel Volle

Michel Volle nous a quittés en juin 2024 à l'âge de 83 ans, et celles et ceux qui le connaissaient savent que c'est une grande perte. Il laisse une œuvre d'un intérêt majeur pour la statistique publique, et dont les principaux enseignements résonnent encore aujourd'hui. Il a réfléchi tant au métier de statisticien qu'à ses outils fondamentaux, sans occulter la question fondamentale de l'usage des statistiques et de leur adaptation aux opportunités et besoins du moment.

Les centres d'intérêt, les lectures de Michel Volle étaient multiples. Il avait la particularité d'avoir construit une expertise et écrit des ouvrages de référence dans les trois métiers de base de l'Insee : la statistique, l'économie et l'informatique. Mais dans son parcours éclectique, il s'est aussi intéressé à l'histoire, à la philosophie, aux télécoms, à la théologie, etc., et a publié sur tous ces sujets de manière régulière sur son blog (www.volle.com), de 1998 à 2018. La seule table des matières de ses billets de blog occupe une cinquantaine de pages ! Mais il aimait à rappeler que c'était bien l'apprentissage de la statistique, et de la production de statistiques, qui lui avait donné une rigueur, un cadre, une exigence, qui lui furent ensuite utiles toute sa vie.

On se restreint ici aux sujets qui concernent le Courrier des statistiques, ce qui écarte plusieurs aspects majeurs de ses travaux, notamment tout ce qui appartient au domaine des télécoms[1], et ne permet d'aborder qu'une petite partie du riche héritage intellectuel qu'il laisse.

S'il quitte l'Insee alors qu'il n'a qu'une quarantaine d'années, il n'en laisse pas moins une trace durable sur des générations de statisticiens. On pense bien entendu à son cours intitulé sobrement « Analyse des données », déjà précis, propre et limpide. On connaît moins « Le métier de statisticien »[2] alors qu'il s'agit d'une référence sur la manière de produire des statistiques, dont de nombreux éléments sont toujours d'actualité, à l'ère de la data science : l'importance des unités, codes ou nomenclatures ; les répertoires et questionnaires ; la collecte, la vérification, l'extrapolation ; l'utilisation de sources d'information d'origine administrative ; la fusion de fichiers ; la publication ; etc.

Il est le co-auteur, avec Bernard Guibert et Jean Laganier, de l'article qui définit le concept de nomenclature, article qui reste aujourd'hui la référence sur cette question[3]. Par la suite, il ne cessera dans ses écrits de souligner, au même titre qu'Alain Desrosières, l'importance de ce sujet, sa difficulté intrinsèque et son caractère structurant pour la statistique.

Aimant convoquer une dimension historique dans ses écrits, il rappellera toujours qu'il n'est pas historien de profession, et qu'il y a des personnes bien plus érudites que lui ;

1 Il fit partie de ceux, avec Pierre Musso notamment, qui au tout début des années 1990 alertèrent les autorités compétentes de l'importance à venir d'Internet.
2 « Le métier de statisticien » (Economica, 1980, 2e édition en 1984), disponible en texte intégral sur *www.volle.com*.
3 « Essai sur les nomenclatures industrielles », Guibert, Laganier, Volle, Économie et Statistique n°20, 1971. Cet article est naturellement cité par les deux derniers articles du Courrier sur les nomenclatures : la nomenclature socioprofessionnelle (N4), la nomenclature des infractions (N7).

mais cela ne l'empêchera pas de mettre en contexte l'activité statistique en rappelant les étapes depuis le XIVe siècle à Florence, ni de produire en 1982 une passionnante Histoire de la statistique industrielle, fondée sur des travaux qu'il a pu mener librement à l'Unité de recherche de l'Insee entre 1975 et 1978 et qui a donné lieu à sa thèse en Histoire. Cette riche chronologie, souvent citée comme référence, illustrée de force portraits, commence aux années 1930 et porte pour la plus grande partie sur la période 1950-1975, période des pionniers qui mettent en place la statistique d'entreprise.

S'il s'intéresse au passé, y compris sur longue période, en veillant à justifier son intérêt pour la question, il porte aussi son regard sur le futur. À titre d'exemple, il publie en 1989 un « Rapport général sur l'évolution à moyen terme de l'appareil statistique français »[4], qui s'ouvre sur cette question fondamentale : « À qui et à quoi sert la statistique ? ». Son constat initial selon lequel le « client » de la statistique ne se laisse pas aisément définir demeure toujours valable. Sur la base de cette question, il organise son rapport autour de quatre questions transversales :

- les nouvelles technologies et la statistique ;
- les articulations marchand-non marchand et public-privé ;
- le rôle de la statistique au sein de l'État ;
- le conditionnement théorique et social de la statistique.

Le discours est suffisamment général pour n'avoir pris que peu de rides : son questionnement sur l'usage des statistiques est toujours d'actualité, et de plus en plus avec le « déluge de données ».

Économiste, il prend conscience très tôt de l'importance des technologies dans l'économie, écrivant en 1999 « Économie des nouvelles technologies », interrogeant les fonctions de production à rendements croissants, puis publiant « E-conomie » en 2000. Il montre le caractère fondamental du phénomène d'informatisation, trop peu étudié par la littérature économique selon lui. Il approfondit sa réflexion sur les pratiques de prédation en économie dans « Prédation et prédateurs » en 2008, livre dans lequel il modélise les mécanismes essentiels de la prédation et du blanchiment. En 2014, il écrit un ouvrage de référence sur le concept d'iconomie[5], c'est-à-dire la société que fait émerger la « troisième révolution industrielle », celle de l'informatisation, et en co-produit le dictionnaire. Il est d'ailleurs intervenu sur ce thème lors d'une table ronde du Cnis tout récemment, en 2018.

Sa contribution majeure en informatique peut paraître plus étonnante : statisticien-économiste, l'informatique n'est pas nécessairement son univers de référence au départ. Le blog qu'il crée en 1998 consacre une large part de ses articles aux systèmes d'information, et représente depuis longtemps une mine pour qui étudie cette matière[6], tant il apporte de lisibilité et de recul théorique sur des questions souvent ardues, techniques. Ses papiers de blog sont percutants, le plus souvent simples à lire et efficaces, en général brefs, pour illustrer une idée (voir par exemple « L'expression des besoins et le système d'information », ou « Éloge du demi-désordre »), parfois plus approfondis (« Évolution du rôle du système d'information : du concept au processus »). Ils constituent autant de briques élémentaires

4 Disponible sur *www.volle.com*.
5 Concept introduit en 2006.
6 Il suffit de voir les commentaires élogieux sur son site...

qui vont lui permettre d'élaborer une véritable somme dont il existe peu ou pas d'équivalent : « De l'informatique » en 2006, toujours chez Economica. Il y aborde une très vaste gamme de sujets : caractéristiques physiques des ordinateurs, modélisation, qualité de service, développement, maîtrise d'ouvrage, *workflow*, histoire de l'informatique et de l'Internet, marketing interne, organisation, réseaux, stratégie d'entreprise, etc.

On l'aura compris, il n'y a jamais eu pour Michel Volle de sujets « nobles » à opposer à des sujets « moins nobles ». Sa démarche, au contraire, est une tentative permanente de comprendre des phénomènes complexes en les abordant dans toutes leurs dimensions, d'en donner à la fois une formalisation générale claire, souvent mathématisée, une solide mise en contexte historique, et des éclairages délibérément simples. Il était capable aussi d'une certaine vigueur dans quelques textes, ne rechignant pas à faire part de ses agacements ou colères.

C'est une personnalité éminente qui nous quitte, infatigable pédagogue, curieux des idées et des êtres, toujours ouvert aux débats, et très attaché au service public.

Pascal Rivière

Statistiques publiques et débat démocratique : de nouvelles attentes et de nouveaux enjeux (1988-2016)

Gaël de Peretti[*] *et Béatrice Touchelay*[**]

À la fin des années 1980, la statistique publique souhaite s'ouvrir à un plus large public. Cette nouvelle esquisse historique, qui se poursuit jusqu'au début des années 2010, décrit comment cette ouverture se traduit sur la production et la diffusion de l'information statistique. Dans un premier temps, les incidences de la construction européenne sur l'appareil statistique ainsi que ses conséquences sur le débat statistique seront abordées. Dans un deuxième temps, à travers quelques exemples, est traitée la façon dont la statistique publique réagit à la demande sociale ou politique de mesures et d'indicateurs statistiques. Quelles orientations sont mises en œuvre pour essayer de répondre aux besoins d'utilisateurs variés et toujours plus exigeants ? Finalement, si l'ouverture du service statistique public à de nombreux usagers semble évidente, la question de la capacité à élargir les discussions sur ce qui doit être compté et comment le mesurer reste posée. Enfin, l'ère numérique modifie les termes du débat, mais c'est une autre histoire…

At the end of the 1980s, Official Statistics wanted to open up to a wider audience. In this new historical sketch from the early 1990s to the 2010s, the aim is to describe how this openness translates into the production and dissemination of statistical information. First, the impact of European construction on the statistical system and its consequences on the statistical debate will be described. Second, we will illustrate through a few examples how public statistics react to the social or political demand for statistical measures and indicators. What directions are being implemented to meet the needs of varied and increasingly demanding users? In the end, if the opening of the public statistical service to many users is obvious, the question of the capacity to open the debate on what should be counted and how to measure it remains raised. Finally, the digital age changes the terms of debate. But that's another story…

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

[**] Professeure des Universités, Université de Lille, Institut de Recherches Historiques de Septentrion (IRHiS).
beatrice.touchelay@univ-lille.fr

Conçu pour être au service du public et de la démocratie économique et sociale conformément au programme du Conseil national de la Résistance de mars 1944, l'insuffisance des moyens budgétaires et humains accordés jusqu'au début des années 1960 ne permet pas de tenir ce programme ambitieux. L'Insee, créé en 1946, destine ses informations à un cercle d'experts et de décideurs. L'arrivée de nouvelles générations de décideurs politiques, l'augmentation du budget, des effectifs de la statistique publique qui essaiment dans les ministères, et le développement de l'informatique favorisent ensuite l'ouverture à un plus vaste public[1].

Avec la réforme des publications, la mise en place d'observatoires économiques régionaux (OER), d'enquêtes de terrain et de recensements démographiques quinquennaux, l'Insee intensifie les rendez-vous avec le public. Mieux connus et utilisés pour orienter les politiques publiques, ses indicateurs (par exemple l'indice des prix ou la mesure du chômage) sont parfois critiqués.

À la fin des années 1980, l'heure est aux interrogations. Le Conseil national de l'information statistique (Cnis)[2] organise un colloque pour réfléchir à l'information statistique des années 2000 (Cnis, 1989). Dans sa synthèse, Michel Volle soulève la question fondamentale (Volle, 1989) : à qui et à quoi sert cette information ? Cependant, les travaux de divers groupes de travail ne répondent pas directement à cette question. Ils étudient plutôt les conséquences des nouvelles technologies sur la production statistique, le rôle de la statistique au sein de l'État, le conditionnement social et théorique de la statistique et les axes de développement de sa qualité.

À cette même période, apparaissent des travaux autour de la « politique des nombres », qui interrogent les liens entre démocratie et nombre, aux États-Unis (Alonso et Starr, 1989), en France (Desrosières, 1987 ; Thévenot, 1981 ; Salais, 1986), au Royaume-Uni, etc. Comment le politique façonne-t-il les nombres ? Et réciproquement comment les nombres influencent-ils le champ politique ? « *How the domain of numbers is politically composed and the domain of politics is made up numerically* » (Rose, 1991, p. 675). Dans tous ces cas, l'interrogation porte sur l'articulation entre façon de penser la société, modalités d'action et modes de description (Desrosières, 2008).

Cette esquisse historique (***figure***) examine les liens entre statistique et débats démocratiques à travers trois focus spécifiques : premièrement l'Europe, deuxièmement ce que nous qualifierons de « statistiques dans l'arène », et troisièmement la volonté d'élargir le champ de diffusion de l'information statistique. Dans ces trois cas, la question principale est le changement entre les quarante premières années de l'Insee et les suivantes avec en particulier, les effets de la volonté de la statistique publique de s'ouvrir toujours plus vers l'extérieur.

1 Voir l'article du Courrier des statistiques N° N9 des mêmes auteurs : « Statistiques publiques et débat démocratique : de la création à la consolidation (1946-1987) ».
2 Le Conseil national de l'information statistique (Cnis) assure la concertation entre les producteurs et les utilisateurs de la statistique publique.

▶ Figure - Les grandes dates à retenir

** Cnis : Conseil national de l'information statistique*

▶ L'Europe ou les « statisticiens sur la brèche »

La fin des années 1980 et le début des années 1990 sont marqués par une intense activité de la Communauté européenne, qui n'est pas sans incidence sur la production statistique. En effet, « la liaison entre description et gestion apparaît nettement quand plusieurs États entreprennent comme c'est le cas aujourd'hui avec l'Europe [...], d'harmoniser leurs législations sociales, fiscales, économiques, afin de rendre possible la libre circulation des personnes, des marchandises et des capitaux » (Desrosières, 1993, p. 17). Cette activité communautaire propulse des statistiques sur le devant de la scène. Ainsi, le traité de Maastricht définit quatre indicateurs de convergence[3] à respecter pour intégrer l'Union économique et monétaire, et donc la zone euro, puis à conserver sous peine de sanctions. Les critères introduits par l'article 121 du traité établissant la Communauté européenne, correspondent à des seuils à ne pas dépasser :

- stabilité des prix : le taux d'inflation des États membres ne doit pas dépasser de plus de 1,5 point celui des trois États membres présentant les meilleurs résultats en matière de stabilité des prix ;
- situation des finances publiques :
  - interdiction d'avoir un déficit public annuel supérieur à 3 % du PIB de l'année précédente ;
  - interdiction d'avoir une dette publique supérieure à 60 % du PIB de l'année précédente ;
- taux de change : interdiction de dévaluer sa monnaie[4] ;
- taux d'intérêt à long terme : ils ne doivent pas excéder de plus de 2 % ceux des trois États membres présentant les meilleurs résultats en matière de stabilité des prix.

Une statistique nationale contrainte par les statistiques européennes

> ***Pour faciliter les comparaisons entre les États membres, la comptabilité nationale du Système européen des comptes (SEC) 95 est mise en place.***

Ces décisions ne sont pas sans incidence sur l'élaboration de statistiques. Pour faciliter les comparaisons entre les États membres, la comptabilité nationale du Système européen des comptes (SEC) 95 est mise en place[5]. Cela implique de disposer d'un système de comptabilité nationale commun, et aussi que tous les membres l'appliquent ; ainsi, le SEC devient un règlement européen le 25 juin 1996[6]. En effet, « Eurostat s'était aperçu que la majorité des pays membres n'utilisaient pas le SEC pour préparer les comptes. Le Système européen n'était utilisé que pour la transmission des données à Eurostat, ce qui introduisait de nombreuses distorsions dans l'interprétation et l'application des règles comptables et des définitions retenues. Ces écarts provoquaient d'importantes disparités

3 Critères de convergence (traité de Maastricht) : *https://www.insee.fr/fr/metadonnees/definition/c1348*.
4 Ceci fut rendu obsolète avec le passage à l'euro pour les pays de la zone euro. En outre, l'État membre doit avoir participé au mécanisme de taux de change du système monétaire européen (SME) sans discontinuer pendant les deux années précédant l'examen de sa situation, sans connaître de tensions graves.
5 Le SEC 95 s'inspire très fortement du Système des comptes nationaux 93 mis en place par l'ONU (Organisation des Nations unies), premier manuel de comptabilité nationale accepté par toutes les grandes organisations internationales.
6 Voir les références juridiques en fin d'article.

entre les résultats des pays membres. Ils ne pouvaient subsister alors que la comptabilité nationale devenait une référence incontournable dans l'application de nombreuses politiques communautaires : Union économique et monétaire, ressources propres, politique régionale, politique sociale, politique agricole, etc. » (Eurostat, 2003, p. 141). S'ajoutent les difficultés inhérentes à la prise en compte, ou à la non prise en compte, de certaines opérations pour réduire le déficit budgétaire. Concernant la France, après la privatisation de France Télécom (FT) en octobre 1996[7], il est nécessaire de savoir s'il faut ou non prendre en compte le versement pour la prise en charge des pensions de ses futurs retraités à l'État.

> ***Au milieu des années 1990 (1994-1997), chaque pays soucieux d'entrer dans la zone euro consacre beaucoup d'énergie pour calculer deux ratios, celui du déficit et celui de la dette publique.***

Ainsi, au milieu des années 1990 (1994-1997), chaque pays soucieux d'entrer dans la zone euro consacre beaucoup d'énergie pour calculer deux ratios, celui du déficit et celui de la dette publique, au détriment parfois d'autres recherches en matière de comptabilité nationale, comme les comptes par ménages.

Dans un même objectif d'harmonisation, Eurostat introduit un indice des prix à la consommation harmonisé (IPCH). Cette harmonisation n'est pas évidente puisque, parmi bien d'autres raisons (habitudes de consommations, conditions climatiques distinctes, etc.), les indices nationaux sont souvent liés à des procédures de revalorisation de contrat. En France par exemple, l'indice des prix hors tabac sert à indexer de nombreux contrats privés, comme les pensions alimentaires ou les rentes viagères et aussi le Smic[8]. Ces deux indicateurs, IPC et IPCH, diffèrent légèrement (Daubaire, 2022) : ils coexistent et comprendre les écarts entre ces deux indicateurs nécessite des explications très techniques.

Par ailleurs, le choix de ces critères très spécifiques orienterait les politiques publiques vers un contrôle des prix et du budget des États au détriment, par exemple, des questions de l'emploi, de l'environnement, des inégalités, etc. (Bournay, 2001).

> ***La création de la Banque centrale européenne (BCE) oriente profondément la production de statistiques.***

Au-delà des critères de Maastricht, la mise en place du système des banques centrales européennes[9] et la création de la Banque centrale européenne (BCE) guident profondément la production de statistiques. En complément du sujet de l'inflation, sur laquelle la BCE est très vigilante, cette institution invite à la production d'indicateurs conjoncturels toujours plus nombreux et dans des délais toujours plus contraints pour disposer quasi en temps réel d'un grand nombre d'indices, afin d'ajuster sa politique monétaire. « À partir du moment où l'on a désigné une Banque centrale européenne indépendante qui a les yeux braqués sur les marchés financiers et qui a donc besoin

7 Voir par exemple, « L'affaire France Télécom : une nuit à la Bundesbank » par Enrico Giovannini, in Eurostat (2003), « *Mémoires d'Eurostat ; Cinquante ans au service de l'Europe* », p.144.

8 Smic : Salaire minimum de croissance. Le Smic est le salaire minimum légal en France. Il se réfère à l'heure de travail. Il a été instauré par une loi du 2 janvier 1970. *https://www.insee.fr/fr/metadonnees/definition/c1006*.

9 *https://fr.wikipedia.org/wiki/Syst%C3%A8me_europ%C3%A9en_de_banques_centrales*.

d'avoir les mêmes indicateurs que les marchés pour dialoguer avec eux, on a vu arriver très tôt, avant même la constitution de la BCE, une demande très forte de production d'indices conjoncturels dans des délais extrêmement rapides. Les services statistiques ont été priés de répondre à la demande à toute allure. » (Durand, 2006).

Plus généralement, comme l'écrit Michel Glaude (2008), « Pour formuler, piloter et évaluer ces différentes politiques communautaires, il a été largement fait appel aux données statistiques et, plus particulièrement, à la constitution de « tableaux de bord » regroupant de nombreux indicateurs. En plus des indicateurs macroéconomiques traditionnels (produit intérieur brut (PIB), inflation, chômage, commerce extérieur, etc.), on a vu apparaître, d'une part, les principaux indicateurs économiques européens (PIEE) dans le domaine conjoncturel, des indicateurs structurels pour suivre la stratégie de Lisbonne[10], mais aussi d'autre part, des séries d'indicateurs relatifs à chaque domaine étudié (emploi, développement durable, inclusion sociale avec les « indicateurs de Laeken[11] », éducation, société de l'information, santé, innovation, etc.). »

Les statistiques au service de la démocratie ou de l'a-démocratie[12] ?

Par construction, cette demande européenne de statistiques contraint les instituts nationaux, mais sans apporter de moyens budgétaires ou humains complémentaires, et sature la capacité de production nationale de statistiques. Si dans un premier temps, la volonté d'harmonisation se traduit par les « outputs », à savoir les indicateurs à produire, cette harmonisation se fait ensuite par les « inputs », c'est-à-dire par des dispositifs d'enquête et/ou de remontées de données administratives totalement encadrés par des règlements. Par conséquent, le rôle du Cnis pour juger de l'opportunité de ces dispositifs est limité, puisqu'il ne peut s'opposer aux règlements européens. De même, la capacité du Comité du label[13] à juger de la pertinence du dispositif, de son protocole et de ses questionnaires est lui aussi restreint dès lors que ces sujets sont encadrés par ces mêmes règlements. Ainsi, la statistique publique est de plus en plus contrainte sur les indicateurs à produire. Et par ailleurs, les marges de manœuvre des utilisateurs et des producteurs pour faire évoluer les dispositifs dans des lieux dédiés aux concertations se réduisent.

Les limitations imposées à la production de statistiques nationales de par cette demande croissante et prioritaire[14] sont parfois interprétées comme une illustration supplémentaire de la « gouvernance par les nombres » (Supiot, 2015 ; Salais, 2007 ; Salais, 2022). Par exemple, à partir de l'évolution des catégories statistiques mises en avant pour étudier l'emploi et le chômage, Salais distingue deux approches. Dans une approche démocratique, le premier intérêt des statistiques est la construction d'un « savoir général » reposant sur

10 La stratégie de Lisbonne est l'axe majeur de politique économique et de développement de l'Union européenne entre 2000 et 2010, décidé au Conseil européen de Lisbonne de mars 2000 par les quinze États membres de l'Union européenne d'alors. L'objectif de cette stratégie est de faire de l'Union européenne « l'économie de la connaissance la plus compétitive et la plus dynamique du monde d'ici à 2010, capable d'une croissance économique durable accompagnée d'une amélioration quantitative et qualitative de l'emploi et d'une plus grande cohésion sociale ».

11 Le sommet de Laeken a conduit à l'adoption d'une liste d'indicateurs de « qualité de l'emploi » (Commission européenne, décembre 2001).

12 Référence à l'expression de Robert Salais (2022).

13 Voir l'article du Courrier des statistiques N5 : « Le Comité du label : un acteur de la gouvernance au service de la qualité des statistiques publiques », Marc Christine et Nicole Roth.

14 Dans le cadre de la révision de la « Loi statistique » (règlement du Conseil du 17 février 1997 relatif à la statistique communautaire), Eurostat avait mis en place le concept de « *First for Europe* » (voir les références juridiques en fin d'article).

> ***Là, au numérateur, on va mettre les actifs de 15 à 64 ans. Cela veut dire qu'il vaut mieux que les jeunes de 15 à 20 ans travaillent plutôt que d'aller à l'école. Même chose pour les gens de 60 à 64 ans.***

une pluralité et une variété de conventions statistiques[15], reconnues par tous pour comprendre le monde dans lequel ils vivent. Une fois le constat réalisé et partagé, il est possible de définir des politiques publiques pour améliorer ce qui doit l'être. Dans une approche qu'il qualifie de gouvernance par la quantification, la politique publique incorpore l'indicateur pour permettre de suivre sa propre mise en place ; cet indicateur est prédéfini et imposé par le « Centre »[16]. Il développe son argumentaire en expliquant comment la Commission européenne (exemple de « centre », voir Salais, 2022) est passée d'un objectif de plein emploi à celui de maximiser le taux d'emploi ; cet indicateur devient l'indicateur phare de la méthode ouverte de coordination[17]. Au-delà du fait que ce qui compte c'est d'être en emploi, et ce quel que soit le salaire, les conditions de travail, la durée ou le contrat, il y a implicitement d'autres éléments embarqués dans ce taux d'emploi des 15-64 ans. Comme le signalait Desrosières (2006) dans une table ronde sur « la statistique au service de la démocratie » : « Là, au numérateur, on va mettre les actifs de 15 à 64 ans. Cela veut dire qu'il vaut mieux que les jeunes de 15 à 20 ans travaillent plutôt que d'aller à l'école. Même chose pour les gens de 60 à 64 ans ». Dit autrement, si l'objectif se restreint au suivi de l'indicateur, cela impose implicitement des choix de politique sur l'emploi des jeunes et des seniors. Et selon Salais, cette approche peut être qualifiée d'a-démocratie, c'est-à-dire « un régime politique qui maintient les procédures formelles de la démocratie, mais entrave toute participation efficace des citoyens et des autres acteurs qui pourraient parler en leur nom » (Salais, 2022). Autrement dit, les façons de décrire la société permettent d'imposer tant la manière de la penser que les politiques à mettre en œuvre. Et surtout, les conventions statistiques sont imposées sans réelle discussion autre que technique. Ce constat sévère porté par Salais, pose la question des lieux de discussion pour échanger sur ce que doit produire la statistique publique.

En France, ce lieu est le Conseil national de l'information statistique (Cnis). Comme expliqué dans le précédent opus, la nécessité de s'ouvrir avait conduit à la création du Conseil national de la statistique (CNS) en 1972, puis à sa transformation en Cnis en 1984[18]. Dans un bilan sur quinze années de ces deux instances, son secrétaire général explique : « Il reste indispensable que le système statistique public joue la carte de l'ouverture vers les divers milieux économiques et sociaux et le reste de l'administration. Ceci n'est pas une exigence technique fondamentale, même si les avis techniques recueillis ne sont pas négligeables, mais à mon avis une exigence démocratique essentielle. L'administration statistique [...] doit s'efforcer à la transparence vis-à-vis de la société, au service de laquelle elle se trouve, ce qui signifie expliquer ce qu'elle fait, exposer ses projets à des interlocuteurs normalement, par hypothèse, moins compétents qu'elle et recueillir leur avis. » (Vanoli, 1989). En Europe,

15 Au sens développé, entre autres par Alain Desrosières, qui parle de convention d'équivalence, mais aussi de la phase de convention qui prévaut à la mesure avant de quantifier un phénomène.
16 Lieu où seraient décidés les indicateurs et les politiques publiques afférentes.
17 La « méthode ouverte de coordination » fait reposer la coordination sur des outils de comparaison entre les États membres.
18 Voir l'article du Courrier des statistiques N6 : « Le Conseil national de l'information statistique : la qualité des statistiques passe aussi par la concertation », Isabelle Anxionnaz et Françoise Maurel.

le comité consultatif européen de la statistique (ESAC[19]) a été mis en place en mars 2008 par la décision 234/2008/CE du Parlement européen et du Conseil[20]. Mais selon les propos de sa présidente en 2011 : « Nous nous considérons comme relativement petits par rapport à l'ampleur des tâches et mandats que nous avons. » (Lievesley, 2011).

▶ Les statistiques dans l'arène

La question mérite d'être posée : le Cnis joue-t-il son rôle ? Il y a toujours des marges de progrès et des besoins de s'adapter au contexte. Si le débat ne naît pas toujours en son sein, il finit parfois par s'y retrouver et conduire à faire évoluer la production statistique. Ainsi au cours de l'été 2004, des économistes publient une tribune dans le journal Le Monde (Concialdi et alii, 2004), intitulée « Cohésion sociale : des politiques à l'aveuglette ». Cette tribune formule un constat très critique sur l'appareil statistique et sa capacité à saisir la progression de la pauvreté et des inégalités[21]. Chose rare, elle suscite trois semaines plus tard une réaction du directeur général de l'Insee, Jean-Michel Charpin, et du président de l'Observatoire national de la pauvreté et de l'exclusion sociale (ONPES), Bertrand Fragonard, qui réfutent les arguments techniques avancés par ces chercheurs : « Qui est pauvre en France ? ». Enfin à la fin de l'été, Libération publie une tribune[22] intitulée « Mieux sonder la pauvreté », qui se réfère explicitement à cet échange entre le collectif de chercheurs et les deux instituts visés. Ils appellent à poursuivre le débat sur la connaissance des inégalités sociales « dans les instances qui y sont spécialement destinées, comme le Cnis : nous avons proposé que cet organisme constitue un groupe de travail en son sein à cet effet. Il doit, simultanément, se poursuivre dans le grand public : cela ne peut qu'être utile à la qualité de nos politiques sociale et économique, ainsi qu'à la capacité du système statistique public de répondre aux attentes des chercheurs et de la société ». Pression syndicale, discussion au sein du Cnis, réflexions à l'Insee : toutes ces étapes seront nécessaires pour la mise en place d'un groupe de travail présidé par Jacques Freyssinet[23]. Ses travaux se déroulent entre novembre 2005 et novembre 2006. Le projet de rapport de ce groupe de travail est débattu au bureau du Cnis en novembre 2006 puis présenté dans sa version finale à l'assemblée plénière de décembre 2006 (Cnis, 2007). Enfin, dans ses vœux du 2 janvier 2007, le directeur général de l'Insee écrit : « Nous avons montré que nous sommes à l'écoute et que nous savons nous remettre en question pour progresser : j'en veux pour exemple le groupe de travail du Cnis sur les niveaux de vie et les inégalités sociales, qui a engagé des débats riches sur un sujet qui préoccupe nos

> ***« Nous avons montré que nous sommes à l'écoute et que nous savons nous remettre en question pour progresser : j'en veux pour exemple le groupe de travail du Cnis sur les niveaux de vie et les inégalités sociales, qui a engagé des débats riches sur un sujet qui préoccupe nos concitoyens. »***

19 *European Statistical Advisory Committee*.
20 Voir les références juridiques en fin d'article.
21 Cette chronique s'appuie sur une communication de Bernard Sujobert dans le cadre du séminaire « Politique des statistiques » organisé à l'EHESS (École des hautes études en sciences sociales) par Isabelle Bruno, Alain Desrosières et Emmanuel Didier en 2012.
22 Rédigée par Nasser Mansouri-Guilani (directeur du centre confédéral d'études de la CGT (Confédération générale du travail)) et Denis Durand (représentant de la CGT au Cnis).
23 Jacques Freyssinet est un économiste français né en 1937 dont les travaux font autorité sur l'emploi et le chômage.

concitoyens. Nous cherchons à favoriser la transparence, y compris sur nous-mêmes : c'est pourquoi, nous publierons pour la première fois en 2006 un rapport d'activité externe[24]. » Les soixante recommandations de ce rapport vont profondément modifier la production statistique sur cette thématique. Ainsi, l'Insee mettra en place un sur-échantillon sur les hauts patrimoines dans l'enquête éponyme, explorera les distributions de revenus et de patrimoines du dernier centile, voire millime, mettra en place une première décontraction du compte des ménages selon les groupes sociaux ou des niveaux de revenus[25], enrichira les données localisées sur les revenus et la pauvreté. Cette réorientation se poursuit, ce qui permet au directeur général suivant Jean-Philippe Cotis de déclarer dans une interview au journal Le Monde (17 novembre 2009) : « La statistique est en train de sortir de la dictature de la moyenne. » Par ailleurs, parmi la cinquantaine d'indicateurs retenus pour étudier les niveaux de vie et les inégalités sociales, si les indicateurs de l'ONPES sont bien sélectionnés, les indicateurs de Laeken (Caussat et alii, 2006) ne le sont pas pour la plupart, comme si « le Cnis souhaitait reprendre son autonomie vis-à-vis de l'Europe » (Sujobert, 2012).

S'ouvrir aux marges de la statistique ?

Un autre exemple emblématique de l'action du Cnis est l'organisation du groupe de travail sur les sans-abri qui a conduit au rapport « Pour une meilleure connaissance des sans-abri et de l'exclusion du logement » (1996). Dans sa préface de l'ouvrage « La rue et le foyer » (2000), Jean-Marie Delarue insiste sur la nécessité de comprendre le fait social « exclusion du logement » : « Il ne suffit pas d'en rester au constat que des personnes n'ont plus de domicile : il faut pouvoir dire pourquoi, comment et combien. » Il insiste aussi sur le rôle du Cnis dans la prise en compte de cet aspect de la réalité sociale. Sept ans séparent la publication du rapport et la première enquête sur les sans-domicile en France réalisée en 2001 par l'Insee. En effet, les défis méthodologiques et organisationnels pour la mise en place de cette enquête ont nécessité plusieurs expérimentations et enquêtes préalables de l'Ined et l'Insee. Parmi les questions que se posent les chercheurs, statisticiens et autres personnes associées aux travaux, il y a celle de la légitimité de mener des enquêtes statistiques auprès des sans-domicile (Firdion et alii, 2000). Les objections à l'enquête statistique sont variées : atteintes à la vie privée, perturbations matérielle et psychologique, résistance à la démarche statistique, utilisation politique des nombres, etc. Faut-il pour autant renoncer à cette démarche statistique car, comme le disait un militant associatif : « Les chiffres, ça ne sert à rien, ce qu'il faut c'est loger les gens »[26] ou au contraire, comme le soulignait le père Wresinski, avoir un certain nombre de connaissances statistiques pour « fonder une politique réaliste et étayer la prise de conscience de la société » (Wresinski, 1987) ? Finalement, une triple légitimité apparaît : scientifique, démocratique, humaine ou humaniste (Firdion et alii, 2000). Scientifique, car il est nécessaire d'échapper aux stéréotypes et aux caricatures en s'attachant à décrire du mieux possible à la fois le continuum de situations de logement, mais aussi les processus qui conduisent les personnes aux marges du logement. Démocratique, car il est anormal d'exclure des citoyens de la « cité statistique » sous prétexte de difficultés méthodologiques pour les interroger. Humaniste, car « parler de soi, même dans un cadre structuré, permet

24 Ce rapport annuel, produit par l'Insee, présente les travaux phares de l'Institut.
25 Voir Bellamy et alii (2009) : il s'agit de mixer des données microéconomiques d'enquête et des données macroéconomiques du compte des ménages afin de décomposer les revenus et consommation de ce compte à partir des niveaux de vie ou des groupes sociaux.
26 Un responsable d'association de solidarité avec les sans-domicile (Paris, février 1995) (Firdion et alii).

d'avoir un regard sur soi, d'échapper quelque peu à la tyrannie du quotidien et de faire reculer le sentiment d'invisibilité sociale ».

Mesurer pour comprendre ?

> ***C'est pourquoi le débat est au sens le plus noble de nature politique. S'il n'appartient pas aux travaux statistiques de trancher des débats politiques et moraux, les données fournies doivent alimenter la réflexion.***

La confrontation entre la demande sociale de statistique et les producteurs peut se faire dans d'autres lieux. Deux exemples différents de la fin des années 2000 sont traités de façon succincte, alors que chacun mériterait de plus amples développements. Le premier est la constitution d'une mission d'information commune sur la mesure des grandes données économiques et sociales par les commissions des Finances, des Affaires sociales et des Affaires économiques fin 2007. Elle rend ses recommandations dans un rapport d'information publié en avril 2008 sous le nom : « Mesurer pour comprendre » (Mariton et Muet, 2008), qui devient la signature de l'Insee en 2013. Cette mission est mise en place à la suite d'une contestation des statistiques officielles considérée comme sans précédent, portant à la fois sur les chiffres du chômage, de l'inflation et du pouvoir d'achat. Elle vise à : « clarifier les termes du débat et proposer des mesures qui permettent de restaurer la confiance dans la statistique publique ». Comme énoncé dans l'introduction du rapport : « le débat ne porte pas tant sur les résultats de la mesure que sur la nature des données mesurées ». Il s'agit donc ici de définir ce qui compte, ce qui doit être compté et mis sur la place publique, car ce qui est sujet à controverse, ce sont les phénomènes économiques et sociaux que l'on veut mesurer. « C'est pourquoi le débat est au sens le plus noble de nature politique. S'il n'appartient pas aux travaux statistiques de trancher des débats politiques et moraux, les données fournies doivent alimenter la réflexion. » L'essentiel des propositions porteront évidemment sur la nécessité de compléter la production d'indicateurs sur le pouvoir d'achat, l'emploi, le halo du chômage et le sous-emploi, de développer les travaux et productions statistiques sur le développement durable, etc. Il s'agit aussi de garantir l'indépendance de la statistique publique en l'inscrivant dans le droit, en chargeant un organisme extérieur[27] de ce sujet et en étendant le Code de bonnes pratiques de la statistique européenne à l'ensemble des services statistiques ministériels.

Aller au-delà du PIB ?

Le deuxième exemple s'intéresse aux indicateurs alternatifs au PIB avec la création de la commission Stiglitz-Sen-Fitoussi et ses effets sur la statistique publique[28]. Lancée par le président de la République Nicolas Sarkozy début 2008, elle publie son rapport en septembre 2009. L'objectif est de réfléchir aux mesures alternatives de l'efficacité économique et du progrès social. Cette commission publiera deux ouvrages, le premier à destination des décideurs publics et des statisticiens (Vers de nouveaux systèmes de mesure, Stiglitz et alii,

27 Dans le rapport, Mariton et Muet évoquent le Cnis.
28 L'essentiel de cette section s'appuie sur la thèse de Félicien Pagnon : « Après la croissance : Controverses autour de la production et de l'usage des indicateurs alternatifs au PIB » (2022).

2009), et le deuxième plus scientifique et critique à l'égard de la notion de croissance et de l'usage du PIB comme indicateur (Richesse des nations et bien-être des individus, Stiglitz et alii, 2009). Un collectif, le Forum pour d'autres indicateurs de richesse (FAIR)[29], participe aux travaux par l'intermédiaire de Jean Gadrey, un des fondateurs du FAIR et expert sur ces sujets. Le FAIR juge assez sévèrement une partie du travail de la commission en particulier sur la faiblesse des recommandations sur le sujet soutenabilité, mais il est satisfait du diagnostic critique porté sur la prédominance du PIB et ses limites. À l'Insee, le rapport contribue à introduire de nouvelles statistiques, comme l'enquête sur le mal-logement en 2010, l'enquête sur les revenus distribués par quintile, des exploitations des enquêtes SILC[30] sur le capital social et le capital humain, des nouvelles questions sur le bien-être subjectif, la sécurité ressentie, etc. À l'OCDE[31], une nouvelle entité est mise en place : la *Better Life Initiative*[32].

▶ La statistique publique et ses usagers

Pour examiner les relations entre statistique et usagers, il est nécessaire de revenir à la question « à quoi cela sert et à qui ça sert ? » (Volle, 1989). Dans la présentation de l'Insee sur son blog, il est écrit : « L'Institut national de la statistique et des études économiques, l'Insee, collecte, produit, analyse et diffuse des informations sur l'économie et la société françaises. Ces informations intéressent les pouvoirs publics, les administrations, les partenaires sociaux, les entreprises, les chercheurs, les médias, les enseignants et les particuliers. Elles leur permettent d'enrichir leurs connaissances, d'effectuer des études, de faire des prévisions et de prendre des décisions. » Dans ce dernier temps, l'objectif est de regarder ce que l'Insee a mis ou pourrait mettre en place pour répondre à cet objectif de diffusion. Tous ces usagers ont des besoins particuliers.

Faut-il ouvrir la boîte noire des conventions ?

Depuis sa création en 1946, l'Insee essaie de s'ouvrir et de développer ses productions pour mieux répondre à la demande sociale. Ces efforts sont-ils suffisants pour répondre aux critiques des usagers ? En 1996, dans un colloque sur l'information économique et sociale, Alain Desrosières, dans un atelier sur « demande sociale et service public de l'information économique et sociale » identifie cinq critiques potentielles :

- on nous cache les informations les plus importantes ;
- ce que l'on a est biaisé, ne correspond pas à la réalité ;
- la statistique est réductrice, ce n'est pas « la vraie vie », la vie c'est autre chose que vos tableaux de chiffres ;
- la statistique exerce un contrôle social abusif ;
- les statistiques sont le produit d'un processus social, de conventions.

29 Chercheurs et associatifs qui travaillent depuis longtemps sur les indicateurs alternatifs au PIB.
30 *Statistics on Income and Living Conditions*, ou Statistiques sur les ressources et conditions de vie (SRCV, *https://www.insee.fr/fr/metadonnees/source/serie/s1220*).
31 OCDE : L'Organisation de coopération et de développement économiques est une organisation intergouvernementale d'études économiques (38 pays membres). *https://www.oecd.org/fr/*.
32 L'indicateur du vivre mieux est l'un des projets de l'initiative du vivre mieux de l'OCDE dont l'objectif est d'aider les gouvernements à placer le bien-être au centre de l'élaboration des politiques publiques.

> ***L'attitude conventionnaliste consiste à considérer que l'activité de base de la statistique est le codage. Le codage est comparable à ce que fait un juge : on prend un cas singulier et on le met dans une classe. Il y a un caractère arbitraire et conventionnel dans cette attribution.***

Derrière ces critiques, se cachent différentes perceptions de la réalité « statistique ». Les deux premiers cas sont qualifiés de réalisme métrologique, avec l'idée qu'il existe une vraie valeur, comme il existerait une vraie altitude du Mont Blanc. Les trois autres critiques sont plus radicales, car elles remettent en cause l'activité statistique.

Plus récemment, surgit l'inquiétude de la « capacité de la statistique à représenter le monde avec précision » (Davies, 2017). En cela, il s'agit plutôt des trois premières critiques. Apparaît une demande, difficile à satisfaire, de statistiques toujours plus précises, toujours plus fines pour s'approcher de la singularité de chacun.

Étudier les marges statistiques, sortir de la dictature de la moyenne, diffuser à des niveaux infra-nationaux voire locaux, proposer des indices des prix personnalisés sont des réponses, sans doute partielles, aux critiques adressées au service statistique public. En revanche, la question des conventions est finalement peu traitée : « L'attitude conventionnaliste consiste à considérer que l'activité de base de la statistique est le codage. Le codage est comparable à ce que fait un juge : on prend un cas singulier et on le met dans une classe. Il y a un caractère arbitraire et conventionnel dans cette attribution. » (Desrosières, 2008). Ces conventions ne sont pas sans effet sur la représentation de la société. Se pose alors la question de la capacité d'intervention de tout un chacun pour participer à l'élaboration de ces conventions. Certains évoquent le « statactivisme » (Bruno et alii, 2014). Une autre forme à explorer serait celle des forums hybrides (Callon et alii, 2001). En effet, se pose la question de la capacité à imposer ce qui compte, ce qui doit être compté et comment cela doit être compté (Latour, 1999). Sans forcément être responsable des choix, le statisticien va contribuer à réifier des catégories qui vont servir à décrire le monde. Ces catégories ou conventions pourraient être discutées dans des forums hybrides, « des espaces ouverts où des groupes peuvent se mobiliser pour débattre de choix techniques qui engagent le collectif ». Hybrides, car ces groupes engagés et leurs porte-paroles sont hétérogènes : experts, profanes, hommes politiques, etc.

Ouvrir les données aux chercheurs

Parmi les multiples usagers de la statistique publique, certains sont plus experts que d'autres et peuvent contribuer à cette phase de convention mais aussi et surtout d'analyse : les chercheurs. Ces derniers ont des besoins particuliers. Ils veulent accéder aux micro-données les plus détaillées possibles pour pouvoir réaliser leurs études en s'affranchissant parfois des nomenclatures usuelles. Pour pouvoir exploiter au mieux ces micro-données, il est nécessaire qu'elles soient documentées et que les métadonnées soient riches et de qualité, ce qui n'est pas sans coût pour les producteurs. De fait, l'Insee et le service statistique public réalisent des enquêtes ou produisent des fichiers à partir de sources administratives sans avoir la capacité d'exploiter pleinement ces ressources. Cette sous-exploitation est évidemment un problème compte tenu du coût de ces opérations. Parmi les critères de qualité mis en avant par l'OCDE, il y a celui de la rentabilité, au sens des informations produites à partir d'un dispositif. Par rapport à la période précédente (1946-1987), diffuser les sources (enquêtes ou

données administratives) pour permettre leur exploitation est une façon d'enrichir le débat social et un nouveau service rendu par la statistique publique. Les progrès en la matière sont importants depuis la fin des années 1980. Tout d'abord en 1986, une convention entre le CNRS[33] et l'Insee via le Laboratoire d'Analyse Secondaire et de Méthodes Appliquées à la Sociologie (Lasmas)[34] est signée. Avant cette convention, l'accès des chercheurs aux données de la statistique publique était parcellaire et plutôt lié à la connaissance de personnes entre elles (Silbermann, 1999). Cette convention, même si l'accès aux chercheurs du CNRS reste limité, est un premier pas. La situation perdure jusqu'au lancement de la mission « Sciences sociales et données » début 1999 par Claude Allègre, alors ministre de l'Éducation nationale, de la Recherche et de la Technologie, pilotée par Roxane Silberman, alors directrice du Lasmas-Iresco[35]. Cette mission identifie trois besoins : accroître la diffusion et l'utilisation des données et mieux associer les chercheurs à la production de données. Cette mission aboutira à la création du Centre Quetelet en 2001 (Chenu, 2003) dont les membres fondateurs sont le CIDSP (actuellement CDSP) qui fournit des enquêtes sociopolitiques, le Lasmas (actuellement Adisp) chargé notamment des données de la statistique publique, et l'Ined[36] : il deviendra en 2005 le Réseau Quetelet (Caporali et alii, 2015). Cette structure gère à la fois le sujet archivage, documentation et contrôle l'accès des chercheurs. Cela conduit à la création des fichiers de production pour la recherche (FPR), « fichiers raisonnablement anonymes, c'est-à-dire où il n'est pas possible d'identifier qui que ce soit, tant que l'on utilise ces fichiers à des fins de recherche scientifique » (Le Gléau, 2014). En ce qui concerne les entreprises, l'impossibilité d'ouvrir les données aux chercheurs incite à modifier la loi de 1951 en 1984[37], pour soumettre cet accès à l'accord du comité du secret statistique. Rapidement, il apparaît que pour des travaux de recherche plus précis, il est nécessaire d'accéder à des fichiers plus détaillés. Cela conduit à une nouvelle modification de la loi de 1951 en 2008 et à l'extension des missions du « comité du secret » des seules enquêtes entreprises aux enquêtes ménages. En parallèle, pour sécuriser l'accès à ces données détaillées, l'Insee avec le Genes (Groupe des écoles nationales d'économie et statistique) lance un projet de centre sécurisé en 2007 qui aboutit à la création en 2010 du Centre d'accès sécurisé aux données (Gadouche, 2019). Cependant, l'accès à ces fichiers est limité à des finalités de recherche.

L'ère numérique et la gratuité

Plus généralement, se pose la question de l'accès aux informations statistiques produites par le service statistique public. À la fin des années 1980, se pose la question de savoir s'il ne faudrait pas se concentrer sur la production statistique susceptible d'être vendue, ce qui permettrait d'identifier les besoins vers lesquels orienter la production (Volle, 1989). Dans les années 1990, l'Insee vend ses publications, des CD-ROM contenant des données plus ou moins détaillées, tout en respectant le secret statistique. Mais le développement d'internet, les possibilités offertes par ce nouveau vecteur pour diffuser l'information produite conduisent l'Insee, mi-2003, à modifier la politique de tarification et de rediffusion dans le sens d'une gratuité totale (Audibert, 2007), suivant ainsi l'exemple de nombreux instituts étrangers. Eurostat fera de même en 2004. Cette nouvelle politique se traduit par une forte

33 CNRS : Le Centre national de la recherche scientifique est le plus grand organisme public français de recherche scientifique.
34 Dirigé à l'époque par Alain Degenne : *https://fr.wikipedia.org/wiki/Alain_Degenne*.
35 Iresco : Institut de recherche sur les sociétés contemporaines.
36 CIDSP : Centre d'informatisation des données sociopolitiques ; CDSP : Centre de données sociopolitiques ; Adisp : Archives de données issues de la statistique publique ; Ined : Institut national d'études démographiques.
37 Voir les références juridiques en fin d'article.

augmentation à la fois de l'offre d'informations statistiques, mais aussi de la demande. Comme le souligne le directeur général de l'Insee en 2007, « le public s'est élargi et diversifié » (Charpin, 2007). Cette décision est très importante et fait de l'Insee un précurseur du développement quelques années plus tard de l'Open data. En effet, en 2011 est créée la mission Etalab qui met en place un portail unique interministériel de données publiques et en 2016, la loi pour une République numérique[38] consacre le principe de l'Open data par défaut[39].

Mais l'accès gratuit et à des informations toujours plus détaillées ne suffit pas. L'information diffusée doit être compréhensible, ce qui exige de documenter et d'accompagner l'usage de fichiers détaillés. Des efforts sont faits en ce sens en centralisant les nombreux sites de l'Insee en un seul, en étoffant les rubriques définitions, méthodes, et plus généralement les métadonnées[40] afférentes et en continuant à produire des analyses de premier niveau, voire plus sophistiquées pour ne pas laisser « l'inseenaute » seul face à la multitude des données. Cette ouverture génère des utilisateurs plus exigeants, ce qui impose de produire en pensant à la cohérence des données diffusées et à leur comparabilité dans le temps et l'espace. En matière d'accompagnement, la mise en place du service « Insee contact » au début des années 2000 permet de répondre aux questions des utilisateurs du site. En parallèle, cette stratégie s'accompagne de la disparition progressive des publications papier. De fait, le site internet insee.fr devient le principal vecteur de diffusion de l'information statistique.

En parallèle de cette augmentation de la diffusion d'informations statistiques via son site internet, l'Insee affiche dans son programme de moyen terme 2016-2025, nommé Insee 2025, la volonté de « faire parler les chiffres et d'aller au-devant de tous les publics ». Derrière cette orientation stratégique, l'objectif est, comme énoncé devant le Cnis en mars 2016, le suivant : « La statistique publique décrit et analyse une réalité de plus en plus complexe, utilise les vecteurs les plus modernes de diffusion, va au-devant de tous les publics dans un langage accessible à chacun et avec des produits adaptés. L'Insee, grâce au travail collectif de tous ceux qui concourent à ses productions, soumet ses chiffres à l'épreuve de la réalité et de la comparaison internationale pour en améliorer la pertinence, la qualité et la cohérence, s'assure de leur utilité pour éclairer les décisions nationales et locales, et veille à ce que la statistique publique couvre un champ cohérent et sans redondance. » Afin d'aller au-devant de tous les publics, l'Insee mobilise différents outils dont un blog, une application mobile, une chaîne YouTube pour élargir ses canaux de diffusion. Et enfin, l'Institut investit les réseaux sociaux comme Twitter (X) et LinkedIn et s'expose ainsi directement aux critiques de ses usagers. Expliquer, diffuser et communiquer sur des informations statistiques à travers ces nouveaux supports permet d'entrer dans une nouvelle arène. Admettre la critique, la recevoir, y répondre que ce soit pour dissiper des incompréhensions ou faire évoluer ses productions, devient une nécessité. Dans cette nouvelle ère de l'Open data, les données sont partout présentes sur internet, ce qui facilite la production de nouvelles statistiques publiques ou non, mais qui ouvre la possibilité à des concurrents moins soucieux de la qualité et la fiabilité de leur production de se développer. Dans un monde où de plus en plus de débats se déroulent sur les réseaux sociaux, avec les dérives que cela peut entraîner, avec la perte de confiance vis-à-vis de l'expertise, la statistique publique doit relever de nouveaux défis pour garder son rôle primordial dans le débat démocratique. Mais ceci est encore une autre histoire…

38 Voir les références juridiques en fin d'article.
39 Voir S. Goëta (2024), *Les données de la démocratie – Open data, pouvoirs et contre-pouvoirs*.
40 Voir l'article de Mauguin et Sagnes, « Faciliter l'accès aux données de l'Insee », dans ce même numéro.

▶ Fondements juridiques

- Règlement (CE) n° 2223/96 du Conseil du 25 juin 1996 relatif au système européen des comptes nationaux et régionaux dans la Communauté. In : *site de l'Union européenne*. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://op.europa.eu/fr/publication-detail/-/publication/c5b88d6f-3be6-4b5f-89d1-95d765f880b4/language-fr*.

- Règlement (CE) n° 322/97 du Conseil du 17 février 1997 relatif à la statistique communautaire. In : *site de l'Union européenne*. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://op.europa.eu/fr/publication-detail/-/publication/96e4c99e-6a09-4ba3-9e8d-a031b680975a/language-fr*.

- Décision n° 234/2008/CE du Parlement européen et du Conseil du 11 mars 2008 instituant le comité consultatif européen de la statistique et abrogeant la décision 91/116/CEE du Conseil. In : *Journal officiel de l'Union européenne*. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://op.europa.eu/fr/publication-detail/-/publication/a09da308-8c26-4dd1-b43b-6040c08ad2b1/language-fr/format-PDF/source-324382681*.

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

▶ Bibliographie

- ALONSO, William et STARR, Paul, 1989. *The Politics of Numbers - Population of the United States in the 1980s: A Census Monograph Series*. Russell Sage Foundation. ISBN 978-0871540157.

- ANXIONNAZ, Isabelle et MAUREL, Françoise, 2021. Le Conseil national de l'information statistique : la qualité des statistiques passe aussi par la concertation. In : *Courrier des statistiques*. [en ligne]. Juin 2021. Insee. N° N6, pp. 123-142. [Consulté le 20 juin 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398693?sommaire=5398695*.

- AUDIBERT, Pierre, 2007. L'expérience du système statistique public français. In : *L'accès à l'information statistique à l'heure d'internet. Les rencontres du Cnis*. 22 janvier 2007. [en ligne]. Rapport n° 104. Juin 2007. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2007_104_information_statistique_internet.pdf*.

- 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]. 1er novembre 2009. Insee. N° G2009/1.1 [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1380884*.

- BOURNAY, Jacques, 2001. Indicateurs statistiques et besoins sociaux. In : *Colloque : Statistique publique, évaluation et démocratie. Session II*. 21 mars 2001.

- BRUNO, Isabelle, DIDIER Emmanuel et PRÉVIEUX Julien, 2014. *Statactivisme : Comment lutter avec des nombres*. Éditions Zones. [en ligne]. 15 mai 2014 [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.editions-zones.fr/livres/statactivisme/*.

- CALLON Michel, LASCOUMES, Pierre et BARTHE, Yannick, 2001. *Agir dans un monde incertain. Essai sur la démocratie technique*. Paris, Le Seuil (collection « La couleur des idées »). [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://journals.openedition.org/developpementdurable/1316*.

- CAPORALI, Arianna, MORISSET, Amandine et LEGLEYE, Stéphane, 2015. La mise à disposition des enquêtes quantitatives en sciences sociales : l'exemple de l'Ined. In : *Population*. 2015/3. Vol. 70, pp. 567-597. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.cairn.info/revue-population-2015-3-page-567.htm*.

- CAUSSAT, Laurent, LELIÈVRE, Michèle, NAUZE-FICHET, Emmanuelle, 2006. Les travaux conduits au niveau européen sur les indicateurs sociaux de pauvreté. In : *Communication au 11e colloque de l'Association de Comptabilité Nationale*. 18-20 janvier 2006. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/2586304/texte_caussat_lelievre_nauze-fichet.pdf*.

- CHARPIN, Jean-Michel, 2007. Conclusion et perspectives. In : *L'accès à l'information statistique à l'heure d'internet. Les rencontres du Cnis*. 22 janvier 2007. [en ligne]. Rapport n° 104. Juin 2007. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2007_104_information_statistique_internet.pdf*.

- CHARPIN, Jean-Michel et FRAGONARD Bertrand, 2004. Qui est pauvre en France ?. In : *Le Monde*. [en ligne]. 21 juillet 2004. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.lemonde.fr/archives/article/2004/07/21/qui-est-pauvre-en-france-par-jean-michel-charpin-et-bertrand-fragonard_373218_1819218.html*.

- CHENU, Alain, 2003. Une infrastructure pour les données en sciences humaines et sociales. In : *Courrier des statistiques*. [en ligne]. Septembre 2003. Insee. N° 107, pp. 29-31. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt1zt/f1.pdf*.

- CHRISTINE, Marc et ROTH, Nicole, 2020. Le Comité du label : un acteur de la gouvernance au service de la qualité des statistiques publiques. In : *Courrier des statistiques*. [en ligne]. Décembre 2020. Insee. N° N5, pp. 39-52. [Consulté le 20 juin 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5008698?sommaire=5008710*.

- CNIS, 1989. L'information statistique en l'an 2000 : continuité et changement. In : *Actes colloque 19-20 avril 1989*.

- CNIS, 1996. Pour une meilleure connaissance des sans-abri et de l'exclusion du logement. In : *Rapport final du groupe de travail sur les sans-abri*. N° 29. Mars 1996.

- CNIS, 2007. Niveaux de vie et inégalités sociales. In : *Rapport final du groupe de travail sur les inégalités sociales*. [en ligne]. Mars 2007. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/11/RAP_2007_103_niveaux_de_vie_inegalites_sociales.pdf*.

- CONCIALDI, Pierre, GADREY, Jean, LÉVY, Catherine et MARIC, Michel, 2004. Cohésion sociale : des politiques à l'aveuglette, par des économistes et une sociologue. In : *Le Monde*. [en ligne]. 1er juillet 2004. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.lemonde.fr/archives/article/2004/07/01/cohesion-sociale-des-politiques-a-l-aveuglette-par-des-economistes-et-une-sociologue_371194_1819218.html*.

- COTIS, Jean-Philippe, 2009. La statistique est en train de sortir de la dictature de la moyenne. In : *Le Monde*. [en ligne]. 17 novembre 2009. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.lemonde.fr/la-crise-financiere/article/2009/11/17/jean-philippe-cotis-la-statistique-est-en-train-de-sortir-de-la-dictature-de-la-moyenne_1268251_1101386.html*.

- DAUBAIRE, Aurélien, 2022. Indice des prix à la consommation vs indice des prix harmonisé au niveau européen : santé et énergie font la différence. In : *Blog Insee*. [en ligne]. 1er mars 2022. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://blog.insee.fr/ipc-vs-ipc-harmonise-sante-et-energie-comptent/*.

- DAVIES, William, 2017. Comment la statistique a perdu son pouvoir - et pourquoi nous devrions craindre ce qui va suivre. In : *Statistique et société*. [en ligne]. Avril 2017. SFdS. vol. 5, n° 1. pp. 11-20. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://statistique-et-societe.fr/index.php/stat_soc/article/view/608*.

- de PERETTI, Gaël et TOUCHELAY, Béatrice, 2023. Statistiques publiques et débat démocratique : de la création à la consolidation (1946 - 1987). In : *Courrier des statistiques*. [en ligne]. Juin 2023. Insee. N° N9, pp. 7-23. [Consulté le 20 juin 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635819?sommaire=7635842*.

- DELARUE, Jean-Marie, 2000. Préface. In : *La rue et le foyer - Une recherche sur les sans-domicile et les mal-logés dans les années 1990*. PUF Diffusion – Ined. ISBN 978-2-7332-0144-2.

- DESROSIÈRES, Alain, 1987. Les nomenclatures de professions et d'emploi. In : *Pour une histoire de la statistique : Tome 2 matériaux*. Éditions Joëlle Affichard, INSEE Economica. pp. 35-36. ISBN 978-2-7178-1261-9.

- DESROSIÈRES, Alain, 1993. *La politique des grands nombres. Histoire de la raison statistique*. La Découverte. ISBN 978-2707165046.

- DESROSIÈRES, Alain, 1996. Analyse des besoins ou analyse des usages. In : *Colloque : L'information économique et sociale aujourd'hui. Besoins, représentations, usages. Atelier 3 - La demande sociale et le service public de l'information économique et sociale*.

- DESROSIÈRES, Alain, 2006. Table ronde : La statistique au service de la démocratie. In : *Colloque : La statistique au service de la démocratie*.

- DESROSIÈRES, Alain, 2008. *Gouverner par les nombres. Argument statistique II*. Presse des Mines. ISBN 978-2-35671-005-5.

- DURAND, Denis, 2006. Table ronde : La statistique au service de la démocratie. In : *Colloque : La statistique au service de la démocratie*.

- EUROSTAT, 2003. *Mémoires d'Eurostat - Cinquante ans au service de l'Europe*. [en ligne]. 15 mai 2003. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/fr/web/products-statistical-books/-/ks-49-02-183*.

- FIRDION, Jean-Marie, MARPSAT, Maryse et BOZON, Michel, 2000. Est-il légitime de mener des enquêtes statistiques auprès des sans-domicile ? - Une question éthique et scientifique. In : *La rue et le foyer - Une recherche sur les sans-domicile et les mal-logés dans les années 1990*. PUF Diffusion – Ined. pp. 127-150. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.researchgate.net/publication/285051906_Est-il_legitime_de_mener_des_enquetes_statistiques_aupres_des_sans-domicile_Une_question_ethique_et_scientifique*.

- 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]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254227?sommaire=4254170*.

- GLAUDE, Michel, 2008. Les instituts nationaux, Eurostat et les organismes internationaux de statistiques : vers une indépendance renforcée. In : *Regards sur l'actualité. Les statistiques publiques en débat*. La Documentation Française. N° 346, décembre 2008.

- LATOUR, Bruno, 1999. *Politiques de la nature. Comment faire entrer les sciences en démocratie*. Éditions La Découverte. Coll. Armillaire. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://journals.openedition.org/questionsdecommunication/6525*.

- LE GLÉAU, Jean-Pierre, 2014. L'accès aux données confidentielles de la statistique publique -De la sensibilité des données économiques à la sensibilité des données de santé. In : *Statistique et société*. [en ligne]. 2 juin 2014. SFdS. Vol. 2, N° 2. pp. 27-32. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://statistique-et-societe.fr/index.php/stat_soc/issue/view/36*.

- LIEVESLEY, Denise, 2011. L'utilité des statistiques du point de vue des utilisateurs, Plénière introductive - La statistique publique, une cible mouvante. In : Colloque : *La statistique publique, un bien public original*.

- MANSOURI-GUILANI, Nasser et DURAND, Denis, 2004. Mieux sonder la pauvreté. In : *Libération*. [en ligne]. 26 août 2004. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.agirensemblecontrelechomage.org/IMG/pdf/BIP40-3.pdf*.

- MARITON, Hervé et MUET, Pierre-Alain, 2008. *Rapport d'information n° 815. Mission d'information commune sur la mesure des grandes données économiques et sociales*. Assemblée nationale, XIII[e] législature. [en ligne]. 16 avril 2008. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.assemblee-nationale.fr/13/rap-info/i0815.asp#P30_354*.

- PAGNON, Félicien, 2022. *Après la croissance : Controverses autour de la production et de l'usage des indicateurs alternatifs au PIB*. Thèse de sociologie. Université Paris sciences et lettres. [en ligne]. Soutenue le 30 novembre 2022. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://theses.fr/2022UPSLD042*.

- ROSE, Nikolas, 1991. Governing by numbers: Figuring out democracy. In : *Accounting, Organizations and Society*. [en ligne]. Vol. 16, N° 7, pp. 673-692. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.sciencedirect.com/science/article/abs/pii/036136829190019B*.

- SALAIS, Robert, 1986. L'émergence de la catégorie moderne de chômeur : les années 1930. In : *L'invention du chômage. Histoire et transformations d'une catégorie en France des années 1890 aux années 1980*. R. Salais, N. Baverez, B. Reynaud. PUF. pp. 77-123.

- SALAIS, Robert, 2007. *Europe and the Deconstruction of the Category of 'Unemployment'*. Archiv für Sozialgeschichte, 47. pp. 371-401. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://library.fes.de/pdf-files/afs/bd47/16_salais.pdf*.

- SALAIS, Robert, 2022. « La donnée n'est pas un donné » : Statistics, Quantification and Democratic Choice. In : *The New Politics of Numbers – Utopia, Evidence and Democracy*. Edited by Andrea Mennicken and Robert Salais. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://link.springer.com/chapter/10.1007/978-3-030-78201-6_12*.

- SILBERMAN, Roxane, 1999. *Les sciences sociales et leurs données*. Ministère de l'Éducation nationale, de la Recherche et de la Technologie. [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.education.gouv.fr/les-sciences-sociales-et-leurs-donnees-12923*.

- STIGLITZ, Joseph E., SEN, Amartya et FITOUSSI, Jean-Paul, 2009. *Vers de nouveaux systèmes de mesure - Performances économiques et progrès social*. Odile Jacob. ISBN 978-2738124630.

- STIGLITZ, Joseph E., SEN, Amartya et FITOUSSI, Jean-Paul, 2009. *Richesse des nations et bien-être des individus*. Odile Jacob. ISBN 978-2738124609.

- SUJOBERT, Bernard, 2012. La société peut-elle intervenir sur le programme de la statistique publique ? Le Cnis en tant que lieu et outil d'élaboration et de confrontation des attentes sociales et des projets de la statistique publique. In : *Séminaire : Politiques des statistiques*. EHESS. Séance du 6 mars 2012.

- SUPIOT, Alain, 2015. *La gouvernance par les nombres. Cours au Collège de France (2012-2014)*. Fayard. Coll. « Poids et Mesures du Monde ». [en ligne]. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://journals.openedition.org/rfsic/3521*.

- THÉVENOT, Laurent, 1981. Les catégories socioprofessionnelles et leur repérage dans les enquêtes. In : *Études méthodologiques*. [en ligne]. Décembre 1981. Insee. N° 38. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.academia.edu/33996982/Les_cat%C3%A9gories_socioprofessionnelles_et_leur_rep%C3%A9rage_dans_les_enqu%C3%Aate*.

- VANOLI, André, 1989. Le Conseil national de l'information statistique. Quinze ans d'expérience (1972-1987) comme secrétaire général. In : *Courrier des statistiques*. [en ligne]. Décembre 1989. Insee. N° 52, pp. 11-18. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06z98fx/f1.pdf*.

- VOLLE, Michel, 1989. Rapport général de synthèse et débats. In : *L'information statistique en l'an 2000 : continuité et changement - Actes colloque 19-20 avril 1989*, Cnis.

- WRESINSKI, Joseph, 1987. Grande pauvreté et précarité économique et sociale. In : *Rapport au Conseil économique et social. Avis et rapports du CES, Journal Officiel*. [en ligne]. Février 1987. [Consulté le 14 mai 2024]. Disponible à l'adresse : *https://www.lecese.fr/sites/default/files/pdf/Rapports/1987/Rapport-WRESINSKI.pdf*.

Faciliter l'accès aux données de l'Insee

Cubes, catalogue et métadonnées

Jocelyne Mauguin * *et Nicolas Sagnes* **

L'Insee publie sur son site une très grande quantité de données couvrant de nombreux thèmes économiques et sociaux, comme la démographie, l'emploi, les comptes nationaux ou encore les indices de prix. Face à la richesse de son offre, l'Insee doit accompagner son public dans le choix et la compréhension de la donnée. Présenter l'ensemble des statistiques de manière simple et cohérente sur *insee.fr* est donc un défi important. Un premier niveau de diffusion, la datavisualisation, permet d'appréhender un sujet donné à travers des indicateurs synthétiques sous forme de tableaux et de visuels simples, clairs et faciles à comprendre. Mais pour aller plus loin dans l'analyse, des données plus détaillées sont mises à disposition. Elles se présentent généralement sous une forme agrégée : des cubes multidimensionnels croisent différentes variables d'intérêt comme le genre, l'âge ou la catégorie socioprofessionnelle dans les enquêtes ménages. L'enjeu est alors de proposer ces informations dans des formats libres et bien normés, mais également de bien les documenter, en s'appuyant sur des standards internationaux. Ces données doivent aussi être bien cataloguées pour en faciliter la découverte. Pour y accéder, les services de consultation de la donnée de l'Insee se modernisent avec la possibilité de naviguer dans ces cubes. Enfin, les données doivent être accessibles tant par les internautes que par les machines qui les moissonnent : l'usage de ces dernières ouvre des perspectives de nouveaux modes de consommation de la donnée grâce à l'intelligence artificielle.

INSEE (National Institute of Statistics and Economic Studies) publishes on its website a vast amount of data covering numerous economic and social themes such as demography, employment, national accounts, and price index. Given the richness of its website, INSEE must guide its users towards their desired data. Presenting all statistical data in a simple and coherent manner on the *insee.fr* website is therefore a significant challenge. A first level of dissemination, which is data visualisation, allows to comprehend a given subject through synthetic indicators presented as simple, clear, and easy-to-understand visuals. However, to go further in the analysis, more detailed data are made available. These are typically presented in an aggregated form: multidimensional cubes that cross-reference various variables of interest such as gender, age, or socio-professional category in household surveys. The challenge then becomes offering these informations in well-standardized and open-source formats, while also thoroughly documenting them, relying on international standards. These data must also be well-catalogued to facilitate discovery. To access them, INSEE's data consultation services are being updated to make it possible to navigate through these cubes. Finally, the data must be accessible both to internet users and to machines that harvest them: the use of the latter opens up new prospects for data consumption modes through artificial intelligence.

* Cheffe de projet statistique, Insee, DDAR.
jocelyne.mauguin@insee.fr

** Directeur de projet, Insee, DDAR.
nicolas.sagnes@insee.fr

L'Insee publie sur son site un très grand nombre de données de référence d'information économique et statistique. Avec le développement de la donnée, un enjeu très fort est de rendre cette offre la plus à jour, lisible et accessible. Pour cela, de nombreux défis sont à relever : ils portent sur l'importance des formats de données, la documentation et ses standards mais aussi sur les services comme la datavisualisation, le catalogage ou encore la navigation dans les données, sans oublier les APIs[1] indispensables à l'utilisation par des machines.

▶ Présenter simplement de très nombreuses données

L'Insee diffuse des statistiques sur de nombreuses thématiques comme la démographie, l'emploi, les comptes nationaux ou encore les indices de prix. Celles-ci constituent une large part des statistiques publiques, les autres étant produites essentiellement par les services statistiques ministériels. Ces statistiques sont essentielles pour l'élaboration d'études économiques, afin d'éclairer des questions structurantes[2], décomposition de l'inflation, facteurs de la pauvreté, etc.

Cette diversité se traduit par un volume très important de données, dans le sillage de l'explosion de la data au cours des dernières décennies[3]. Quelques chiffres : l'Insee diffuse chaque année sur son site environ 5 000 fichiers XLSX[4] ou encore 70 000 séries historiques (par exemple la série du produit intérieur brut depuis 1949 ou les séries mensuelles des indices de prix à la consommation).

Face à une telle richesse de l'offre, l'Insee doit accompagner ses utilisatrices et utilisateurs, dénommés par la suite « inseenautes », dans le parcours vers la donnée souhaitée. La donnée doit être facilement trouvée et comprise. Pour cela, l'Insee s'efforce de suivre les grands principes du Code de bonnes pratiques de la statistique européenne[5], pierre angulaire du cadre qualité commun aux instituts statistiques européens. La cohérence/comparabilité et l'accessibilité/clarté sont les principes essentiels de la diffusion de statistiques :

> ***L'Insee accompagne ses utilisatrices et utilisateurs dans le choix et la compréhension de la donnée.***

- **Cohérence et comparabilité** : la comparaison des données sur une période raisonnable est possible ; les statistiques sont élaborées sur la base de normes communes pour les définitions, les unités et les nomenclatures dans les différentes enquêtes et sources de données.
- **Données accessibles et claires** : les statistiques sont présentées avec une documentation pour les interpréter correctement et les comparer utilement ; des technologies, des méthodes et des plateformes d'information

1 *Application Programming Interface* ou Interface de Programmation d'Application en français. On parle aussi de service web ou web service. Le site insee.fr propose actuellement un service web dont le résultat respecte le standard international SDMX.
2 Voir (European Commission, 2015).
3 Voir par exemple le site *https://project.opendatamonitor.eu/*.
4 XLSX est une extension de nom de fichier pour tableur au format Office Open XML utilisé par Microsoft Office à partir de la version 2007.
5 *https://www.insee.fr/fr/information/4140105*.

et de communication modernes sont utilisées ; des normes de données ouvertes sont proposées, avec un accès dans un format non propriétaire (Ubaldi, 2013 ; Emilsson et alii, 2020).

Pour appliquer ces principes, il faut aussi tenir compte de la grande diversité des profils et attentes des inseenautes. Citons quelques exemples : une étudiante doit réaliser un exposé sur la comptabilité nationale et a juste besoin de consulter un tableau des grands agrégats comptables (PIB, valeur ajoutée, etc.) sur une page web ; un particulier loue son appartement et veut obtenir tous les ans l'indice de référence des loyers pour réévaluer le loyer ; une chercheuse souhaite analyser les migrations résidentielles entre communes et pour cela télécharger le fichier de données du recensement de la population à un niveau fin, etc. L'Insee choisit d'aller au-devant de tous ces publics et à ce titre, doit proposer différents modes d'accès aux données, en commençant par la datavisualisation (De Jonge et Ten Bosch, 2012).

▶ Des figures pour faciliter l'accès aux données

Pour connaître l'essentiel sur une thématique, l'Insee propose ses chiffres clés, souvent présentés sous forme d'infographies et de tableaux synthétiques : la datavisualisation, c'est-à-dire un ensemble d'indicateurs synthétiques sous forme de visuels simples et faciles à appréhender (Lagarenne et alii, 2023). Elle est privilégiée pour accompagner l'inseenaute dans sa lecture et lui permettre de s'approprier plus facilement les résultats d'une étude. Ainsi, une série chronologique[6] représentée graphiquement sous forme de courbe selon les périodes disponibles satisfait les besoins de tous les publics sur la plupart des thématiques (indices des prix à la consommation ou de production, chiffres du chômage, emploi salarié, etc.).

Un autre exemple est le Tableau de Bord de l'Économie Française (TBEF), service multi-thématique de datavisualisation sur le site insee.fr. Toutes les informations essentielles des différents domaines du débat public (économie, pouvoir d'achat, démographie, société, salaires, entreprises, développement durable, etc.) sont présentées et cela selon trois volets géographiques (Europe, France, territoires) (***figure 1***). L'institut de statistique du Danemark[7] propose, quant à lui, une arborescence thématique dans la rubrique « Trouver des statistiques » de son site : une fois le domaine choisi, les données statistiques sont présentées sous forme de figures avec options de téléchargement de ces données et d'analyse plus fine du domaine.

▶ Télécharger des données pour les réutiliser

Sur le site web de l'Insee, les figures de datavisualisation comprennent systématiquement une option de téléchargement des données. Cela peut servir aux étudiants pour étayer leurs présentations ou aux professeurs d'économie ou de sciences sociales pour préparer des cours. Les journalistes s'intéressent également aux mises à jour de ces indicateurs ou

6 Par exemple cette page regroupant les principaux indices et séries chronologiques : *https://www.insee.fr/fr/statistiques/3530678*.
7 *https://www.dst.dk/en*.

▶ Figure 1 - Vue du Tableau de Bord de l'Économie Française

1. Entrée par thème
2. Mise en perspective territoriale, nationale et européenne
3. Résumé des indicateurs du thème, avec rubrique « En savoir plus » pour accéder à des données complémentaires
4. Pour pouvoir télécharger le tableau des données
5. Datavisualisation des indicateurs synthétiques

résultats d'enquête pour préparer un article ; les data journalistes téléchargent notamment les séries chronologiques pour analyser un ensemble de données en support ou complément d'un article de fond.

Au-delà de la datavisualisation, des fichiers avec des volumes de données plus importants sont disponibles, notamment au format XLSX. Ceux-ci portent sur des niveaux de détail plus fins ou regroupent toutes les informations disponibles sur un thème donné et pas seulement un extrait, comme c'est le cas pour une figure de datavisualisation. Cette offre de fichiers à télécharger s'adresse aux inseenautes qui veulent exploiter directement les données pour leur propre analyse, comme les bureaux d'études, les chercheurs ou certains acteurs locaux. Ainsi, un Conseil régional peut étudier l'activité économique de sa région en utilisant

> ***L'offre de fichiers à télécharger s'adresse aux inseenautes qui exploitent les données pour leur propre analyse.***

les fichiers des créations d'entreprises très finement agrégées selon la localisation géographique, l'activité, la taille et la catégorie juridique des entreprises. Le niveau de détail peut parfois aller jusqu'à des données individuelles comme sur les naissances, les mariages ou les décès issus de l'état civil. Le Conseil régional peut alors faire ses propres agrégations et évaluer les besoins d'installations d'équipement en fonction de la population de sa région.

▶ Organiser l'offre de fichiers de données

Compte tenu de la variété des thématiques et de la diversité des utilisations, l'Insee doit organiser au mieux son offre de fichiers, en commençant par définir leur contenu. Toute la difficulté est de créer des fichiers de données avec des **axes d'analyse** (aussi appelés variables) pertinents pour des inseenautes aux profils différents. Par exemple, concernant la thématique des salaires, si un journaliste s'intéresse aux inégalités de genre, il compare les salaires en privilégiant le genre alors qu'une chargée d'études qui suit l'évolution des salaires tout au long de la carrière professionnelle privilégie plutôt l'âge. Il est donc pertinent de proposer un jeu de données sur les salaires moyens croisés selon les axes d'analyse « genre » et « âge » afin de satisfaire ces deux besoins.

La taille des fichiers de données est aussi un élément important de l'offre. Les fichiers ne doivent être ni trop gros (difficilement exploitables par les inseenautes), ni trop petits (nécessité d'en consulter beaucoup pour analyser un sujet). Par exemple, un fichier de données issues du recensement de la population contenant toutes les informations de l'Insee sur la population française serait beaucoup trop gros et l'inseenaute s'y perdrait facilement. Il doit être fractionné selon des thématiques comme le logement, la famille ou la population étrangère et immigrée. Un découpage peut aussi se faire par le degré d'information : un fichier sur le logement avec les informations principales à connaître, complété par un fichier contenant des informations complémentaires, à destination d'inseenautes plus spécialistes[8].

▶ De la nécessité de normer les fichiers pour les utiliser facilement

Afin d'en faciliter l'exploitation, le format des fichiers de données est généralement normé. Les formats dits plats sont utilisés, au premier rang desquels le format CSV ou plus récemment Parquet (Dondon et Lamarche, 2023) car ils sont facilement lisibles dans un langage de programmation[9], voire dans un tableur[10] si le fichier n'est pas trop volumineux.

8 Un tel découpage reflète la manière dont le recensement de la population est conçu par l'Insee : une exploitation principale et une exploitation complémentaire.
9 R ou Python.
10 Par exemple, Calc de la suite Libre Office.

Le contenu statistique des fichiers est également normé. D'une part, chaque colonne du fichier correspond à une variable déclinée selon ses modalités. Ensuite, les fichiers ne contiennent pas de libellés au niveau des titres de colonnes ou des lignes mais des codes, lesquels sont plus faciles à utiliser quand on veut exploiter le fichier : le titre de colonne est un code relatif à une variable (par exemple le code AGE pour l'âge) et chaque cellule de cette colonne est un code relatif à des modalités de cette variable (par exemple le code « Y35T39 » qui représente la tranche d'âge de 35 à 39 ans). Enfin, les valeurs dans chaque colonne sont dans un même format. Les principaux formats sont la date, la chaîne de caractère, ou le format numérique. Comme le format de chaque colonne est fixe, le contenu de cette dernière peut être exploité plus rapidement par des outils informatiques d'analyse de données.

En accompagnement du fichier, les codes de variables et de leurs modalités sont documentés dans un dictionnaire de codes où ils sont associés à des libellés et regroupés dans des listes de codes. Par exemple, le code de variable AGE a pour libellé Âge et possède une liste de codes formée de codes comme Y35T54 (de libellé « de 35 à 54 ans ») ou Y_GE75 (« 75 ans ou plus ») (***figure 2***). Les variables du fichier peuvent également être attachées à des concepts sémantiques bien définis. Dans l'exemple, la variable de code

▶ Figure 2 - Modélisation de la variable âge

AGE sera attachée à un concept d'âge qui précise s'il s'agit de l'âge en années révolues ou calendaires. De même, des informations générales des tableaux de données comme des précisions dans le titre, l'unité de mesure, le caractère provisoire ou révisé des données sont formalisées et regroupées dans des variables et des listes de codes.

> ***Les variables doivent être comparables d'un jeu de données à un autre lorsqu'elles ont le même sens.***

L'ensemble de ces descriptions des données, nommé métadonnées de structure, est essentiel pour comprendre les données (Bonnans, 2019). Lorsqu'un inseenaute s'intéresse à un sujet, il souhaite généralement obtenir toutes les informations disponibles sur celui-ci. Il faut donc que les variables soient comparables d'un fichier de données à un autre lorsqu'elles ont le même sens. D'où l'importance, la nécessité même, d'harmoniser les métadonnées identiques des différents fichiers de données, pour rendre cohérentes les données entre sources. Pour ce faire, une norme de description en conformité avec les standards internationaux est utilisée par l'Insee.

▶ Des données structurées sous la forme de cubes multidimensionnels

Une notion structurante de la diffusion est celle de « jeu de données » (ou « dataset » en anglais) qui renvoie aux informations contenues dans le fichier de données. Il convient de bien dissocier cette notion de celle de fichier, un même jeu de données pouvant se présenter dans plusieurs fichiers de format différent.

> ***Les jeux de données statistiques sont structurés sous forme d'hypercubes.***

Les jeux de données vont être structurés sous forme de « cubes multidimensionnels » ou « hypercubes » dont les dimensions sont les axes d'analyse. On dénombre plusieurs centaines de tels axes dans toute la diffusion de l'Insee ; les plus fréquents sont l'âge, le sexe, la catégorie socioprofessionnelle, le secteur d'activité ou la catégorie d'entreprise (d'un point de vue juridique ou selon la taille). Au croisement des dimensions de ces cubes, on trouve les valeurs des indicateurs, comme le nombre d'habitants, le nombre d'entreprises ou le revenu.

Ces cubes multidimensionnels sont décrits via le standard international SDMX[11] [12], et plus particulièrement son modèle d'information. Il est utilisé par l'Office statistique de l'Union européenne Eurostat dans ses échanges de données avec les États membres et par les

11 SDMX signifie *Statistical Data and Metadata eXchange*. L'initiative SDMX, lancée en 2002, établit des normes pour faciliter l'échange de données statistiques et de métadonnées entre les organisations internationales et leurs pays membres, à l'aide des technologies modernes de l'information. Ce format est parrainé par sept organisations internationales : la Banque des règlements internationaux (BRI), la Banque centrale européenne (BCE), l'Office statistique de l'Union européenne (Eurostat), le Fonds monétaire international (FMI), l'Organisation de coopération et de développement économiques (OCDE), la Division de statistiques des Nations Unies (DSNU) et la Banque mondiale. Pour plus de détails, voir (SDMX, 2012).

12 Un autre standard s'appuyant sur le modèle d'information du SDMX est le standard de web sémantique Datacube.

Nations Unies pour les indicateurs des objectifs de développement durable[13]. Les portails de ces sites explicitent distinctement cette norme descriptive et son format d'utilisation sur des pages dédiées de la rubrique sur les données. Le dictionnaire de codes du cube y est appelé Définition de Structure de Données (ou « *Data Structure Definition* » en anglais, abrégé en DSD). Les variables du cube sont de trois types : les mesures, les dimensions et les attributs. Ces composants sont définis comme suit :

- **les mesures**

Les mesures représentent un phénomène observé via une statistique (population, opérations comptables en statistique d'entreprise ou comptabilité nationale, nuitées dans les hôtels, indices de prix à la consommation ou à la production industrielle, etc.).

- **les dimensions**

Les dimensions correspondent aux axes d'analyse du phénomène observé. Si l'on s'intéresse à une population, il peut être intéressant de décliner cette mesure selon les dimensions telles que le genre, l'âge ou le statut d'emploi. Deux dimensions ont un statut particulier dans la diffusion : la période temporelle (typiquement l'année de référence des données) et le niveau géographique (la région par exemple).

- **les attributs**

Ils apportent des informations qui ne sont pas indispensables à la valeur mesurée mais nécessaires à la compréhension de ce qui est mesuré. Ils peuvent spécifier par exemple les unités de mesure (personnes physiques ou équivalent temps plein), les facteurs d'échelle (unités ou milliers) et le statut de la valeur (définitive ou provisoire).

Avec cette modélisation, un tableau de la population nantaise ventilée selon différents axes correspond à un cube multidimensionnel où la mesure est la population, où les dimensions sont le sexe, l'âge, la catégorie socioprofessionnelle, la commune, l'année et où l'attribut « nombre de personnes » indique que la population est mesurée en unités et non pas en milliers de personnes par exemple (***figure 3***[14]). Autre exemple : à partir du tableau « Chiffres-clés » sur l'enquête Cadre de vie et sécurité sur le nombre de victimes d'agression ou de vol hors ménage selon l'âge et le sexe[15], la mesure est le nombre de victimes d'agression ou de vol hors ménage. Il est mesuré selon trois dimensions que sont le sexe, l'âge et le type de violences. Les unités de mesure (valeurs en milliers de personnes, taux de plainte en pourcentage) sont informatives et constituent donc des attributs.

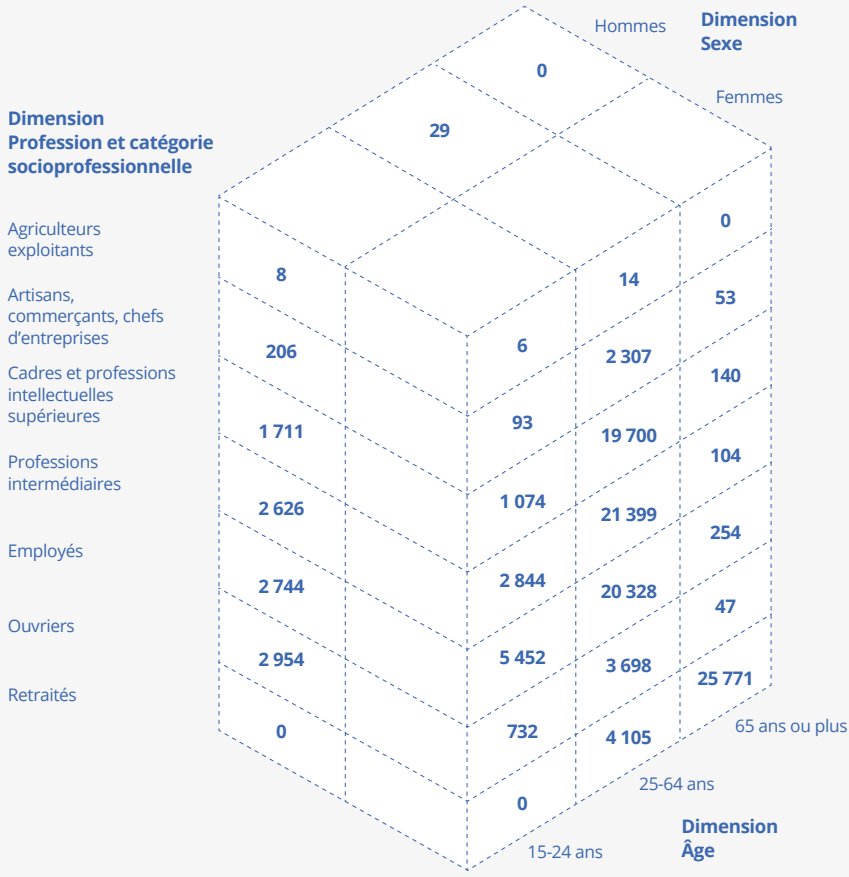
Les dimensions et leurs listes de codes sont réutilisables d'un jeu de données à l'autre, ce qui permet de filtrer les jeux de données qui contiennent la dimension recherchée (par exemple l'âge), voire d'aller plus loin en filtrant plus précisément ceux qui contiennent tel code de cette dimension (concrètement une tranche d'âge particulière). C'est une fonction de recherche très utile pour un catalogue.

13 *https://unstats.un.org/sdgs/dataportal*.
14 On peut ici faire une représentation graphique du cube, car il n'a que trois dimensions.
15 *https://www.insee.fr/fr/statistiques/2525801*.

▶ Figure 3 - Un cube de données sur la population active de 15 ans ou plus à Nantes en 2020 selon le sexe, l'âge et la catégorie socioprofessionnelle

Mesure = Population
Dimension géographique = Nantes
Dimension temporelle = 2020

Unité = Nombre de personnes
= Attribut



Source : Insee, recensement de la population 2020.
Lecture : À Nantes en 2020, 732 femmes de 15 à 24 ans sont ouvrières et 29 hommes de 25 à 64 ans sont agriculteurs exploitants. Les valeurs sont affichées pour les croisements de dimensions visibles.

▶ Un catalogue pour découvrir les jeux de données

Concevoir et structurer les jeux de données n'est pas suffisant. Encore faut-il que l'inseenaute en connaisse l'existence ! Pour cela, ils sont généralement présentés dans un catalogue dédié. Grâce à cet outil, l'inseenaute recherche, selon différents critères, le jeu de données de son choix puis obtient des informations et accède aux données associées. Les critères de recherche sont très importants pour permettre de trouver les fichiers efficacement ; et plus la description des jeux de données est claire, plus le résultat de la recherche sera précis.

Pour bien les décrire, des standards internationaux sont mobilisables comme DCAT[16]. Ce standard décrit les métadonnées de catalogage, c'est-à-dire les champs pertinents d'un jeu de données qui sont autant de critères de recherche possibles : par exemple, sa date de création, son thème, son millésime, sa maille géographique (commune, département, région, etc.) ou encore sa source[17]. Le standard DCAT aide notamment à assurer la comparaison internationale entre les jeux des différents instituts nationaux de statistique (INS). Au final, un jeu aura donc deux types de métadonnées : ses métadonnées de catalogage et ses métadonnées de structure (***figure 4***).

Une fois les jeux de données décrits, ils peuvent être présentés dans une interface Web de catalogue afin d'y accéder facilement. Cette interface présente l'ensemble des jeux de données et permet à l'internaute de les filtrer selon les critères de recherche. Elle affiche également des informations supplémentaires sur chaque jeu (résumé ou couverture temporelle des données).

En pratique, les catalogues disponibles sur les sites internet de statistiques publiques organisent majoritairement leurs jeux de données par une entrée thématique (démographie, emploi, etc.) puis une arborescence plus fine des thèmes pour obtenir le jeu de données souhaité. Un catalogue est disponible sur le site d'Eurostat[18] pour consulter les différents jeux des données statistiques européennes[19].

C'est également le cas de l'institut de statistique allemand Destatis[20] qui met à disposition ses données statistiques via son catalogue Genesis. Comme souvent, ce site sépare le catalogue des autres informations statistiques (tableaux, publications, etc.). Les jeux sont accessibles via la déclinaison de chaque thème. En sélectionner un permet de le visualiser avant de le télécharger. Il en est de même pour le site Agreste[21] du service statistique du ministère de l'Agriculture qui propose dans la rubrique « Chiffres et analyses », l'accès aux tableaux interactifs par une arborescence thématique.

16 DCAT signifie Data Catalog Vocabulary. La Commission européenne s'est attachée à décrire un cadre mutualisé pour cataloguer les informations, dans le cas des catalogues de données. Ces derniers peuvent tout autant être ceux de fournisseurs de données (instituts statistiques, administrations publiques, opérateurs) que des portails d'agrégation proposant des regroupements d'information.

17 *https://www.insee.fr/fr/metadonnees/sources*.

18 *https://ec.europa.eu/eurostat/databrowser/explore/all/all_themes?subtheme=demo&display=list&sort=category*.

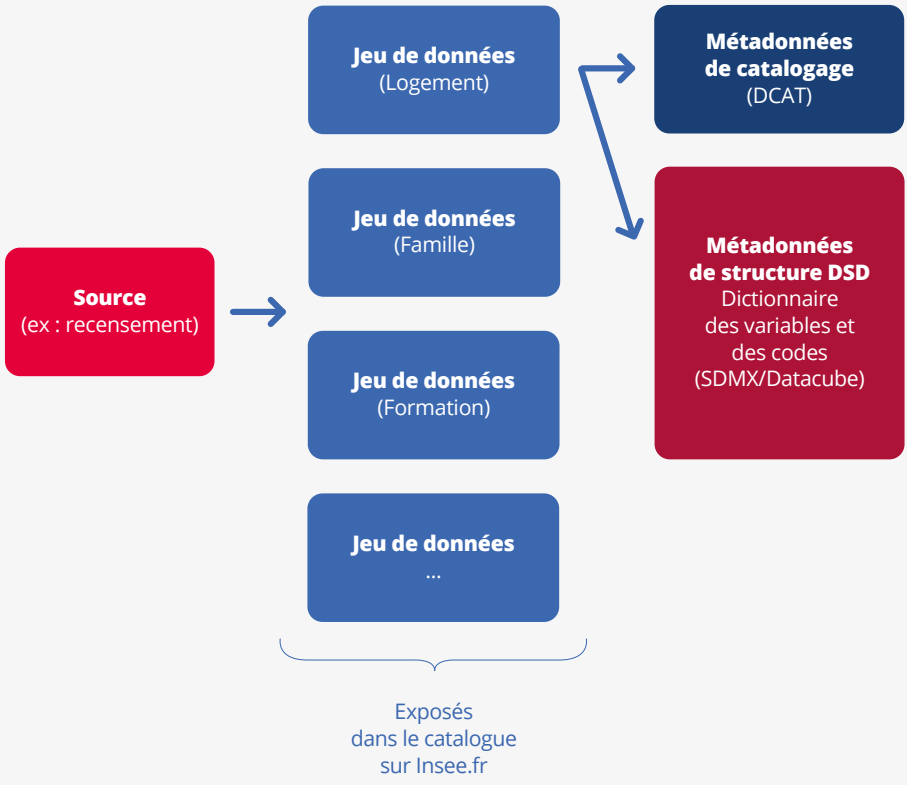
19 Un autre exemple d'accès est la plateforme ouverte des données publiques françaises : *https://www.data.gouv.fr/fr/pages/thematiques-a-la-une/*. La sélection d'un thème permet d'accéder à la documentation détaillant le sujet et aux jeux de données. L'accès peut se faire également via une arborescence restreinte de thèmes plus précis.

20 *https://www.destatis.de/EN/Home/_node.html*.

21 *https://agreste.agriculture.gouv.fr/agreste-web/disaron/!searchurl/4b54e171-2bf3-4c8b-93b9-06e41472066c!cda8b080-3e9e-4368-b41d-7a29c1da0be6/search/*.

L'Insee dispose d'un tel catalogue[22] qui permet un accès plus direct à la donnée via toutes les facettes de recherche (***figure 5***). Conçu dans le cadre d'un projet de modernisation de la diffusion, ce catalogue a vocation à intégrer l'ensemble des données statistiques de l'Insee, tout en répondant aux deux critères d'accessibilité et de clarté du Code de bonnes pratiques de la statistique européenne.

▶ Figure 4 - Les métadonnées d'un jeu de données



22 *https://catalogue-donnees.insee.fr.*

Figure 5 - Catalogue des jeux de données de l'Insee en ligne
(*https://catalogue-donnees.insee.fr*)

Lecture : Les facettes à gauche permettent de filtrer selon différents critères. Les jeux de données sont alors affichés à droite. En sélectionnant le jeu souhaité, on obtient sa description. On peut télécharger le jeu de données au format CSV, et parfois sous forme de fichiers XLSX.

▶ Naviguer dans les cubes pour analyser les données

Une fois le jeu de données choisi dans le catalogue, il est intéressant de l'explorer dynamiquement et de construire ses propres extractions de tableaux. Les INS sont nombreux à proposer ce moyen souple d'exploration. Sur le site de l'Institut italien Istat[23], l'internaute choisit un jeu de données, le visualise directement et accède à la documentation de chaque variable et modalité en cliquant sur les multiples points d'information. L'internaute personnalise ensuite le jeu de données par une sélection des variables et/ou modalités. À l'inverse, le site de l'institut néo-zélandais[24] propose tout de suite de construire son tableau avant de le prévisualiser et de l'exporter. Les explorateurs de ces deux sites sont particulièrement complets concernant le choix des caractéristiques des données (unités, présence ou non des lignes ou colonnes vides de valeurs pour les modalités sélectionnées, etc.) et proposent plusieurs formats d'export des données sélectionnées, pouvant inclure des informations sur les données (provisoire, révisée, etc.).

En France, le site Agreste du ministère de l'Agriculture présente ses cubes en ligne et permet, par exemple, de consulter les cubes issus du recensement agricole sur les exploitations[25]. De même, l'explorateur attaché au catalogue de données sur insee.fr est similaire et permet d'extraire une partie d'un jeu de données en sélectionnant les modalités pertinentes des différents axes. Par exemple, une chargée d'études d'une mairie étudiant les logements locaux pourra filtrer les données de recensement de la population sur sa commune et les communes avoisinantes.

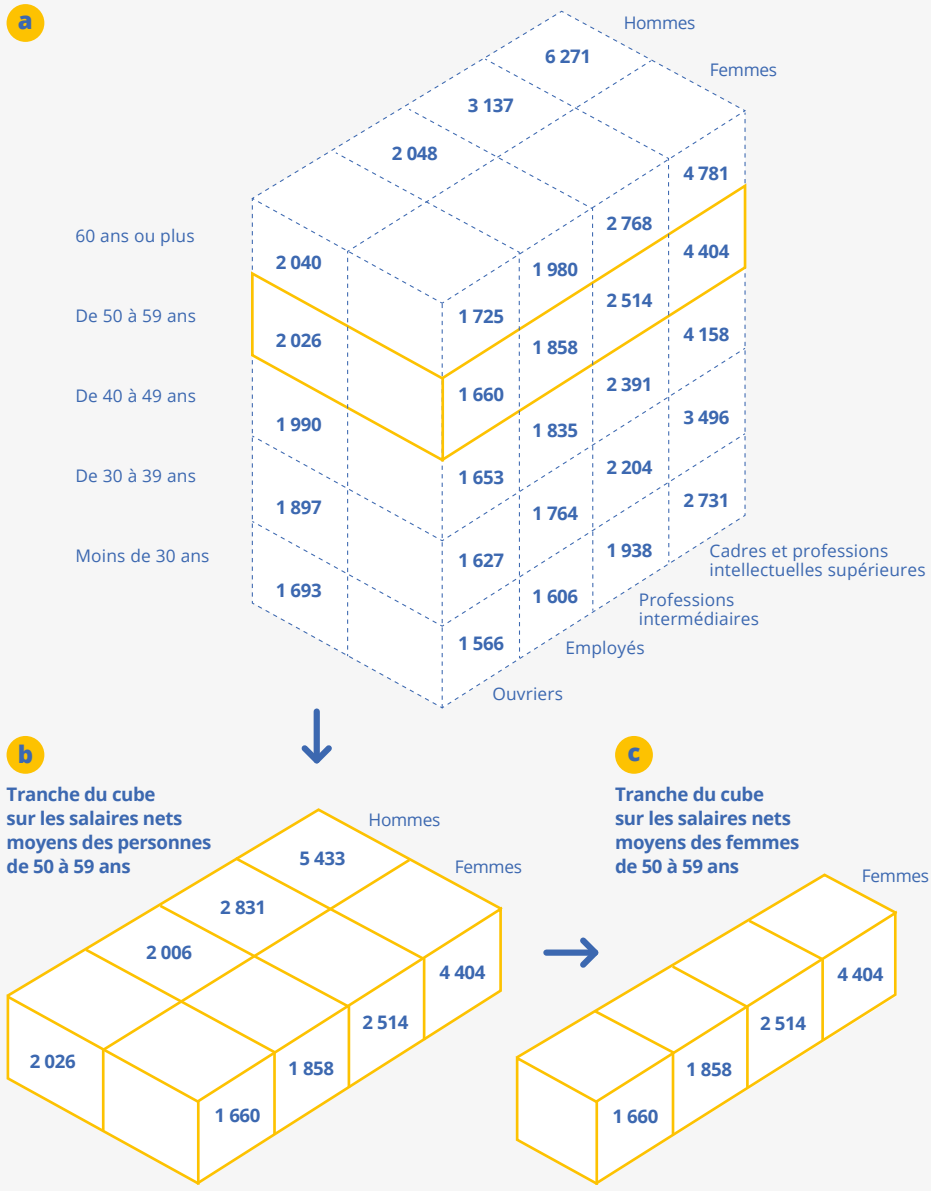
Ces services permettent différents modes d'exploration de cubes multidimensionnels qui peuvent se résumer comme suit[26] :

- **le découpage en tranches** : on fige une dimension à une valeur (en anglais « *slice* » pour tranche) en laissant varier les autres dimensions. Dans l'exemple des salaires moyens selon le sexe, l'âge et la catégorie socioprofessionnelle[27] (***figure 6a***), on s'intéresse aux données spécifiquement sur les personnes de 50 à 59 ans : on tranche ici selon l'âge en figeant la dimension AGE à la modalité « De 50 à 59 ans ». On obtient alors la ventilation des salaires des personnes de 50 à 59 ans selon leur catégorie socioprofessionnelle (***figure 6b***). Si l'on souhaite regarder les écarts de salaires entre les hommes et les femmes, on constitue une tranche plus fine selon le sexe en fixant la dimension SEXE à la modalité « Femme » pour obtenir un cube sur les salaires moyens des femmes de 50 à 59 ans (***figure 6c***).

- **le découpage en sous-cubes** : on croise cette fois-ci plusieurs dimensions entre elles selon certaines modalités (« *dice* » en anglais), pour obtenir un sous-ensemble de données du cube. Sur ce même exemple, on extrait le salaire moyen des femmes ouvrières de 50 à 59 ans.

23 *https://www.istat.it/en/analysis-and-products/databases/statbase*.
24 *https://infoshare.stats.govt.nz/*.
25 Par exemple, le cube sur les exploitations par taille économique et par orientation : *https://agreste.agriculture.gouv.fr/agreste-web/disaron/RA2020_001/detail/*.
26 Ceci est expliqué dans la littérature des architectures de données sous le terme de OLAP, acronyme d'« Online Analytical Processing ». C'est une technologie de base de données optimisée pour les requêtes et les rapports, plutôt que pour le traitement des transactions (Codd et alii, 1993).
27 *https://www.insee.fr/fr/outil-interactif/5369554/index.html*.

▶ Figure 6 - Cube et tranche de cubes sur les salaires nets moyens en 2021 selon le sexe, l'âge et la catégorie socioprofessionnelle



Source : Base Tous salariés 2021.
Lecture : À partir du cube sur les salaires nets moyens en 2021, on peut extraire une tranche de cube sur les salaires nets moyens des personnes de 50 à 59 ans ; on peut ensuite cibler une tranche de cube sur les salaires nets moyens des femmes de 50 à 59 ans.

• **le forage vers le haut ou vers le bas** : il est possible de zoomer et dézoomer (en anglais « *drill up* » et « *drill down* ») sur la donnée. Ceci est particulièrement intéressant dans le cas de niveaux d'agrégation emboîtés, notamment pour des nomenclatures, afin d'étudier les données sur des granularités différentes. Ainsi, sur la population de tranche d'âges 50 à 59 ans, on zoome sur cette tranche d'âges pour cibler les populations sur les deux tranches 50 à 54 ans et 55 à 59 ans, voire pour chaque année de 50 à 59 ans. Autre exemple : on dézoome selon des niveaux géographiques allant de la commune au pays.

Ces services d'exploration de données sont destinés à tous les publics, des particuliers pour rechercher des informations à titre personnel aux professionnels traitant de nombreuses données. Le plus souvent, les personnes intéressées ont toutefois un profil de statisticien ou d'économiste, et les professionnels qui exploitent les données de manière automatique et régulière ont besoin d'autres moyens plus techniques.

▶ Le moissonnage des données par les machines

La consommation de données se fait de plus en plus de machine à machine via des traitements automatisés. C'est le cas, en particulier, des sociétés qui veulent intégrer directement les données de l'Insee dans leur propre système d'information. L'Insee met ses données à disposition via une API, service web pouvant alimenter des applications clientes directement à partir de ses bases de données. Le principe de fonctionnement est le suivant : l'application cliente de l'API est programmée pour interroger régulièrement l'API afin de détecter les mises à jour de données, et le cas échéant, récupérer les dernières informations via une requête (***encadré***). Ce mode de consommation est particulièrement intéressant, car il évite de télécharger manuellement des fichiers sur le site insee.fr et permet grâce au paramétrage de la requête de récupérer uniquement les données d'intérêt (Jacobson et alii, 2011). On parle d'interface machine-machine, car la récupération se fait automatiquement par le programme client, sans aucune intervention manuelle.

De nombreux organismes proposent des APIs (Boyd et alii, 2020). Par exemple, la Cnav (Caisse Nationale d'Assurance Vieillesse) propose une API[28] qui permet de lire des données mises à disposition, comme le nombre de retraites au 31 décembre selon le genre, le montant global de la retraite au 31 décembre par type de droit ou le montant mensuel moyen de la retraite.

C'est le cas aussi de l'OCDE[29] ou de l'institut canadien StatCan[30]. De la même manière, l'Insee offre déjà aujourd'hui des APIs pour différents domaines comme la Banque de données macroéconomiques (BDM) pour les séries macroéconomiques ou la Diffusion de Données Locales (DDL) pour les données locales : elles seront remplacées par une unique API appelée Melodi[31] grâce à une modernisation de la diffusion à l'Insee.

28 *https://data.cnav.fr/api/explore/v2.1/console*. Cet exemple illustre le « *swagger* », cette page internet présentant l'ensemble des requêtes possibles et le format du résultat des requêtes.
29 *https://data.oecd.org/fr/api/*.
30 *https://www.statcan.gc.ca/fr/debut*.
31 Mon Espace de Livraison des données en Open Data de l'Insee.

▶ Encadré. Comment utiliser une API ?

L'exploration par API consiste à utiliser des adresses internet appelées aussi URL* pour interroger le jeu de données. L'API envoie directement le contenu (dans la page du navigateur internet ou dans l'application cliente) sous un format de fichier standard appelé JSON**.

La structure de l'URL est normalisée comme suit :

Nom de l'API / Méthode / Nom / Filtre de la requête.

Les « méthodes » usuelles sont DATA (pour indiquer qu'on récupère des données) et STRUCTURE (pour avoir le détail des métadonnées). Le nom est ensuite celui du jeu de données (pour la méthode DATA) ou de sa métadonnée de structure (pour la méthode STRUCTURE).

Par exemple, le jeu de données DS_TICM*** sur l'équipement des ménages en technologies de l'information et de la communication propose le taux d'équipement internet à domicile et la part des personnes ayant le haut débit fixe ou mobile à domicile. Cette information est recherchée par une entreprise pour évaluer le marché de production de matériel électronique.

Par simplification, supposons que le début de l'URL soit insee.api****. L'entreprise collecte l'ensemble des données du jeu de données dans son navigateur Internet à l'URL suivante : insee.api/DATA/DS_TICM.

L'entreprise peut aussi extraire une partie du jeu de données en filtrant sur les dimensions de celui-ci. Si elle recherche uniquement les taux d'équipement Internet des femmes en 2022, elle ajoutera le filtre correspondant dans la requête API :

insee.api/DATA/DS_TICM?MESURE=EQUIP_INT&SEXE='F'&ANNEE=2022.

MESURE est la dimension de mesure figée au code EQUIP_INT (taux d'équipement) ; SEXE est la dimension du sexe figée au code F (femme) ; ANNEE est la dimension période temporelle figée à 2022.

À noter enfin que seule la récupération des données est possible, tout calcul doit se faire chez le client à partir des données obtenues par l'API.

* *Sigle de l'anglais « uniform resource locator », localisateur universel de ressources. Adresse qui précise la localisation d'une ressource Internet en indiquant le protocole à adopter, le nom de la machine, le chemin d'accès et le nom du fichier : https://www.insee.fr/fr/accueil est une URL.*

** *JavaScript Object Notation (JSON) est un format de données textuel dérivé de la notation des objets du langage JavaScript.*

*** *DS pour dataset et TICM pour l'enquête TIC ménages.*

**** *La future API unique de l'Insee sera http://api-diffusion-catalogue-donnees-externe.insee.fr.*

> ***Les APIs démultiplient le potentiel de réutilisation des données statistiques.***

Cette forme de mise à disposition de données est particulièrement intéressante pour leur diffusion, car elle démultiplie le potentiel de réutilisation des données statistiques. En effet, les outils de datavisualisation s'appuient généralement sur les APIs. Ainsi, l'outil de datavisualisation des salaires sur insee.fr[32] permet d'interroger les données de salaires sous différents angles tels que le métier, la catégorie socioprofessionnelle ou encore le sexe. Lorsque l'internaute choisit une profession pour en connaître le salaire moyen, une requête est faite à l'API de l'Insee qui trouve le chiffre recherché dans la base de données de diffusion de l'Insee et l'envoie à l'outil, qui l'affiche. L'API envoie la valeur disponible la plus fraîche possible puisqu'il a accès directement à la base de données de diffusion. Cet outil sur les salaires reçoit la valeur et l'affiche.

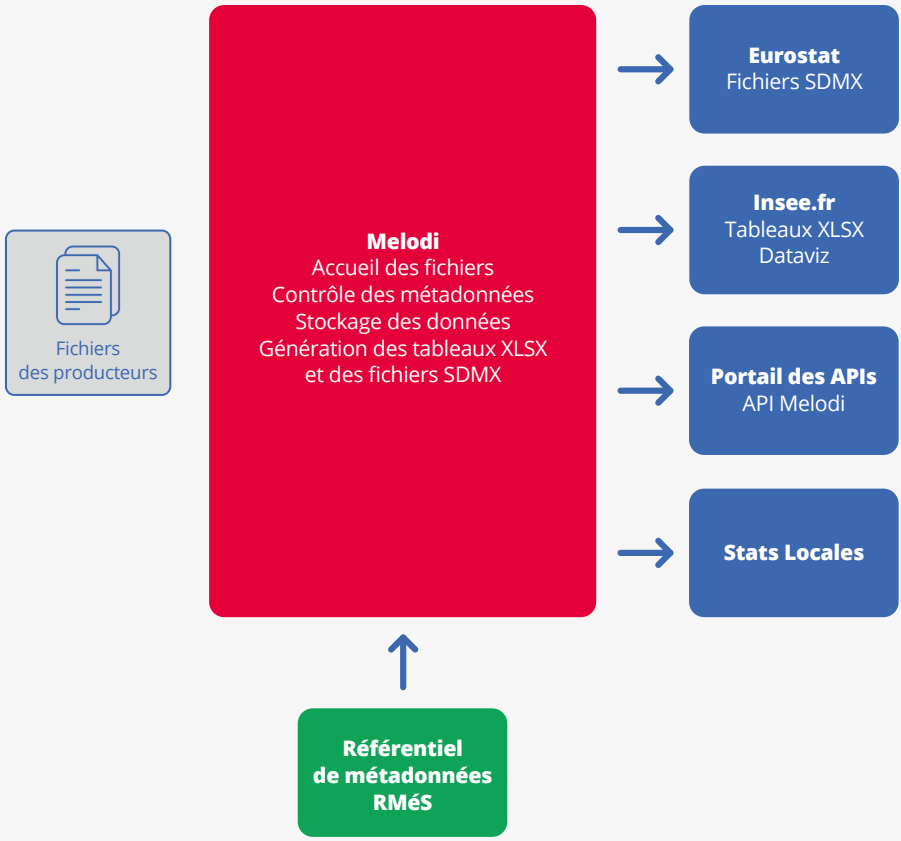
32 *https://www.insee.fr/fr/outil-interactif/5369554/index.html.*

La nouvelle offre de diffusion de l'Insee

Pour offrir les services décrits précédemment (catalogue, description des cubes, explorateur de données, API), l'Insee s'est engagé dans un projet de modernisation appelé Melodi[33]. Ce vaste projet de transformation numérique, qui conduit à une standardisation des données diffusées à l'Insee, repose sur plusieurs principes :

- le premier est de centraliser toutes les données à diffuser dans un même espace appelé entrepôt statistique[34] de données et de réaliser tous les produits de données en ligne (fichiers téléchargeables, tableaux web, fichiers envoyés à Eurostat) à partir des données de cet entrepôt, via des outils mutualisés pour l'ensemble de la diffusion (***figure 7***).

Figure 7 - Le nouveau processus de diffusion des données à l'Insee



33 Mon Espace de Livraison des données en Open Data de l'Insee.
34 Voir (Goossens, 2012) pour une présentation détaillée d'un entrepôt statistique de données.

• le second principe est de décrire ces données selon des métadonnées standardisées (modèle d'information du SDMX/Datacube pour les métadonnées de structure et DCAT pour les métadonnées de catalogage). À ce titre, le processus Melodi s'appuie sur le référentiel de métadonnées statistiques de l'Insee, appelé RMéS. Cette organisation a de fortes implications pour les équipes de production de données de l'Insee qui construisent les données à diffuser et assurent leur livraison dans l'entrepôt Melodi. Elles doivent fournir des fichiers au format attendu[35] et conformes aux métadonnées préalablement décrites dans le référentiel RMéS.

• un troisième principe, « Dites-le-nous une fois », évite que les équipes de production livrent les mêmes données dans différents canaux de diffusion et réduit fortement le risque d'incohérence de données.

Par ailleurs, la mise en place de Melodi constitue une réelle opportunité de revoir l'offre actuelle de données. Tout d'abord, cela conduit à revoir le contenu statistique de la diffusion : décider si des fichiers très peu téléchargés sont maintenus et à l'inverse développer des thèmes très demandés ou nouveaux. Ensuite, il s'agit de redessiner l'offre autour du catalogue de jeux de données, qui constitue un point d'accès central, et de son explorateur. On peut réduire l'offre de fichiers XLSX en la recentrant sur les indicateurs les plus demandés, et inviter les inseenautes qui cherchent des données plus spécifiques ou plus détaillées à consulter l'explorateur pour construire leurs propres tableaux ou alors télécharger les fichiers contenant l'ensemble du jeu de données.

▶ Et demain ?

Ce besoin d'utilisation massive de données statistiques s'avère de plus en plus important et nécessite de mener à bien des évolutions conceptuelles et techniques pour y répondre. On pense notamment à la technologie données ouvertes connectées (*Linked Open Data* ou LOD). Le principe est de structurer les données autour de métadonnées qui sont des ressources universellement utilisées. Par exemple, la région Nouvelle-Aquitaine serait référencée sous la forme d'une « ressource » Internet unique et toute donnée portant sur cette région pointerait vers cette ressource. À la différence d'aujourd'hui où chaque producteur de données est libre de codifier cette région comme il veut, à l'avenir il devrait faire référence à cette codification universelle. Ce recours à des métadonnées universelles permettrait d'assurer une comparabilité entre jeux de données.

Au-delà de la sphère de la statistique publique, l'intelligence artificielle (IA) ouvre la voie à de nouveaux services d'interrogation de la donnée pour la rendre encore plus accessible. La description des métadonnées associées aux données facilite grandement leur compréhension par des algorithmes d'intelligence artificielle. C'est particulièrement utile pour des outils de type chatbot/statbot où l'internaute pose une question – par exemple quel est le dernier taux de chômage ? – question ensuite interprétée par un algorithme d'IA pour interroger la base de données et envoyer la réponse ; la qualité de la description des données sera alors un facteur déterminant dans la capacité de l'IA à répondre de manière pertinente.

35 Beaucoup de livraisons des producteurs se font en SAS ou XLSX dans les processus actuels. Melodi impose des formats dits plats comme le CSV ou le Parquet, adaptés aux fichiers très volumineux.

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 6 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168396?sommaire=4168411*.

- BOYD, Mark, GATTWINKEL, Dietmar, POSADA, Monica et VACCARI, Lorenzino, 2020. An Application Programming Interface (API) framework for digital government. In : *Publications Office of the European Union, Luxembourg*. ISBN : 978-92-76-18980-0. [en ligne]. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120715/final_version_pdf_version.pdf*.

- CODD, Edgar Franck, CODD, Sharon B. et SALLEY, Clynch T., 1993. Providing OLAP to User-Analysts: An IT Mandate. In : *E. F. Codd & Associates*. [en ligne]. [Consulté le 6 février 2024]. Disponible à l'adresse : *http://www.estgv.ipv.pt/paginaspessoais/jloureiro/esi_aid2007_2008/fichas/codd.pdf*.

- DE JONGE, Edwin et TEN BOSCH, Olav, 2012. Visualising official statistics. In : *Site de Statistics Netherlands*. [en ligne]. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://www.researchgate.net/publication/267856653_Visualising_official_statistics*.

- DONDON, Alexis et LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 86-103. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.

- EMILSSON, Cecilia, RIVERA PÉREZ, Jacob A. et UBALDI, Barbara-Chiara, 2020. OECD Open, Useful and Re-usable data (OURdata) Index: 2019. In : *Site de l'OCDE*. [en ligne]. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://web-archive.oecd.org/2020-03-10/547558-ourdata-index-policy-paper-2020.pdf*.

- EUROPEAN COMMISSION, 2015. Creating Value through Open Data. In : *Portail officiel des données européennes*. [en ligne]. Novembre 2015. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf*.

- GOOSSENS, Harry, 2012. The statistical data warehouse: a central data hub, integrating new data sources and statistical output – Contributed Paper at the UNECE Conference of European Statisticians. In : *Site de l'UNECE*. [en ligne]. 8 octobre 2012. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP18.pdf*.

- JACOBSON, Daniel, BRAIL, Greg et WOODS, Dan, 2011. APIs: A Strategy Guide. In : O'Reilly Media, Inc. ISBN : 9781449308926.

- LAGARENNE, Christine, MINODIER, Frédéric et SAMSON, Odile, 2023. Comment présenter nos données pour mieux communiquer ? – La datavisualisation : synthèse et simplicité. In : *Courrier des statistiques*. [en ligne]. 11 décembre 2023. Insee. N° N10, pp. 7-29. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7722079?sommaire=7722116*.

- SDMX, 2012. SDMX 2.1 User Guide. In : *Site de SDMX*. [en ligne]. 19 septembre 2012. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf*.

- UBALDI, Barbara, 2013. Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. In : *OECD Working Papers on Public Governance*, N° 22, OECD Publishing, Paris. [en ligne]. Mai 2013. [Consulté le 6 février 2024]. Disponible à l'adresse : *https://sharingcitiesalliance.knowledgeowl.com/help/open-government-data-towards-empirical-analysis-of-open-government-data-initiatives*.

Les statistiques publiques de l'énergie

Enjeux passés, présents et futurs

Ronan Le Saout[*], *Nicolas Riedinger*[**] *et Bérengère Mesqui*[***]

Les statistiques de l'énergie se sont historiquement construites autour des bilans énergétiques annuels, qui décrivent les différentes étapes du cycle de vie de l'énergie : depuis son extraction de la nature jusqu'à ses différents usages en passant par sa transformation et son transport. Cette comptabilité physique s'appuie sur de nombreuses conventions, qui ont une forte incidence sur certains indicateurs tels que le taux d'indépendance énergétique ou la part de la consommation d'énergie d'origine renouvelable. Des comptes monétaires ont été introduits plus récemment en France, afin notamment d'éclairer les enjeux associés à l'énergie en matière de pouvoir d'achat des ménages et de compétitivité des entreprises.

Les statistiques de l'énergie constituent une source essentielle de la comptabilité des émissions de gaz à effet de serre (GES) et permettent d'éclairer l'évolution de ces dernières, en les croisant avec les statistiques d'autres domaines. La transition énergétique demande une adaptation continue des dispositifs d'observation, pour décrire les nouveaux usages (comme la voiture électrique), les nouvelles formes d'énergie (comme l'hydrogène) mais aussi les leviers de cette transition, comme la rénovation des bâtiments. Des données locales de plus en plus fines sont mises à disposition des acteurs. Les données issues des compteurs communicants offrent des perspectives intéressantes en matière d'évaluation des politiques publiques.

Historically, energy statistics have been built around annual energy balances, which describe the various stages in the life cycle of energy: from its extraction, through its transformation and transport, to its various uses. This physical accounting is based on numerous conventions, which have a major impact on certain indicators such as the energy independence rate or the share of renewable energy consumption. Monetary accounts have been introduced more recently in France, in particular to shed light on the issues associated with energy in terms of household purchasing power and business competitiveness.

Energy statistics are an essential source of greenhouse gas (GHG) emissions accounting and can be used to shed light on trends in GHG emissions by cross-referencing them with statistics from other fields. The energy transition calls for ongoing adaptation of observation systems, to describe new uses (such as electric cars), new forms of energy (such as hydrogen) and also the drivers of this transition, notably building renovation. Increasingly detailed local data is becoming available. Data from smart meters offers interesting possibilities for the evaluation of public policies.

* À la date de la rédaction, expert méthodes statistiques en économie de l'énergie, SDES.
ronan.le-saout@ensai.fr

** Directeur du département du développement durable et du numérique à France Stratégie.
nicolas.riedinger@strategie.gouv.fr

*** Sous-directrice des statistiques de l'énergie, SDES.
berengere.mesqui@developpement-durable.gouv.fr

Les politiques de l'énergie ont de multiples objectifs qui se sont enrichis avec le temps. Au départ, ces politiques cherchaient à garantir la sécurité d'approvisionnement, l'indépendance énergétique, la compétitivité des entreprises et le pouvoir d'achat des ménages. Avec la préoccupation croissante liée au dérèglement climatique, la réduction des émissions de gaz à effet de serre issus de la combustion d'énergie s'est peu à peu imposée comme l'un des objectifs principaux de la politique énergétique autour du projet de « transition énergétique ». Les statistiques nationales de l'énergie ont suivi l'évolution des besoins pour aider à la décision publique et alimenter le débat public. Elles sont largement utilisées dans les débats sur l'évolution des prix, l'origine de nos importations énergétiques ou le nombre de rénovations énergétiques réalisées chaque année.

Ces statistiques présentent la particularité de porter sur un bien pouvant faire l'objet d'une mesure physique. Historiquement, elles se sont construites autour du bilan physique de l'énergie, avec des conventions définies au niveau international. Si certains indicateurs économiques, comme le poids de l'énergie dans le budget, sont suivis depuis longtemps, ce n'est qu'en 2017 qu'un bilan monétaire a été mis en place, enrichissant et présentant dans un cadre cohérent les statistiques sur les prix et les dépenses énergétiques. Les politiques publiques climatiques et environnementales et les évolutions technologiques donnent lieu à des extensions régulières du champ des statistiques de l'énergie, avec des chiffres sur les émissions associées, les énergies renouvelables, la rénovation énergétique, l'hydrogène, etc.

> ***Les statistiques de l'énergie ont fait l'objet d'une institutionnalisation tardive, avec la création de l'Observatoire de l'énergie en 1982.***

Relativement à d'autres domaines, leur institutionnalisation a été tardive. Le premier service labellisé « statistique publique » de l'énergie (l'Observatoire de l'énergie) a été créé en 1982, à la suite des chocs pétroliers. De nombreux autres producteurs de données, parfois créés antérieurement, diffusent toujours de l'information de nature statistique (Ceren[1], CPDP[2], gestionnaires de réseaux[3], Ademe[4], etc.). Ceci conduit à réinterroger continuellement le périmètre de ce qui doit relever de la statistique publique, dans un contexte de besoins évolutifs et de plus en plus importants. Les transitions énergétique et climatique vont encore accroître ces besoins, créant des défis majeurs pour ce système statistique. À cet égard, l'intégration des statistiques publiques de l'énergie au sein du service statistique du ministère chargé de la transition écologique constitue un atout, permettant de bénéficier de synergies avec celles de l'environnement, des transports et du logement, ces deux derniers constituant des secteurs fortement utilisateurs d'énergie et émetteurs de gaz à effet de serre.

Après quelques définitions et conventions statistiques autour du concept d'énergie physique, les développements plus récents en matière de statistiques économiques liées à l'énergie sont présentés avant d'aborder les nouveaux enjeux d'observation dans le contexte de la transition énergétique.

1 Ceren : Centre d'études et de recherches économiques sur l'énergie.
2 CPDP : Comité Professionnel Du Pétrole.
3 Les principaux gestionnaires de réseaux sont RTE et Enedis pour l'électricité, et GRTgaz et GRDF pour le gaz.
4 Ademe : Agence de l'environnement et de la maîtrise de l'énergie, appelée Agence de la transition écologique depuis juin 2020.

▶ De quelle « énergie » parle le statisticien public ?

> ***Le statisticien appréhende l'énergie différemment du physicien.***

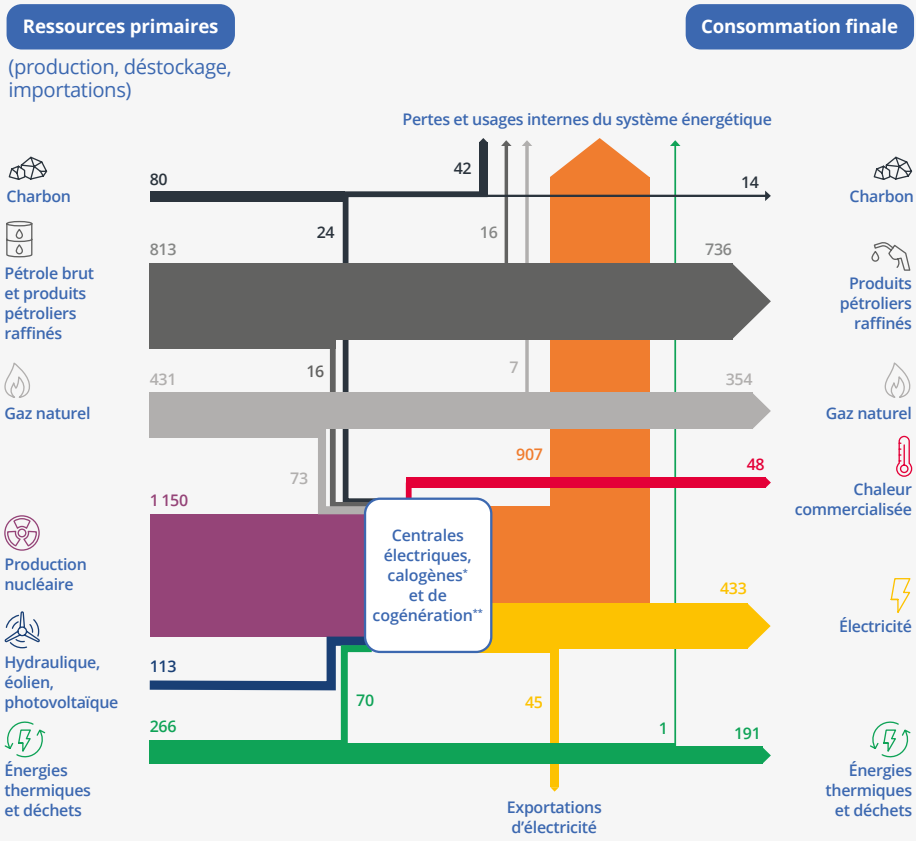
Le statisticien de l'énergie emprunte parfois les unités de mesure du physicien (joule, wattheure et leurs dérivés). Cependant, il n'appréhende pas l'énergie exactement comme ce dernier. Pour preuve, les concepts de production et de consommation, centraux dans les bilans énergétiques, sont étrangers à la thermodynamique[5], voire en contradiction avec son premier principe, selon lequel il n'existe jamais que des transferts d'énergie. La principale raison en est probablement que le statisticien adopte un point de vue anthropocentré et ignore en particulier « *l'énergie existant dans la nature et n'ayant aucun impact direct sur la société* », suivant les « *Recommandations internationales pour les statistiques énergétiques* » (ONU[6], 2020). Cette définition « par la négative » apparaît toutefois vague et incomplète[7] et on peut s'étonner que le périmètre de ce qui est observé ne fasse pas l'objet de principes généraux plus clairs. Pour la France, comme pour les autres pays membres de l'Union européenne ou de l'AIE[8], ce périmètre correspond de fait à la liste limitative des formes d'énergie et des flux énergétiques figurant dans le manuel sur les statistiques de l'énergie AIE/Eurostat[9] de 2005.

Il ressort de l'examen de cette liste que les flux énergétiques considérés sont liés à une intervention humaine visant in fine à rendre des services à des consommateurs, après un éventuel processus de transformation. Le **bilan énergétique**, principale source statistique sur l'énergie, a pour objet de décrire ces flux, en distinguant les ressources et les emplois, dont les totaux doivent en théorie s'équilibrer pour chaque forme d'énergie[10], permettant ainsi de répondre aux deux grandes questions, sur un périmètre géographique donné : comment s'approvisionne-t-on en énergie ? et qui consomme quoi ?

Les approvisionnements sont constitués de la production primaire, c'est-à-dire de l'énergie directement extraite de ressources naturelles, des importations nettes des exportations ainsi que des variations de stocks (***figure 1***). Une partie de ces ressources est directement consommée par les utilisateurs finaux (par exemple, le gaz naturel directement brûlé dans les chaudières des logements), tandis que le reste est transformé en énergie secondaire (par exemple, en électricité) avant utilisation finale. Au cours de ces transformations ainsi que du transport de l'énergie, une partie de cette énergie secondaire est perdue : la consommation finale est donc toujours inférieure à la consommation primaire (cette dernière correspondant au total des approvisionnements par équilibre ressources-emplois).

5 Branche de la physique qui étudie les propriétés des systèmes où interviennent les notions de température et de chaleur.
6 L'Organisation des Nations Unies coordonne la définition des recommandations internationales pour les statistiques énergétiques.
7 En particulier, l'objet des statistiques de l'énergie ne couvre pas du tout la comptabilisation de la quantité totale de chaleur absorbée par la Terre et son atmosphère, qui a pourtant un impact très important sur les sociétés humaines, comme c'est le cas avec le changement climatique.
8 L'Agence internationale de l'énergie (AIE ou IEA en anglais) est une organisation intergouvernementale autonome rattachée à l'Organisation de coopération et de développement économique (OCDE). Elle est composée de 30 pays membres, pour la plupart importateurs de pétrole.
9 Eurostat, l'Office statistique de l'Union européenne, est chargé de publier des statistiques et des indicateurs européens, permettant d'effectuer des comparaisons entre les pays et les régions.
10 À noter qu'à la différence des comptes nationaux par exemple, dont l'équilibrage repose sur des arbitrages entre les estimations des ressources et celles des emplois, le bilan de l'énergie ne cherche pas à réconcilier les deux grandeurs, faisant ainsi apparaître un « écart statistique » pour chaque forme d'énergie.

▶ Figure 1 - Le diagramme de Sankey



Lecture : Un diagramme de Sankey permet de visualiser les flux énergétiques, la largeur des flèches étant proportionnelle au flux physique représenté. Les ressources primaires (importations, production nationale, variations des stocks) se trouvent à gauche, la consommation finale à droite. Le milieu représente la transformation en énergie secondaire et les pertes associées (flux vers le haut).

* *Calogène : réacteur nucléaire utilisé comme source de chaleur.*
** *La cogénération consiste à produire en même temps et dans la même installation de l'énergie thermique à flamme et de l'énergie mécanique.*

▶ Les principales conventions de comptabilisation de l'énergie

Qui dit statistiques dit agrégation, qui dit agrégation dit échelle d'équivalence, et qui dit échelle d'équivalence dit convention(s). Ainsi, tandis que le physicien « mesure » selon des lois de la nature, le statisticien « quantifie » en s'appuyant sur des conventions, pour reprendre la distinction explicitée dans l'étude sociologique de la statistique (Desrosières et Kott, 2005). Les conventions retenues dans les bilans énergétiques, tant pour l'énergie finale que l'énergie primaire, ne doivent pas être perdues de vue par l'utilisateur, au risque d'erreurs d'interprétation potentiellement importantes.

En ce qui concerne l'énergie finale, on pourrait idéalement s'attendre à ce qu'une même quantité d'énergie finale apporte un service équivalent aux consommateurs indépendamment de la forme d'énergie. Cette hypothèse n'est pas toujours vérifiée. D'une part, il existe des usages spécifiques à certaines formes d'énergie, notamment l'électricité, sans possibilité de substitution par d'autres formes d'énergie. D'autre part, en pratique, la consommation d'énergie finale correspond le plus souvent à la quantité achetée par le consommateur. Or, celui-ci peut lui-même transformer l'énergie avant de l'utiliser (par exemple, en brûlant un combustible dans une chaudière), ce qui génère des pertes de transformation, qui ne sont pas considérées comme telles dans le bilan. Il en résulte, par exemple, qu'1 kWh d'énergie finale sous forme de fioul ou de bois à usage de chaleur rend un service moins élevé qu'1 kWh d'énergie finale fournie par un réseau de chaleur. Pour éviter ce biais, dans son rapport sur les statistiques de l'énergie de 1983, l'ONU recommandait de comptabiliser non seulement l'énergie « livrée » mais aussi l'énergie « utile » (ONU, 1983, p. 31). Cette recommandation n'a pas été suivie, probablement pour des raisons de complexité de mise en œuvre.

> ***La comptabilisation des diverses énergies fait l'objet de débats.***

En ce qui concerne l'énergie primaire, la comptabilisation du nucléaire a historiquement fait l'objet de débats. Les organisations internationales ont retenu comme méthode officielle celle du « contenu en énergie ». Cette méthode consiste à comptabiliser la chaleur nucléaire dégagée par la réaction de fission nucléaire, estimée à trois fois la quantité d'électricité produite. Elle a été préférée à la méthode de la « substitution partielle », reposant sur l'estimation de la quantité d'énergies fossiles qui aurait été nécessaire pour produire autant d'électricité que les centrales nucléaires. Cette dernière méthode a le mérite de répondre plus directement à la question de savoir quelle quantité d'énergies fossiles permet d'économiser le recours au nucléaire, mais a l'inconvénient de nécessiter des hypothèses sur les rendements de centrales thermiques fossiles n'existant pas dans la réalité. Faire de telles hypothèses dépasse les compétences habituelles du statisticien, dont le métier est d'observer et non pas de concevoir un monde alternatif (ou un « scénario contrefactuel » comme disent les économistes).

Le développement du photovoltaïque et de l'éolien a relancé le débat sur ces conventions comptables. La convention officielle retenue est de comptabiliser l'énergie primaire éolienne et photovoltaïque à hauteur de la quantité d'électricité produite. En effet, les notions de rendements et de pertes sont jugées peu pertinentes pour ces énergies issues du vent et du soleil, seule une infime part de l'énergie disponible naturellement étant convertie en électricité. Il en résulte que la contribution « comptable » des énergies renouvelables à la baisse de la consommation d'énergies fossiles et, ce faisant, des émissions de CO_2, est sensiblement inférieure à celle qui découlerait de la mise en œuvre de la méthode de substitution partielle, qui n'apparaît cependant pas exempte de défauts non plus (Mesqui et Théron, 2022).

▶ Territorialité et indépendance énergétique : d'où provient l'énergie ?

Le bilan énergétique d'un État donné est censé couvrir les flux énergétiques survenant sur le territoire de cet État[11]. Cependant, l'application de ce principe de territorialité repose sur le choix normatif de ce qui est considéré ou non comme une forme d'énergie (***encadré 1***). Cette convention a un impact sur l'appréciation du taux d'indépendance énergétique, défini comme le ratio entre la production nationale d'énergie primaire et la consommation primaire.

> ***Le taux d'indépendance énergétique de la France chuterait si l'uranium était considéré comme une énergie.***

Les principales formes d'énergie primaire autres que le nucléaire sont les énergies fossiles (pétrole, gaz, charbon) et les énergies renouvelables. Ces énergies peuvent être utilisées directement ou pour produire de l'électricité, et le combustible énergétique est directement comptabilisé dans la consommation primaire. Pour l'électricité nucléaire, ce n'est pas le cas. En effet, dans les conventions internationales, les combustibles nucléaires (uranium et plutonium) ne sont pas considérés comme de l'énergie. C'est la chaleur issue de la réaction, produite là où est installé le réacteur, qui est considérée comme énergie primaire. Cette convention est l'une des plus controversées des bilans énergétiques. Elle a une incidence très forte sur le taux d'indépendance énergétique de la France, qui importe en totalité l'uranium qu'elle utilise. Ce taux est estimé officiellement à 50,6 % en 2022, mais chuterait à 13 %, si l'on considérait l'uranium comme une énergie (SDES, 2023). Bien qu'il ne soit pas explicite dans la documentation méthodologique internationale[12], le principal argument à l'appui de cette convention semble être que l'uranium abonderait sur Terre et en particulier dans des pays bien disposés envers les pays consommateurs. Le facteur véritablement limitant du recours au nucléaire serait donc l'accès à la technologie et non au combustible[13]. Quoiqu'on pense de cette convention, elle souligne que les bilans énergétiques ne peuvent être lus indépendamment du contexte, notamment géopolitique.

Des contraintes en matière d'observation conduisent par ailleurs à certaines entorses au principe de territorialité. Ainsi, la consommation finale de combustibles et de carburants est en pratique attribuée au pays dans lequel ils sont achetés. C'est une difficulté pour croiser la consommation de carburants avec les statistiques de circulation, en particulier des poids lourds, dont certains peuvent traverser la France sans y acheter de carburants. Au-delà du taux d'indépendance énergétique, se pose aussi la question des pays d'origine des importations. Leur détermination est complexe et obéit aussi à des conventions, le cheminement des électrons et des molécules de gaz naturel au sein des réseaux étant en particulier difficilement traçable.

11 C'est une différence avec le volet PEFA (*Physical Energy Flow Accounts*) des comptes de l'environnement ; ce volet obéit, comme les comptes nationaux, au principe de résidence (c'est-à-dire s'intéresse aux flux énergétiques des unités résidentes). En comptabilité nationale, une unité est dite résidente lorsqu'elle a un centre d'intérêt économique sur le territoire économique de ce pays.

12 Le manuel AIE/Eurostat de 2005 mentionne l'existence d'un débat mais sans en poser les termes, tandis que les « Recommandations internationales pour les statistiques énergétiques » de l'ONU de 2019 l'ignorent complètement. C'est d'autant plus frappant que le rapport sur les statistiques de l'énergie de l'ONU de 1983 consacrait un long développement au cycle de l'uranium et à la comptabilité du nucléaire.

13 Un autre argument est que l'uranium est plus facilement stockable en grande quantité que le gaz ou le pétrole.

▶ Encadré 1. Les principales sources de données

Les statistiques de l'énergie reposent principalement sur l'exploitation de données recueillies par le SDES*. D'une part, ce service conduit des enquêtes statistiques au sens de la loi de 1951** (notamment sur la production d'électricité, les réseaux de chaleur, le charbon et les prix du gaz et de l'électricité). D'autre part, il collecte des données prévues par le code de l'énergie. La collecte la plus notable porte sur des données locales de consommation d'énergie. Ses résultats sont mis à disposition du public à une maille très fine (jusqu'au niveau du bâtiment pour les consommations d'énergie résidentielles).

Ces sources propres au SDES sont complétées par des sources externes. Les principales d'entre elles sont issues de la statistique publique, comme l'enquête annuelle sur la consommation d'énergie dans l'industrie (EACEI) de l'Insee ou les statistiques de commerce extérieur du service statistique des Douanes par exemple. Les autres proviennent d'organismes tels que l'Ademe, la Commission de régulation de l'énergie ou la Direction générale de l'énergie et du climat par exemple.

Les sources annuelles utilisées pour le bilan de l'énergie sont décrites plus précisément dans la note méthodologique associée***.

* *SDES : Service des données et études statistiques. Le SDES est rattaché au Commissariat général au développement durable (CGDD), au sein du ministère de la Transition écologique et de la Cohésion des territoires.*

** *Voir fondements juridiques.*

*** *Sous-direction des statistiques de l'énergie (2023), « Méthodologie du bilan énergétique de la France », note méthodologique.*

▶ Les comptes monétaires de l'énergie : un outil complémentaire du bilan physique

La définition de comptes monétaires exprimés en euros offre une autre échelle d'équivalence, reposant sur une logique économique de coûts et non plus de quantité physique d'énergie. Ce suivi monétaire est utile pour éclairer deux des objectifs de la politique énergétique : soutenir la compétitivité des entreprises et le pouvoir d'achat des ménages. De tels comptes monétaires permettent de construire des indicateurs tels que le poids de l'énergie dans le budget des ménages ou dans les charges des entreprises. Ceux-ci sont plus parlants pour le grand public que des statistiques physiques en TWh[14]. Ces comptes constituent en outre un outil d'aide au calibrage de modèles d'évaluation micro- ou macro-économiques.

L'intérêt pour les statistiques monétaires de l'énergie croît dans les périodes de forte hausse des prix de l'énergie, voire de crise énergétique comme celle récente liée à la guerre en Ukraine. Il est donc logique de trouver la proposition d'un « compte satellite[15] » de l'énergie dès le début des années 1980, après les deux chocs pétroliers, dans les archives de l'Observatoire de l'énergie. Le contre-choc pétrolier de 1985 et le maintien d'un bas prix du pétrole jusqu'à la fin du siècle dernier ont probablement contribué à la mise en sommeil de ce projet. Le retour de la question du coût et de la fiscalité de l'énergie au centre des débats a ravivé la demande de statistiques dédiées et a conduit à la mise en place d'un « bilan monétaire » de l'énergie en 2017, novateur au plan international. Ce bilan monétaire s'apparente à un compte satellite. Il est cependant construit dans un cadre un peu différent de celui des comptes nationaux, afin de privilégier la cohérence avec le bilan physique de l'énergie avec lequel il est conjointement diffusé.

14 TWh : Symbole du térawatt-heure, unité de mesure d'énergie. 1 TWh représente l'énergie fournie en 1 heure par une puissance de 1 000 milliards de watts.

15 De manière générale, les comptes satellites visent à fournir des informations dans un domaine particulier, plus fines que les comptes nationaux dans un cadre cohérent avec ces derniers. La France a été pionnière dans l'élaboration de tels comptes, la première commission sectorielle des comptes, celle des transports, ayant été mise en place en 1955.

Le bilan monétaire décrit les flux en euros associés aux flux énergétiques présentés dans le bilan physique pour les énergies faisant l'objet d'échanges marchands (pétrole, gaz, électricité, chaleur, charbon, biocarburants, bois). Il prend la forme, comme le bilan physique, d'un équilibre ressources-emplois. La dépense nationale en énergie en constitue le principal agrégat. Pour chaque forme d'énergie, cette dépense peut être ventilée en emplois, par secteur consommateur, et en ressources, suivant les différentes composantes des prix (importations, production nationale, gestion des réseaux, marges de commerce, taxes, etc.).

Comme c'est usuellement le cas pour les comptes satellites, les bilans physique et monétaire de l'énergie font l'objet d'une présentation et d'une discussion au sein d'une instance dédiée, formée d'experts et de représentants des parties prenantes. Cette instance, dénommée « formation énergie-climat », a été mise en place au sein de la Commission de l'économie du développement durable (CEDD), créée en 2021, aux côtés de trois autres formations ayant succédé aux commissions des comptes de l'environnement, du logement et des transports. Elle a aussi vocation à examiner des travaux réalisés par des acteurs extérieurs à la statistique publique, par exemple sur la précarité énergétique (dont un observatoire est piloté par l'Ademe) ou en matière d'évaluation des politiques publiques de l'énergie (fiscalité, chèque énergie).

▶ Une observation des prix de l'énergie plus délicate avec la libéralisation des marchés

Le lien entre bilan physique et bilan monétaire s'établit en grande partie par la mesure des prix de l'énergie. Leur observation est source de défis propres à chaque forme d'énergie. Le gaz et l'électricité relèvent des économies de réseaux, avec des marchés historiquement intégrés verticalement (de la production à la fourniture d'énergie). La libéralisation de ces deux marchés à la fin des années 2000 a complexifié le travail d'observation des prix. Il existe une enquête dédiée sur la transparence des prix[16].

> ***L'observation des prix de l'énergie est source de défis.***

Les autres formes d'énergie ont chacune leurs spécificités, nécessitant des sources variées. Deux exemples peuvent être mentionnés :

- l'enquête annuelle sur les réseaux de chaleur et de froid (EARCF) pour le prix de la chaleur ;
- les données du Centre d'Études de l'Économie du Bois (CEEB) pour le prix du bois (mesure néanmoins complexe, le bois étant largement échangé sur le marché informel).

Mesurer les dépenses énergétiques nécessite toujours d'associer aux consommations d'énergie les prix correspondants. Pour l'électricité et le gaz, plusieurs prix existent, associés à la chaîne de valeur de ces marchés. Les producteurs, soumis à la concurrence, peuvent vendre de gré à gré[17] ou sur un marché de gros avec la définition d'un prix de marché. Dans le cas particulier de l'électricité nucléaire produite par EDF[18], un tarif et des conditions de vente spécifiques existent à travers le dispositif ARENH « Accès Régulé à l'Électricité Nucléaire

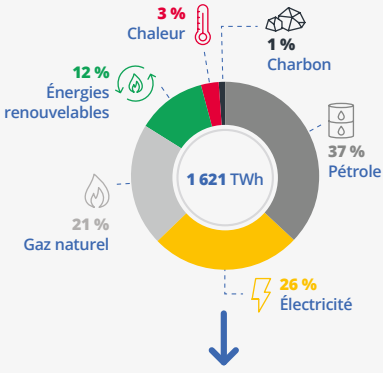
16 Enquête semestrielle « Transparence des prix du gaz et de l'électricité ».
17 Vente directe du producteur sans recourir à un marché.
18 Électricité de France (EDF), entreprise publique française de production et de fourniture d'électricité.

▶ Encadré 2. Les chiffres clés de l'énergie en 2021

Indépendance énergétique

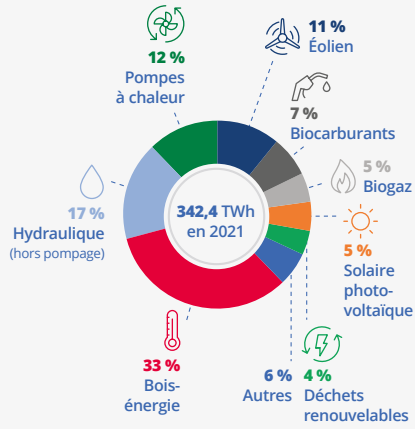
55,1 % de l'énergie consommée est **produite sur le territoire**

Consommation finale à usage énergétique par énergie



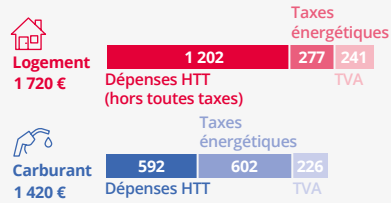
+ 9 % en 2021 de la **consommation finale** à usage **énergétique**

Production primaire d'énergies renouvelables

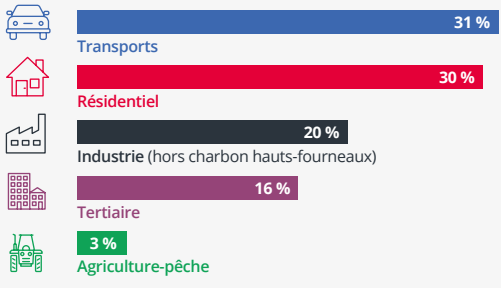


Facture des ménages en énergie

3 141 € en 2021



Consommation finale énergétique par secteur



En 2021, **+ 12,1 %** de la **consommation d'énergie finale** pour le secteur des transports

Source : Bilan énergétique de la France pour 2021. Chiffres Clés de l'Énergie, Édition 2022, DataLab SDES.

Historique » de la loi NOME (Nouvelle Organisation du Marché de l'Électricité) de 2011. Des discussions sont en cours sur le dispositif de remplacement qui sera mis en œuvre à la suite de l'arrêt de l'ARENH prévu fin 2025. Les gestionnaires des réseaux en situation de monopole naturel sont rémunérés par un forfait spécifique pour l'accès aux réseaux. Les fournisseurs d'énergie, soumis à la concurrence, peuvent proposer différents contrats aux consommateurs finaux, notamment : tarif réglementé (pour les fournisseurs historiques d'électricité), tarif fixe sur la durée du contrat, ou tarif indexé sur le prix des marchés de gros du gaz et de l'électricité. Des politiques publiques peuvent également fixer un prix plafond de manière provisoire (comme cela a été le cas en 2022 avec le « bouclier tarifaire ») ou mener des politiques redistributives auprès des ménages modestes (chèque énergie). Ces contrats incluent généralement un abonnement, des tarifs différenciés selon la puissance souscrite et la période Heures Pleines/Heures Creuses (ce que justifie la théorie économique, Percebois, 2001, pour une synthèse).

L'observation des prix n'est pas directe et demeure complexe quelle que soit la forme d'énergie. Mieux mesurer les conséquences des crises énergétiques sur les comptes des agents (marges des entreprises, pouvoir d'achat et revenus des ménages) est un sujet d'actualité. La statistique publique offre des produits complémentaires pour l'analyse des questions économiques : le bilan monétaire, qui a l'avantage d'être cohérent avec le bilan physique, et des enquêtes telles que l'EACEI qui relie statistiques d'entreprises et consommation d'énergie ou l'Enquête Logement pour les ménages (***encadré 2***).

▶ Les statistiques de l'énergie au cœur de la transition écologique et climatique

L'Union européenne et la France se sont engagées à atteindre la neutralité carbone à l'horizon 2050[19]. L'énergie est responsable d'environ 70 % des émissions de gaz à effet de serre (GES) en France. Cet objectif très ambitieux implique donc une transition énergétique, reposant sur la baisse de la consommation d'énergie et le recours croissant à des sources d'énergie décarbonées.

> ***Les dispositifs d'observation des différents champs n'ont pas été initialement conçus pour être interopérables.***

À côté des statistiques d'émissions, dont la description dépasse le cadre de cet article (Carnot et alii, 2023), les statistiques de l'énergie constituent des données essentielles au bon pilotage de la Stratégie nationale bas-carbone (SNBC). La compréhension des dynamiques d'émissions liées à l'énergie requiert de croiser ces statistiques énergétiques avec d'autres, issues des champs économique, démographique, des transports et du logement (Mesqui et Théron, 2022). Ces croisements ne sont pas toujours aisés et peuvent nécessiter certains retraitements. En effet, les dispositifs d'observation des différents champs n'ont pas été initialement conçus pour être interopérables. Les statistiques de circulation et celles de

19 À horizon 2050, les émissions résiduelles de gaz à effet de serre de la France devront être inférieures aux capacités d'absorption par les puits de carbone (sols, forêts, océans, puits technologiques).

consommation de carburants présentent par exemple des différences de périmètre. Le système de comptabilité économique et environnementale, élaboré au niveau de l'ONU, présente à cet égard l'avantage d'offrir un cadre de description des émissions et des flux énergétiques cohérent avec celui des comptes nationaux. Les comptes des flux physiques d'énergie (« *Physical Energy Flow Account*s », PEFA), dont la transmission à Eurostat est obligatoire depuis l'exercice 2014, s'inscrivent dans ce cadre. Ils permettent, en théorie, de rapprocher les consommations d'énergie des différents secteurs économiques avec les grandeurs usuelles des comptes nationaux (production, valeur ajoutée, etc.). Cependant, leur utilisation est encore limitée par la faible profondeur temporelle des séries.

> ***Le besoin des acteurs publics en statistiques locales sur l'énergie est un des enjeux.***

Un autre enjeu est le besoin des acteurs publics en statistiques locales sur ces domaines, pour les aider à réaliser leurs différents exercices de planification (plans climat locaux par exemple). Des données locales de consommation d'électricité, de gaz, de chaleur et de froid, de carburants et de combustibles sont mises à disposition sur le site du service statistique du ministère chargé de la transition écologique, en vertu de l'article 179 de la loi de transition énergétique pour la croissance verte (TEPCV) du 17 août 2015[20]. Pour l'électricité et le gaz, elles incluent des données au niveau des bâtiments (hors bâtiments résidentiels de moins de dix logements pour garantir la protection des données individuelles). Le service statistique ministériel est ainsi placé dans la position relativement inhabituelle de contrôler la qualité des millions de données fournies par d'autres, en l'occurrence, les gestionnaires de réseaux. Ceci pose la délicate question de l'ampleur et de la nature de ce contrôle, nécessairement limité.

▶ Nouveaux usages énergétiques, nouvelles méthodes statistiques

La transition énergétique demande par ailleurs une adaptation continue de l'observation statistique, du fait de l'émergence de nouveaux usages (comme l'autoconsommation électrique par exemple) ainsi que de nouvelles formes d'énergie (***encadré 3***). L'observation des énergies renouvelables donne lieu à une publication spécifique de chiffres clés. Au-delà d'énergies renouvelables traditionnelles comme le bois, l'hydraulique ou les déchets, ce recueil s'est enrichi au cours des vingt dernières années pour prendre en compte l'éolien, le solaire, les pompes à chaleur ou les biocarburants. Les flux de certaines de ces nouvelles formes d'énergie, souvent décentralisées, ne sont pas toujours directement observables. Leur estimation peut alors requérir de combiner des compétences statistiques et des compétences d'ingénieurs. En général, la méthodologie est discutée et définie au niveau international[21]. Un nouveau défi de taille est le suivi statistique de l'hydrogène. À la différence des énergies prémentionnées et comme l'électricité, l'hydrogène n'est pas une source primaire. C'est un vecteur énergétique : il peut être produit de diverses manières et utilisé à des fins variées (production d'électricité notamment). Un tout premier bilan de

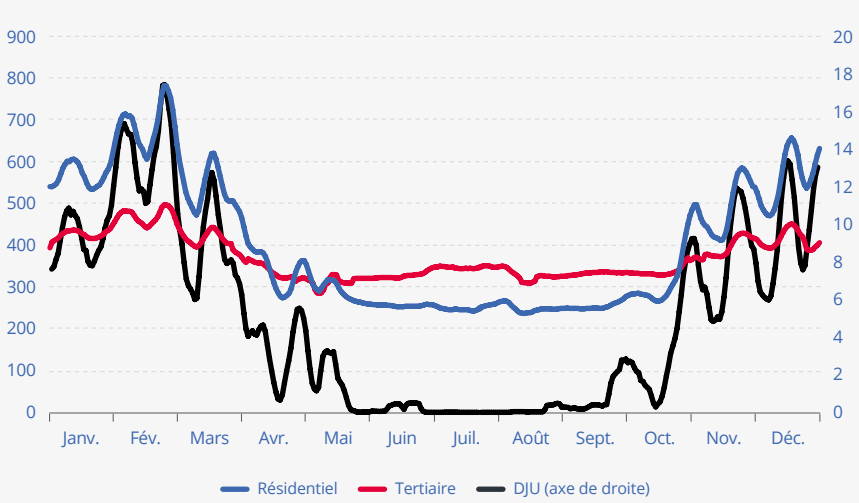
20 Voir fondements juridiques.
21 Par exemple, pour les pompes à chaleur, la chaleur extraite de l'air ou du sol est estimée en multipliant leur puissance électrique par le nombre estimé d'heures de fonctionnement, plus un facteur de performance.

▶ Encadré 3. Qu'est-ce que la correction des variations climatiques ?

La principale source des variations des consommations d'énergie à très court terme est le climat. En effet, plus il fait froid, plus on consomme et réciproquement. Même si à long terme, l'action de l'homme intervient, la variation des conditions météorologiques est largement exogène et les politiques publiques n'ont pas de prise sur cette variation de court terme. Pour identifier le rôle des facteurs socio-économiques dans l'évolution des consommations et des émissions (part des différentes énergies utilisées, efficacité énergétique, comportements des ménages et sobriété, prix des énergies), il est nécessaire de disposer de consommations d'énergie corrigées des variations climatiques. Les consommations d'énergie sont dites thermosensibles. Elles font l'objet de correction dans le cadre du bilan annuel de l'énergie ainsi que dans la conjoncture mensuelle et trimestrielle (en complétant, pour cette dernière, par les traditionnelles corrections CJO-CVS*), à partir d'un indicateur de climat, les degrés jours unifiés (DJU). Sur la période de chauffage, on comptabilise la somme des degrés chaque jour inférieurs à 17°C. Sur la période de climatisation, on comptabilise la somme des degrés chaque jour supérieurs à 21°C. Cet indicateur avait été introduit au XIX^{e} siècle pour des études sur les rendements agricoles (*Rahman, 2011*). L'utilisation des DJU permet également, à l'aide de méthodes économétriques adaptées, de distinguer la part relevant de l'usage de chauffage dans les consommations d'énergie.

* *CJO-CVS : La correction des variations saisonnières et des effets de jours ouvrables est un traitement statistique de la série brute qui vise à en éliminer les composantes cycliques (saisonnalité, nombre de jours ouvrables, année bissextile) pour améliorer l'analyse économique (Ladiray et Quartier-La-Tente, 2018).*

Consommations quotidiennes d'électricité et conditions météorologiques



Exemple sur l'année 2018. Moyenne mobile sur sept jours.
Calcul des auteurs à partir des données Enedis en Open data et des données de température de Météo-France.
Résidentiel : Ensemble des logements du territoire français ;
Tertiaire : Le secteur tertiaire recouvre un vaste champ d'activités qui s'étend du commerce à l'administration, en passant par les transports, les activités financières et immobilières, les services aux entreprises et services aux particuliers, l'éducation, la santé et l'action sociale. Le périmètre du secteur tertiaire est de fait défini par complémentarité avec les activités agricoles et industrielles (secteurs primaire et secondaire).

l'hydrogène a été publié en décembre 2023 (Andrieux, 2023).

La transition énergétique renforce aussi les besoins de données permettant d'en mesurer les effets économiques. Il est nécessaire d'approfondir le suivi statistique de l'investissement dans la production, les réseaux et l'efficacité énergétiques, dans ses dimensions à la fois physique et monétaire (Carnot et alii, 2023). Si certains domaines, comme les énergies renouvelables, font déjà l'objet d'un tel suivi, ce n'est pas encore le cas du nucléaire par exemple.

Mesurer l'amélioration de l'efficacité énergétique des logements

Dans le secteur résidentiel, un facteur explicatif de la baisse de la consommation et des émissions est la plus grande efficacité thermique des logements (***figure 2***). Les normes pour le logement neuf se sont progressivement durcies (Réglementation Thermique 2012, Réglementation Environnementale 2020[22]). Les incitations à la rénovation énergétique du parc existant se sont renforcées. Par exemple, l'interdiction dès 2023, de mise en location de logements très énergivores a pour objectif de supprimer les logements dits « passoires énergétiques ». Un Observatoire national de la rénovation énergétique (ONRE) a été créé en 2019, afin de suivre à la fois les performances énergétiques du parc de logements, les travaux de rénovation énergétique et l'efficacité des politiques publiques afférentes. Le pilotage de cet observatoire a été confié au service statistique du ministère chargé de la transition écologique, qui a l'avantage d'être compétent à la fois sur le logement et l'énergie.

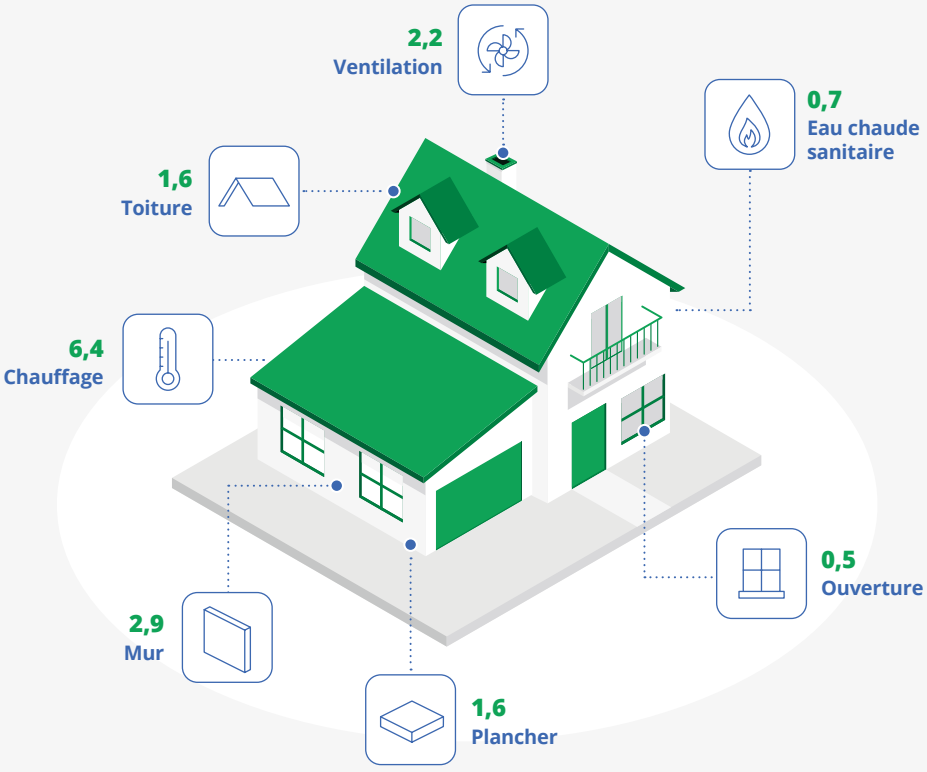
> ***Un Observatoire national de la rénovation énergétique a été créé afin de suivre les performances énergétiques du parc de logements, les travaux de rénovation énergétique et l'efficacité des politiques publiques afférentes.***

Plusieurs sources de données sont disponibles pour effectuer ce suivi. La performance énergétique du parc de logements est mesurée à travers les diagnostics de performance énergétique (DPE). La réalisation d'un DPE est obligatoire lors de la vente ou de la location d'un logement. Ce qui est observé dans les données de la base DPE de l'Ademe, qui collecte l'ensemble des DPE réalisés, n'est donc pas représentatif du parc et nécessite de redresser les données (Le Saout, 2023). De plus, ce qui est mesuré dépend évidemment de la norme de définition du DPE en vigueur. En particulier, la méthode de calcul du DPE a évolué en 2021. La classification de A à G n'est plus basée sur la seule consommation des logements, mais aussi sur leurs émissions de gaz à effet de serre.

Le suivi des travaux de rénovation énergétique des logements (Kraszewski et Le Jeannic, 2023) s'effectue principalement en comptabilisant les aides à la rénovation énergétique, dont les principales sont actuellement MaPrimeRenov' et les Certificats d'économie d'énergie (CEE). Ces aides évoluent dans le temps, concernant les gestes couverts et

22 Voir fondements juridiques.

Figure 2 - Gains moyens de consommation d'énergie en MWh/an lors de travaux de rénovation énergétique dans une maison individuelle



Champ : France métropolitaine, ménages en maison individuelle ayant réalisé des travaux d'économie d'énergie en 2019.
Source : Enquête TREMI 2020, exploitation SDES.

les ménages ciblés (modestes et aisés). Ce qui est mesuré est donc aussi associé aux modalités administratives de définition des aides.

Des enquêtes sur la rénovation énergétique des logements sont donc nécessaires à un pas de temps régulier afin de pouvoir mesurer la part « cachée » de la rénovation énergétique, c'est-à-dire non financée par des aides à la rénovation énergétique. L'enquête TRELO 2023[23] représente à cet égard une source d'information très utile. Alors que les enquêtes précédentes ne couvraient jusqu'à présent que les maisons individuelles (enquêtes TREMI[24]), cette enquête intègre aussi les appartements. Cette extension représente toutefois un défi méthodologique (Le Saout et Rathle, 2023),

23 Enquête sur les travaux de rénovation énergétique dans les logements.
24 Enquête sur les travaux de rénovation énergétique dans les maisons individuelles.

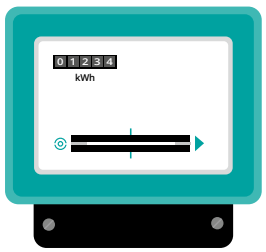
notamment car il s'agit d'interroger une nouvelle unité statistique pour la statistique publique, la copropriété. Au sein d'une copropriété, les travaux de rénovation peuvent en effet concerner à la fois les parties privatives et les parties communes.

L'observation statistique de la rénovation énergétique ne se limite pas au parc de logements et au secteur résidentiel. Des travaux sont engagés sur la mesure de la rénovation énergétique dans le secteur tertiaire, pour lequel la consommation énergétique est importante (immeubles de bureaux, bâtiments recevant du public (écoles, centres commerciaux, etc.)) et fait l'objet d'objectifs de réduction spécifiques (décret Tertiaire du 23 juillet 2019[25]).

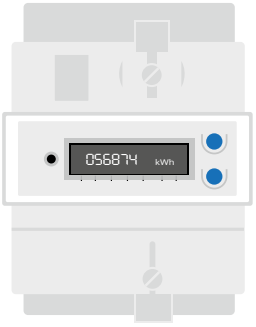
Des données de compteurs communicants pour évaluer les politiques énergétiques

Dans le cadre du suivi des aides ou de l'estimation de la performance énergétique du parc des logements, l'approche est dite « conventionnelle » : la consommation théorique (respectivement la baisse de consommation théorique) est calculée en fonction des caractéristiques techniques (isolation, ensoleillement, équipement de chauffage, etc.) et géographique (zone climatique) du logement. Or cette consommation « conventionnelle » peut s'écarter largement des consommations réelles, du fait de comportements de restriction de consommation des ménages modestes, des potentiels effets rebonds (c'est-à-dire le changement de comportements des ménages en matière de consommation d'énergie à la suite de la mise en place de travaux – Bair et alii, 2017) ou de la qualité inobservée des travaux, par exemple si la qualité des isolants thermiques utilisés est moindre qu'annoncée (Giraudet et alii, 2018). Les politiques de rénovation énergétique visent également des objectifs sociaux de réduction de la précarité énergétique. En cas de forte restriction initiale de l'usage du chauffage pour des raisons financières, lorsque

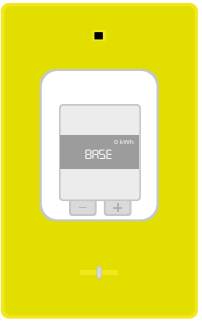
Figure 3 - Les différentes générations de compteurs électriques



Compteur électromécanique



Compteur électronique



Compteur communicant

25 Voir fondements juridiques.

des travaux sont entrepris, potentiellement cela génère un gain de confort et non une baisse de la consommation. Les études économétriques (Fowlie et alii, 2018 ; Penasco et Diaz, 2023 ; Webber et alii, 2015) évaluent ces politiques dans d'autres pays que la France et donnent des résultats contrastés sur les effets de court et long terme.

Un des enjeux pour la statistique publique est de dépasser ce cadre conventionnel pour établir une évaluation réelle des politiques publiques de rénovation énergétique. Pour ce faire, des données mensuelles de consommation individuelle de gaz et d'électricité issues des compteurs communicants (Linky pour l'électricité, Gazpar pour le gaz) seront utilisées (***figure 3***). Cette transmission de données au service statistique du ministère chargé de la transition écologique est autorisée (arrêté du 10 février 2023[26]), après avoir informé les ménages. Elle concerne un échantillon d'un million de ménages ainsi que les ménages répondant à des enquêtes de la statistique publique. Bien que les données mobilisées ne soient pas l'ensemble des courbes de charge individuelles heure par heure (les consommations sont mensuelles), leur utilisation présente déjà un défi pour ce qui est du volume des données et des méthodes statistiques à mobiliser.

> ***La création d'un identifiant unique des logements et des bâtiments dans le futur devrait permettre de fortes avancées dans l'observation statistique.***

Des start-ups (HelloWatt, Homeys, etc.) et l'Institut Français pour la performance du bâtiment (IFPEB) développent des méthodes alternatives basées sur la modélisation prédictive des courbes de charge à partir de données individuelles à des pas de temps très fins (toutes les 30 minutes), et à partir de questionnaires sur l'usage de l'énergie. Ces approches donnent des résultats très détaillés (au niveau de chaque ménage), mais nécessitent l'accord des usagers et présentent donc des biais de sélection difficiles à corriger pour obtenir des résultats agrégés. Dans le futur, l'utilisation de données avec une granularité plus fine sera un des enjeux pour la statistique publique.

L'appariement des données locales de l'énergie avec d'autres sources (données sur le logement, données fiscales, SIRENE) nécessite de développer des méthodologies ad hoc et innovantes d'appariement, au vu de la diversité des formats d'adresse. Hors statistique publique, le CSTB[27] a initié un tel travail d'appariement de sources à travers la Base Nationale des Bâtiments (disponible en open data) et l'usage de méthodes d'apprentissage automatique (machine learning[28]), mais avec une qualité des appariements difficile à évaluer. La création d'un identifiant unique des logements et des bâtiments dans le futur devrait permettre de fortes avancées dans l'observation statistique.

26 Voir fondements juridiques.
27 CSTB : Centre Scientifique et Technique du Bâtiment.
28 L'apprentissage automatique (*machine learning* en anglais) est un champ d'étude de l'intelligence artificielle qui vise à donner aux machines la capacité d'« apprendre » à partir de données, via des modèles mathématiques.

En guise de conclusion

Lors d'une assemblée générale du Cnis[29] en mars 2023, une communication sur les enjeux de la crise énergétique pour le système statistique publique a été présentée (Tavernier, 2023). Pour améliorer le suivi des politiques nationales et européennes (notamment le plan de sobriété énergétique et le bouclier tarifaire), plusieurs sujets étaient identifiés : l'origine des approvisionnements, l'évolution des consommations et des prix, et leurs conséquences sur les comptes des agents (entreprises et ménages). Un défi majeur pour l'avenir est d'adapter le système statistique pour éclairer et évaluer les politiques publiques en lien avec la transition écologique, notamment l'évolution du parc automobile, la disparition des passoires thermiques, la production et les usages de l'hydrogène ou encore le stockage de l'électricité.

Réaliser ces objectifs ambitieux requiert un travail sur la mise à disposition de nouvelles sources de données, à la fois pour l'observation de nouvelles thématiques (hydrogène, voitures électriques, rénovation énergétique dans le tertiaire) mais également avec une granularité plus fine, tant géographique (données locales) que temporelle. Les compteurs communicants peuvent permettre de mobiliser en effet des données très précises, limitées au gaz et à l'électricité ; néanmoins, se posent des défis spécifiques en matière de conditions juridiques d'accès, de volume de données et de méthodes statistiques.

29 Cnis : le Conseil national de l'information statistique est chargé de la concertation entre les producteurs et les utilisateurs de la statistique publique.

▶ Fondements juridiques

- 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 4 mars 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

- Loi relative à la transition énergétique pour la croissance verte (TEPCV) du 17 août 2015. [en ligne]. [Consulté le 4 mars 2024]. Disponible à l'adresse : *https://www.ecologie.gouv.fr/loi-relative-transition-energetique-croissance-verte-tepcv*.

- Décret n° 2019-771 du 23 juillet 2019 relatif aux obligations d'actions de réduction de la consommation d'énergie finale dans des bâtiments à usage tertiaire. In : *site de Légifrance*. [en ligne]. [Consulté le 4 mars 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000038812251*.

- Arrêté du 10 février 2023 concernant la collecte de données à des fins statistiques prévue à l'article L. 142-1 du code de l'énergie. In : *site de Légifrance*. [en ligne]. [Consulté le 4 mars 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000047254496*.

- Réglementation thermique 2012. [en ligne]. [Consulté le 4 mars 2024]. Disponible à l'adresse : *https://www.ecologie.gouv.fr/reglementation-thermique-rt2012*.

- Réglementation environnementale 2020. [en ligne]. [Consulté le 4 mars 2024]. Disponible à l'adresse : *https://www.ecologie.gouv.fr/reglementation-environnementale-re2020*.

▶ Bibliographie

- AIE/Eurostat/OCDE, 2005. Manuel sur les statistiques de l'énergie. [en ligne]. In : *site de Eurostat*. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/fr/web/products-manuals-and-guidelines/-/nrg-2004*.

- ANDRIEUX, Virginie, 2023. L'hydrogène pur : première évaluation des ressources et des usages en France en 2022. In : *site datalab du SDES*. [en ligne]. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.statistiques.developpement-durable.gouv.fr/lhydrogene-pur-premiere-evaluation-des-ressources-et-des-usages-en-france-en-2022*.

- BAIR, Sabrine, BELAïD, Fateh et TEISSIER, Olivier, 2017. Quels enseignements tirer de l'enquête Phébus sur la question de l'effet rebond ? In : *Les ménages et la consommation d'énergie, Théma, Service de l'observation et des statistiques (SOeS)*, pp. 101-113. [en ligne]. Mars 2017. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.ecologie.gouv.fr/sites/default/files/Th%C3%A9ma%20-%20Les%20m%C3%A9nages%20et%20la%20consommation%20d%E2%80%99%C3%A9nergie.pdf*.

- CARNOT, Nicolas, LARRIEU, Sylvain et RIEDINGER, Nicolas, 2023. Les incidences économiques de l'action pour le climat - Indicateurs et données. In : *Rapport thématique de mission, France Stratégie et Insee*. [en ligne]. Mai 2023. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.strategie.gouv.fr/publications/incidences-economiques-de-laction-climat*.

- DESROSIÈRES, Alain et KOTT, Sandrine, 2005. Quantifier. In : *Genèses*. 2005/1, n° 58, pp. 2-3. [en ligne]. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.cairn.info/revue-geneses-2005-1-page-2.htm*.

- FOWLIE, Meredith, GREENSTONE, Michael et WOLFRAM, Catherine, 2018. Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program. In : *The Quarterly Journal of Economics*. Volume 133, n° 3, pp. 1597-1644. [en ligne]. Août 2018. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://academic.oup.com/qje/article/133/3/1597/4828342?login=true*.

- GIRAUDET, Louis-Gaëtan, HOUDE, Sébastien et MAHER, Joseph, 2018. Moral Hazard and the Energy Efficiency Gap: Theory and Evidence. In : *Journal of the Association of Environmental and Resource Economists*. Volume 5, n° 4, pp. 1-68. [en ligne]. Juillet 2017. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://hal.science/hal-01420872/document*.

- KRASZEWSKI, Marlène, LE JEANNIC, Thomas, 2023. Les rénovations énergétiques aidées du secteur résidentiel entre 2016 et 2020. In : *Document de travail de l'ONRE*. [en ligne]. Février 2023. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.statistiques.developpement-durable.gouv.fr/les-renovations-energetiques-aidees-du-secteur-residentiel-entre-2016-et-2020*.

- LADIRAY, Dominique, QUARTIER-LA-TENTE, Alain, 2018. Du bon usage des modèles Reg-ARIMA en désaisonnalisation. In : *Communication aux Journées de Méthodologie Statistique 2018*. [en ligne]. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *http://jms-insee.fr/jms2018s05_1/*.

- LE SAOUT, Ronan, 2023. Un exemple de redressement de données administratives : les diagnostics de performance énergétique. In : *Communication au Colloque Sondages 2022*. [en ligne]. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://drive.google.com/file/d/1f19JrSRx2PxEXgbC3J5wmODgh7q1ttEZ/view*.

- LE SAOUT, Ronan et RATHLE, Jean-Philippe, 2023. L'enquête sur les travaux de rénovation énergétique des logements (Trelo) 2023. In : *Communication au Colloque Sondages 2022*. [en ligne]. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://drive.google.com/file/d/1jyY73D9tcN3fcKMw-10xM2FNzsqFTbYe/view*.

- MESQUI, Bérengère et THÉRON, Guilhem, 2022. Les facteurs d'évolution des émissions de CO_2 liées à l'énergie en France de 1990 à 2020. In : *DataLab, ministère de la Transition écologique et de la Cohésion des territoires*. [en ligne]. Septembre 2022. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.statistiques.developpement-durable.gouv.fr/sites/default/files/2022-09/datalab_106_les_facteurs_d%E2%80%99evolution_des_emissions_de_co2_liees_a_l_energie_en_france_de_1990_a_2020_septembre2022.pdf*.

- ORGANISATION DES NATIONS UNIES, 1983. Concepts et méthodes d'établissement des statistiques de l'énergie et notamment des comptes et bilans énergétiques. In : *Rapport technique, Département des affaires économiques et sociales internationales, Bureau de statistique, Nations Unies*. [en ligne]. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://catalogue.bnf.fr/ark:/12148/cb348238955*.

- ORGANISATION DES NATIONS UNIES, 2020. Recommandations internationales pour les statistiques énergétiques. In : *Handbooks and guidelines*. [en ligne]. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://seea.un.org/fr/document-type/handbooks-and-guidelines*.

- PEÑASCO, Cristina et DÍAZ ANADÓN, Laura, 2023. Assessing the effectiveness of energy efficiency measures in the residential sector gas consumption through dynamic treatment effects: Evidence from England and Wales. In : *Energy Economics*. Volume 117. [en ligne]. Janvier 2023. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.sciencedirect.com/science/article/pii/S0140988322005643*.

- PERCEBOIS, Jacques, 2001. Énergie et théorie économique : un survol. In : *Revue d'économie politique*. Décembre 2001. Volume 111, pp. 815-860. [en ligne]. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.cairn.info/revue-d-economie-politique-2001-6-page-815.htm*.

- RAHMAN, Sumit, 2011. Temperature correction of energy statistics. In : *Rapport du Methodology Advisory Service, UK Office for National Statistics*. [en ligne]. Juin 2011. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.gov.uk/government/publications/temperature-correction-of-energy-statistics-from-the-office-of-national-statistics*.

- SOUS-DIRECTION DES STATISTIQUES DE l'ÉNERGIE, 2023. Méthodologie du bilan énergétique de la France. In : *Note méthodologique*. [en ligne]. Mars 2023. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.statistiques.developpement-durable.gouv.fr/sites/default/files/2023-03/methodologie_bilan_energie_france_mars2023.pdf*.

- TAVERNIER, Jean-Luc, 2023. Éclairer la crise énergétique et ses conséquences sur l'économie française : quels enjeux pour le service statistique public ? In : *Communication à l'assemblée plénière du Cnis*. [en ligne]. 24 janvier 2023. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.cnis.fr/evenements/assemblee-pleniere-2023/?category=1070*.

- WEBBER, Phil, GOULDSON, Andy et KERR, Niall, 2015. The Impacts of Household Retrofit and Domestic Energy Efficiency Schemes: A large scale ex-post evaluation. In : *Energy Policy.* Volume 84, pp. 35-43. [en ligne]. Septembre 2015. [Consulté le 29 janvier 2024]. Disponible à l'adresse : *https://www.researchgate.net/publication/276298840_The_Impacts_of_Household_Retrofit_and_Domestic_Energy_Efficiency_Schemes_A_large_scale_ex-post_evaluation*.

Le Répertoire Statistique des Individus et des Logements (Résil)

Un nouvel univers de référence pour les statistiques démographiques et sociales

Olivier Lefebvre *

**L'Insee construit actuellement un répertoire statistique des individus et des logements (Résil), pour moderniser son dispositif de production des statistiques démographiques et sociales, notamment en tirant davantage parti des données administratives.
Ce projet, déjà bien avancé, doit aboutir fin 2025 et être utilisé dès début 2026.
Ce répertoire permettra, à l'image de son homologue pour les entreprises et leurs établissements, de construire des bases de sondage ou encore de vérifier la couverture des données administratives, mais aussi de construire plus simplement et de façon plus assurée et sécurisée des fichiers enrichis par appariement de sources diverses. Différents processus de production de données pourront ainsi répondre à des exigences croissantes de couverture, de rapidité de traitement, de réactivité.
Pour le construire, divers prérequis sont nécessaires : recourir à différentes sources pour tendre vers l'exhaustivité, pouvoir y appliquer des traitements statistiques, souvent innovants, pour en assurer la qualité, bénéficier d'un cadre juridique pour protéger les données traitées, et enfin disposer d'un « mandat social », c'est-à-dire d'une légitimité, au-delà de la capacité technique ou juridique.**

**INSEE is currently setting up a Statistical Register of Individuals and Dwellings (Résil) in order to modernise its system for producing demographic and social statistics, in particular by making greater use of administrative data.
This project is already well advanced. It should be completed by the end of 2025 and be operational from the beginning of 2026.
Like its counterpart for enterprises and their local units, this register will make it possible to set up sampling bases or to check the coverage of administrative data, as well as to create files enriched by matching different sources in a simpler and more reliable way. The various data production processes will thus be able to respond to the growing demands for coverage, processing speed and responsiveness.
There are a number of prerequisites for setting up such a system: the need to use different sources in order to aim for exhaustiveness; the ability to apply statistical processing, which is often innovative, in order to ensure quality; the need for a legal framework to protect the data processed; and finally, the need for a "social mandate", i.e. a legitimacy that goes beyond technical or legal capacity.**

* Maître d'ouvrage du programme Résil, Insee.
olivier.lefebvre@insee.fr

Pour établir une statistique, il est nécessaire de collecter des données (Dupont, 2023), que ce soit de manière directe (par voie d'enquête) ou indirecte (en mobilisant des données administratives, voire des données détenues par des acteurs privés). Il faut également assurer la qualité de couverture de cette collecte : couvre-t-elle bien toutes les unités statistiques de notre champ d'intérêt ? Sans doublons ni omissions ? Les informations recueillies sont-elles relatives aux « bonnes » unités statistiques ?

Pour cela, disposer d'une liste de toutes les unités statistiques du champ d'observation, sans unités présentes à tort, s'avère extrêmement utile. En effet, on peut alors tirer un échantillon d'enquête dans cette base de sondage, ou encore vérifier que les données administratives utilisées sont exhaustives sur le champ, caractériser leur représentativité et les corriger le cas échéant afin d'éviter un biais lié à un défaut de couverture. De telles listes sont appelées « univers de référence ».

▶ Sous certaines conditions, les répertoires peuvent constituer un univers de référence

Un répertoire, c'est une liste exhaustive d'objets, avec très peu de variables (Rivière, 2022). S'il fallait dessiner un répertoire, il serait à la fois très haut, car il contient potentiellement toutes les observations d'un champ, mais très étroit, car peu de variables y sont gérées.

Les variables présentes dans un répertoire doivent permettre d'identifier sans ambiguïté les unités qu'il contient, notamment pour éviter des doublons, faciliter les mises à jour et permettre de les relier avec d'autres éléments du système d'information. Le répertoire forme la colonne vertébrale du système d'information. Un répertoire est un objet vivant, car mis à jour le plus souvent en continu ; il est néanmoins possible d'en extraire une photo, reflétant la situation un jour donné (souvent au 1er janvier) qui constituera la base de l'univers de référence.

L'Insee s'inscrit dans une longue histoire et une longue expérience en matière de répertoires. Il gère ainsi le Répertoire national d'immatriculation des personnes physiques (RNIPP) depuis 1946 (Espinasse et Roux, 2022). Plus récemment, en 2019, l'Institut a construit et pris en charge le Répertoire électoral unique (REU) pour la gestion des listes électorales (Desmotes-Mainard, 2019).

Dans le domaine des entreprises, il a été confié en 1973 à l'Insee la gestion du répertoire Sirene[1] (Système Informatique pour le Répertoire des ENtreprises et des Établissements).

Ces répertoires, notamment Sirene et le RNIPP, pourraient-ils servir d'« univers de référence » ? Pas directement, car les informations concernant les sorties du champ (départ du territoire national pour les individus, fin d'activité pour une entreprise), ne sont pas connues ou le sont avec retard. En effet, le RNIPP ne contient pas l'adresse des personnes et il n'y a aucune obligation de signaler des départs hors de France. Dans le domaine des entreprises, la cessation officielle intervient souvent bien après l'arrêt effectif de l'activité économique.

1 *https://www.insee.fr/fr/information/6675111.*

> ***L'Insee a considéré qu'il fallait aller plus loin, en créant des répertoires statistiques situés en aval de ces répertoires administratifs.***

L'Insee a donc considéré qu'il fallait aller plus loin, en créant en aval des répertoires administratifs, des répertoires « statistiques ». De tels répertoires permettent de mettre en œuvre des traitements pour des usages statistiques, sans conséquence sur les personnes ou entreprises concernées par les traitements : par exemple, un traitement statistique peut conclure qu'une personne ne vit plus sur le territoire national, mais cela n'aura pas de conséquence sur son affiliation à un régime de retraite.

La statistique d'entreprises est en avance par rapport à la statistique démographique et sociale sur ce point. L'Insee a, en effet, mis en œuvre en 2012 le Système d'immatriculation au répertoire des unités statistiques (Sirus), un répertoire statistique des entreprises et d'établissements, enrichi de quelques informations collectées ou construites par la statistique publique – contours des groupes, niveau d'activité des entreprises, etc. (Hachid et Leclair, 2022).

En matière de statistiques démographiques et sociales, la situation était plus complexe : si on dispose d'une base de sondage construite à partir des fichiers fiscaux pour les enquêtes auprès des ménages, couvrant l'ensemble du champ des logements ordinaires[2], l'absence d'identifiant partagé et la non-exhaustivité des sources rendaient impossible de la considérer comme un « univers de référence » auquel comparer l'ensemble des sources administratives.

Faire face à de nouveaux défis et à de nouvelles opportunités : créer le répertoire statistique des individus et des logements

Plusieurs facteurs ont convergé pour aller plus loin en matière de statistiques démographiques et sociales.

D'une part, des besoins croissants concernant l'exploitation des données administratives, assortis de nouvelles opportunités. Côté besoins, il était nécessaire de répondre à de nouvelles questions, ou plus rapidement à des questions récurrentes, ou encore de mieux rendre compte de la diversité des situations notamment territoriales, via la production de données plus fines que celles obtenues sur la base d'échantillons ; par exemple, le dispositif Filosofi[3] apporte une information finement localisée sur les niveaux de vie, à partir de déclarations fiscales et de données sur les prestations sociales. Côté opportunités, il existe des sources plus nombreuses, plus accessibles, mieux structurées et documentées, de meilleure qualité, ainsi que des capacités de traitement informatique permettant de traiter de très gros volumes de données de manière sécurisée et dans des délais courts (exemple du PASRAU[4]). Comme la plupart des instituts nationaux de statistique, l'Insee investit donc fortement sur ce champ (***encadré 1***).

2 Un logement ordinaire est un logement défini par opposition à un logement en résidence offrant des services spécifiques (résidences pour personnes âgées, pour étudiants, de tourisme, à vocation sociale, pour personnes handicapées, etc.).
3 FIchiers LOcalisés SOciaux et FIscaux.
4 Le dispositif PASRAU (Prélèvement À la Source pour les Revenus AUtres) résulte de travaux de simplification et de rationalisation des déclarations sociales et de la nécessité de transmettre à la DGFiP des informations nécessaires au prélèvement à la source.

▶ Encadré 1. Regard sur les instituts nationaux de statistiques (INS) étrangers

Dans les instituts nationaux de statistiques étrangers, si on constate un besoin partagé d'une production de données statistiques plus riche et réactive, s'appuyant davantage sur des données administratives, on observe des réponses différentes, en fonction du contexte technique, organisationnel, culturel, ou encore juridique. On identifie plusieurs grands modèles :

– un système statistique très intégré, qui utilise de longue date des registres de population administratifs et un identifiant partagé

Les pays dotés de registres de population administratifs reposant sur une obligation de déclaration de changement de domicile et ceux qui s'appuient sur un identifiant personnel partagé par toutes les administrations ont pu construire un système statistique très intégré. L'utilisation courante de la formulation « registre de population » (« register » en anglais) induit une certaine confusion dans la mesure où il ne précise pas sa finalité (administrative ou statistique).

En Finlande par exemple, la production de données statistiques s'appuie à 95 % sur des données issues de registres ou de sources administratives. Le recensement de la population finlandais est, depuis 1980, entièrement fondé sur ce type de données. Le système d'information sur la population rassemble de nombreuses données caractérisant les personnes tout en permettant une interconnexion avec d'autres fichiers (voir le site internet de l'institut de statistique finlandais*). Il s'inscrit dans un État et une société où l'interconnexion de fichiers individuels ne soulève pas de problèmes techniques ou organisationnels majeurs, ni de problème d'acceptation par la population. Il n'est ni envisageable ni envisagé que la France s'oriente vers un tel dispositif.

– aux Pays-Bas : un système très intégré avec registres, sources administratives et données d'enquêtes

L'institut de statistique hollandais (CBS**) s'inscrit dans une démarche générale d'utilisation des sources de données disponibles, le « System of social statistical datasets (SSD) » (Bakker et alii, 2014). C'est un système de registres et d'enquêtes interconnectés et normalisés. Il contient une mine d'informations sur les personnes, les ménages, les emplois et les prestations, les pensions, l'éducation, les hospitalisations, les rapports de criminalité, les logements, les véhicules, etc.

Aux Pays-Bas, il s'agit de la plus importante source de statistiques sociales officielles, sous forme de résultats agrégés qui préservent la confidentialité des données utilisées. Les données individuelles très détaillées auxquelles accèdent les chercheurs restent dans l'environnement sécurisé géré par le CBS, qui vérifie systématiquement que les données ou résultats exportés ne présentent pas de risque de rupture de confidentialité***.

- en Nouvelle-Zélande : un système d'identification statistique des individus et des logements afin de faciliter les appariements, en particulier dans un objectif de recherche

Il n'existe ni identifiant individuel partagé ni répertoire administratif mobilisable (Bycroft et alii, 2022). Un premier répertoire statistique d'individus a été constitué à partir des données d'état civil (naissances et décès), des données aux frontières (immigration-émigration) et des données fiscales. Il permet de relier, à la demande, plusieurs sources de données administratives entre elles, à des finalités statistiques ou de recherche. Ce dispositif a été très utile pour le recensement en 2018, perturbé par des difficultés de collecte susceptibles d'entraîner des biais dans les données produites. Il a permis d'obtenir des résultats statistiques en combinant des données (administratives ou d'enquêtes) sur l'éducation, le marché du travail, les prestations, la justice, la santé et la sécurité, les migrations et les données commerciales.

Depuis 2021, une rénovation est en cours pour disposer d'un véritable répertoire statistique des individus et des logements plus intégré et plus facile à gérer.

L'INS néo-zélandais est très soucieux de la communication réalisée sur ces techniques et la mobilisation des données : des informations sont faites pour rassurer les utilisateurs sur la qualité et pour obtenir l'adhésion de toute la population face à ce nouveau type de collecte, en particulier la population autochtone maori.

* *https://dvv.fi/en/personal-data.*

** *https://www.cbs.nl/nl-nl/.*

*** *https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research.*

D'autre part, la suppression de la taxe d'habitation (TH) sur les résidences principales a accéléré le processus et imposé de trouver de nouvelles solutions pour les statistiques construites à partir de celle-ci.

L'Insee s'est appuyé depuis de longues années sur un fichier issu de la gestion de la TH, avec des utilisations de plus en plus riches. Ce fichier a d'abord permis d'établir une liste de logements dans chaque commune, afin de préparer et de contrôler les enquêtes annuelles de recensement. Il a aussi permis de calculer la population des communes en actualisant

> ***La suppression de la taxe d'habitation sur les résidences principales a imposé de trouver de nouvelles solutions pour les statistiques construites à partir de celle-ci.***

les résultats des enquêtes de recensement. Il a ensuite été utilisé comme base de sondage pour les enquêtes auprès des ménages et servi d'ossature au Fichier Démographique sur les Logements et les Individus (Fidéli) et de référence pour constituer les contours des ménages, préalable essentiel pour le calcul des niveaux de vie, y compris à des niveaux géographiques fins (Lamarche et Lollivier, 2021).

L'objectif avec Résil est donc de construire un dispositif rendant a minima les mêmes services que le fichier issu de la TH, c'est-à-dire réaliser le recensement, échantillonner les enquêtes auprès des ménages, reconstituer les niveaux de vie et créer des fichiers composites d'étude. L'Insee en a profité pour aller plus loin, sur trois axes : renforcer la pérennité du dispositif (ne pas revivre l'épisode de la suppression de la TH), progresser sur le contrôle du champ couvert et rendre le « service d'univers de référence » pour les processus de collecte et d'exploitation de données administratives.

La cible est donc de construire un dispositif plus robuste, à plusieurs titres :

- créer un répertoire, avec des unités identifiées sans ambiguïté et stables dans le temps, donc de meilleure qualité ;
- étendre le champ aux logements non couverts par la taxe d'habitation, notamment les logements des communautés (Ehpad[5], internats, etc.) ;
- mobiliser plusieurs sources pour alimenter et mettre à jour ce répertoire, pour une couverture plus complète du champ d'observation, en assurant la continuité en cas de changement sur les données. La suppression de la taxe d'habitation a fait prendre conscience à l'Insee que ce risque était réel.

Il est également possible, grâce au caractère central du répertoire, de développer les appariements de données (et donc de production multi-sources) et de contrôler la couverture des sources administratives, au regard de leur utilisation à des fins statistiques.

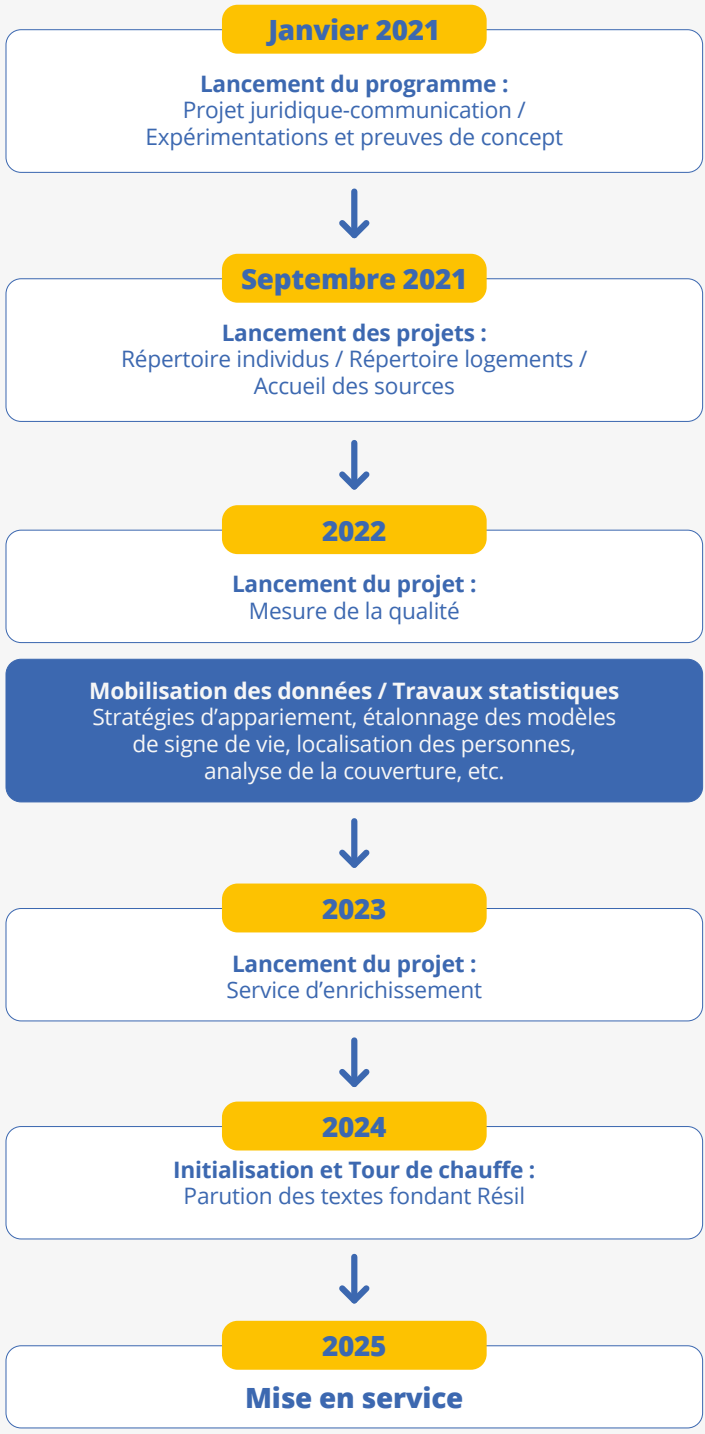
> ***La cible est le répertoire statistique des individus et des logements (Résil), pour créer les univers de référence attendus pour la statistique démographique et sociale et faciliter les appariements.***

La cible est le répertoire statistique des individus et des logements (Résil), permettant de créer les univers de référence attendus pour la statistique démographique et sociale, mais aussi de faciliter les appariements de données administratives entre elles ou avec d'autres données, principalement d'enquêtes. L'objectif est d'y parvenir en 2025 (***figure 1***).

Selon les termes du décret créant Résil[6], ce dernier « a pour finalité, en vue de contribuer au débat public ainsi qu'à l'élaboration et

5 Ehpad : Établissement d'hébergement pour personnes âgées dépendantes.
6 Voir les références juridiques en fin d'article.

▶ Figure 1 - Résil : un projet qui se construit étape par étape



à l'évaluation des politiques publiques, de renforcer la capacité de l'Institut national de la statistique et des études économiques et des services statistiques ministériels à produire des données et études statistiques, en permettant l'établissement d'un répertoire national de la population et des logements et en facilitant les appariements de données administratives avec d'autres sources de données ».

Des usages variés, essentiels pour la construction des statistiques démographiques et sociales

Résil est donc une infrastructure de production permettant de répondre à plusieurs objectifs exclusivement statistiques.

Il offrira, uniquement au service statistique public, un service d'appariement de données avec des usages multiples[7] et pouvant notamment servir (Dupont, 2023) à :

- alléger la collecte d'information par enquête en ne posant des questions que sur des aspects non couverts par les données administratives ;
- apparier des données administratives entre elles, afin de produire des statistiques à une échelle fine, impossible à produire sur échantillon ;
- enrichir un fichier par des variables complémentaires permettant d'approfondir les analyses (par exemple l'ajout d'informations sur le revenu dans l'enquête sur les ressources et les conditions de vie (SRCV)) ;
- éclairer des aspects méthodologiques particuliers : par exemple, apparier les fichiers de l'enquête Emploi et le fichier historique des demandeurs d'emploi afin de mesurer la différence entre les concepts de chômeur au sens du BIT[8] et demandeur d'emploi inscrit à France Travail ;
- évaluer des politiques publiques (suivi de trajectoire de bénéficiaires d'aides particulières).

Ces appariements seront meilleurs et moins coûteux grâce à la présence d'identifiants communs ; on pourra par ailleurs en mesurer plus facilement la représentativité et la qualité.

> ***Résil permettra également de mesurer la qualité des sources administratives qui constituent une des ressources principales de la statistique publique.***

Résil permettra également de mesurer la qualité des sources administratives qui constituent une des ressources principales de la statistique publique. Il sera possible de comparer le champ effectivement couvert par une source statistique à la liste des individus ou logements présents dans Résil, et ainsi de détecter d'éventuels défauts de couverture.

Dans le prolongement des dispositifs existants, la base de sondage dans laquelle tirer des échantillons pour les enquêtes réalisées par le service statistique public auprès des ménages sera issue de Résil. La

7 Voir l'article de Koumarianos, Lefebvre et Malherbe sur les appariements dans ce même numéro.
8 BIT : Bureau International du Travail.

couverture sera mieux assurée qu'auparavant, à la fois par la diversité des sources, l'ajout des personnes vivant dans les communautés et la prise en compte plus rapide des décès ; il sera également possible d'ajouter à la base de sondage des variables d'autres origines que les sources fiscales, ce qui rendra l'échantillonnage plus précis.

Résil fournira l'information permettant de préparer et réaliser les enquêtes annuelles de recensement, et d'en extrapoler les résultats, à l'instar de l'utilisation actuelle des fichiers issus de la taxe d'habitation (***encadré 2***). Résil pourra ainsi être mobilisé pour faciliter la production d'indicateurs démographiques plus précoces.

Enfin, Résil permettra de progresser sur la cohérence des traitements et des données produites, mais aussi sur l'efficience de ces traitements, via l'utilisation d'outils partagés et performants. En facilitant le rapprochement de données, mais aussi leur confrontation, en unifiant les données de référence et les marges de calage[9], il permet de « casser les silos » du système d'information actuel organisé par source.

▶ Encadré 2. Résil et le Recensement de la population, un partenariat gagnant-gagnant

Résil va produire, en remplacement de la taxe d'habitation, les données dont le recensement a besoin pour préparer, contrôler et extrapoler les enquêtes de recensement. Il pourrait permettre en particulier de fournir des estimations plus précoces, comme demandé par de nombreux utilisateurs et par Eurostat.

Inversement, le recensement permet d'évaluer la qualité des sources utilisées en entrée de Résil, qu'il s'agisse des individus ou des logements.

Par ailleurs, des méthodes statistiques dites d'estimation par système dual (Zhang et Dunne, 2017) devraient permettre à terme de mesurer la couverture de Résil d'une part, du recensement d'autre part, en confrontant ces deux sources et donc d'identifier des biais de couverture.

Comme dans d'autres pays, Résil devrait permettre de moderniser le système de recensement ; cependant, il est trop tôt pour dire comment se traduira cette modernisation et dans quel calendrier elle se fera.

▶ Que contiendra Résil ?

Concrètement, Résil sera constitué de deux sous-répertoires statistiques distincts, reliés l'un à l'autre : un répertoire des individus et un répertoire des logements. Ils seront mis à jour régulièrement sur les naissances et les décès et avec les sources fiscales et sociales (données sur les prestations sociales et familiales de la Cnaf[10], DSN[11], PASRAU) ou d'autres sources sur un public spécifique, comme les fichiers d'inscription dans l'enseignement supérieur.

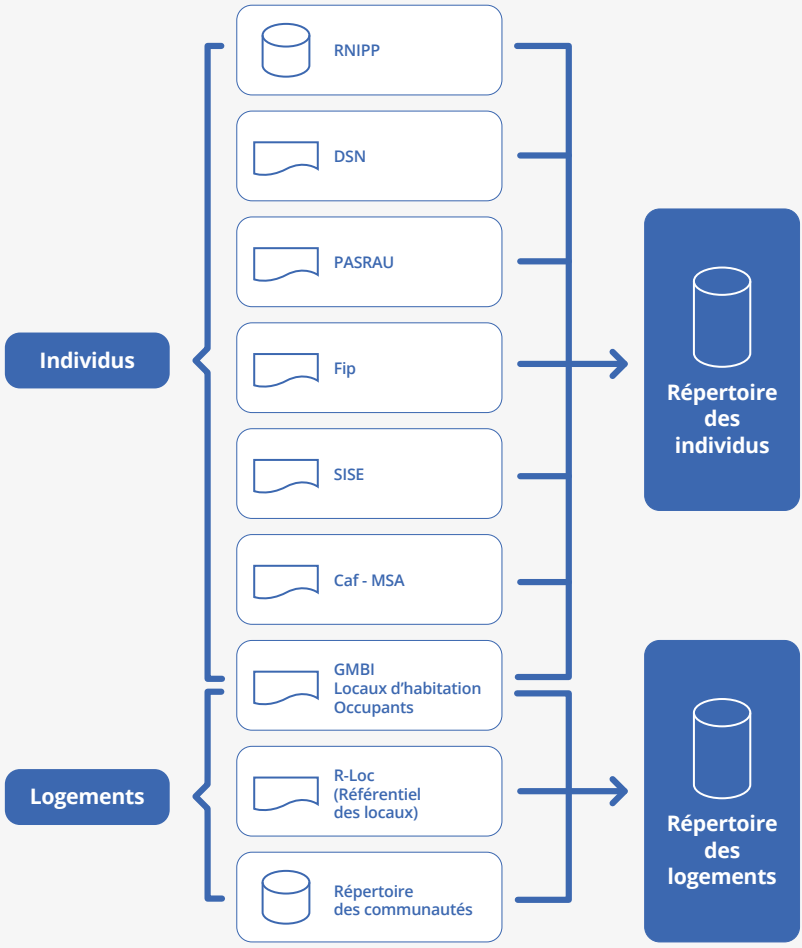
Un point essentiel : dans ces sources, on ne retient que des données d'identification des individus et des logements, des données d'adresse et des liens entre individus et logements (***figure 2***).

9 Le calage sur marges est une technique statistique visant à améliorer la précision des enquêtes par sondage. Elle consiste à modifier les poids de sondage des individus de l'échantillon afin que les totaux pondérés sur l'échantillon de certaines variables correspondent aux totaux connus pour ces variables sur l'ensemble du champ d'observation (la population, le parc de logements, les entreprises) (Deville et alii, 1992).

10 Cnaf : Caisse nationale d'allocations familiales.

11 DSN : Déclaration sociale nominative.

▶ Figure 2 - Résil et ses sources d'alimentation*



RNIPP : Répertoire National d'Immatriculation des Personnes Physiques
DSN : Déclaration sociale nominative
Dispositif PASRAU : Prélèvement À la Source pour les Revenus AUtres
Fip : Fichier d'imposition des personnes
SISE : Système d'information sur le suivi de l'étudiant
Caf : Caisses d'allocations familiales
MSA : Mutualité sociale agricole
GMBI : Gérer mes biens immobiliers

** Cette liste de sources peut évoluer dans le temps, sous réserve d'un avis favorable de la Cnil et du Cnis.*

À partir de ces répertoires seront produites annuellement des « photographies » composant l'univers de référence :

- la liste des individus présents sur le territoire national au 1er janvier ;
- la liste des logements situés sur le territoire national au 1er janvier, et leur statut (résidence principale, secondaire, logement vacant) ;
- la liste des ménages[12] et leur composition au 1er janvier.

Les listes des individus et des logements serviront de référence pour la statistique démographique et sociale. Les listes de ménages sont indispensables pour construire des données telles que les niveaux de vie ou pour réaliser des enquêtes.

Les informations contenues dans les répertoires Résil seront essentiellement des clés d'identification pour assurer le rôle de liste de référence (éviter les oublis et les doublons) et pour réaliser les appariements :

> ***Le numéro d'inscription au RNIPP (le NIR ou numéro de sécurité sociale) ne sera pas stocké dans Résil.***

- des identifiants d'individus : le code statistique non signifiant (CSNS) (Espinasse et alii, 2023) créé par la loi pour une République numérique en 2016[13] pour faciliter les appariements entre sources au sein du service statistique public, un identifiant spécifique et strictement interne à Résil, pérenne pour la gestion de l'historique, les identifiants des sources utilisées dans Résil (pour les individus et les foyers). Le numéro d'inscription au RNIPP (le NIR ou numéro de sécurité sociale) ne sera pas stocké dans Résil ;
- des identifiants de logements : un identifiant spécifique à Résil pérenne, les identifiants des sources utilisées pour Résil ;
- chaque logement aura un identifiant d'adresse issu du référentiel d'adresses de l'Insee du type « BZ140JD » (et non l'adresse en clair « 8 rue Zéphyrin Brioché à Gleux-lès-Lure, département de la Haute-Saône »), ce qui le rend inexploitable en dehors de l'Insee ;
- les données d'état civil : nom, prénom, date et lieu de naissance, le cas échéant date du décès ;
- des liens entre les individus et leur logement ou leur adresse[14] d'habitation, avec, le cas échéant, plusieurs logements possibles pour un même individu selon les sources ; in fine une résidence principale sera déterminée pour chaque individu.

Quelques autres variables permettant la gestion du répertoire et la mesure de la qualité :

- dates de mise à jour des données ;
- dates d'effet (date de début et date de fin) pour certaines variables de Résil pour lesquelles on souhaite conserver un historique ;

12 Un ménage regroupe l'ensemble des personnes partageant un même logement.
13 Voir les références juridiques en fin d'article.
14 Dans certains cas, il ne sera pas possible de faire la distinction entre plusieurs logements situés à la même adresse.

- indicateur de présence sur le territoire français ;
- présence de la personne ou du logement dans chaque source administrative (oui/non).

Résil ne contiendra aucune autre information. Les données telles que le revenu, l'état matrimonial, la profession, la surface des logements, etc. figureront dans des bases spécifiques indépendantes de Résil, et ne seront mobilisées qu'à la demande dans le cadre d'un traitement distinct. Résil ne sera donc pas une « méga-base » contenant tout ce que l'on sait sur chaque individu ou chaque logement.

Le répertoire Résil s'appuie sur quatre piliers majeurs : des données d'origines diverses, des traitements statistiques permettant de les transformer en un répertoire statistique de qualité, un fondement juridique solide et un « mandat social ».

▶ Quatre piliers pour Résil

Premier pilier : des sources d'informations diverses pour un résultat robuste

Résil utilisera plusieurs sources de données pour assurer :

- la meilleure couverture possible de la population (aucune source administrative n'est exhaustive, et aucune n'est parfaitement conforme aux concepts statistiques de population résidente[15]) ;
- une localisation plus précise des individus et une meilleure appréhension des résidences multiples ;
- la pérennité du dispositif au défaut d'une source, voire sa transformation ou sa disparition, pour ne pas subir de rupture de collecte (exemple de la suppression de la taxe d'habitation).

Les résultats des premières expérimentations confirment l'intérêt de mobiliser chacune de ces sources, en sus de la seule source fiscale, en ce qui concerne la couverture de la population. Si on prend l'Enquête Annuelle de Recensement (EAR) comme référence, le gain global de couverture[16] est de l'ordre de 2 points chez les individus de plus de 18 ans, mais il s'élève à 10 points pour les 21-25 ans. La couverture par âge est ainsi plus homogène qu'avec la seule source fiscale.

Pour les personnes vivant en communauté[17] (environ 1,3 million de personnes dans des maisons de retraites, cités universitaires, internats, foyers de travailleurs, établissements pénitentiaires, communautés religieuses, etc.), le taux de couverture progresse de 10 points, de 80 % à 90 %.

15 La population résidente comprend toutes les personnes résidant en France, quelle que soit leur nationalité et leur situation, à partir du moment où elles sont en France depuis au moins un an, ou, si elles viennent d'arriver, qu'elles ont l'intention d'y rester pour au moins un an. En revanche, les personnes de passage (touristes, travailleurs saisonniers ou étudiants étrangers venant pour une année scolaire de 9 mois) n'y figurent pas. Cette définition correspond aux règles internationales et permet ainsi des comparaisons entre pays.
Les personnes sans résidence habituelle dans un autre pays sont comptées dans la population résidente de la France si elles s'y trouvent à la date de référence du calcul de cette population.

16 Le taux de couverture est estimé par la part des personnes recensées retrouvées dans la source fiscale ou les autres sources.

17 Voir définition (au sens du recensement de la population) : *https://www.insee.fr/fr/metadonnees/definition/c1134*.

▶ Tableau des sources utilisées dans le processus d'alimentation de Résil

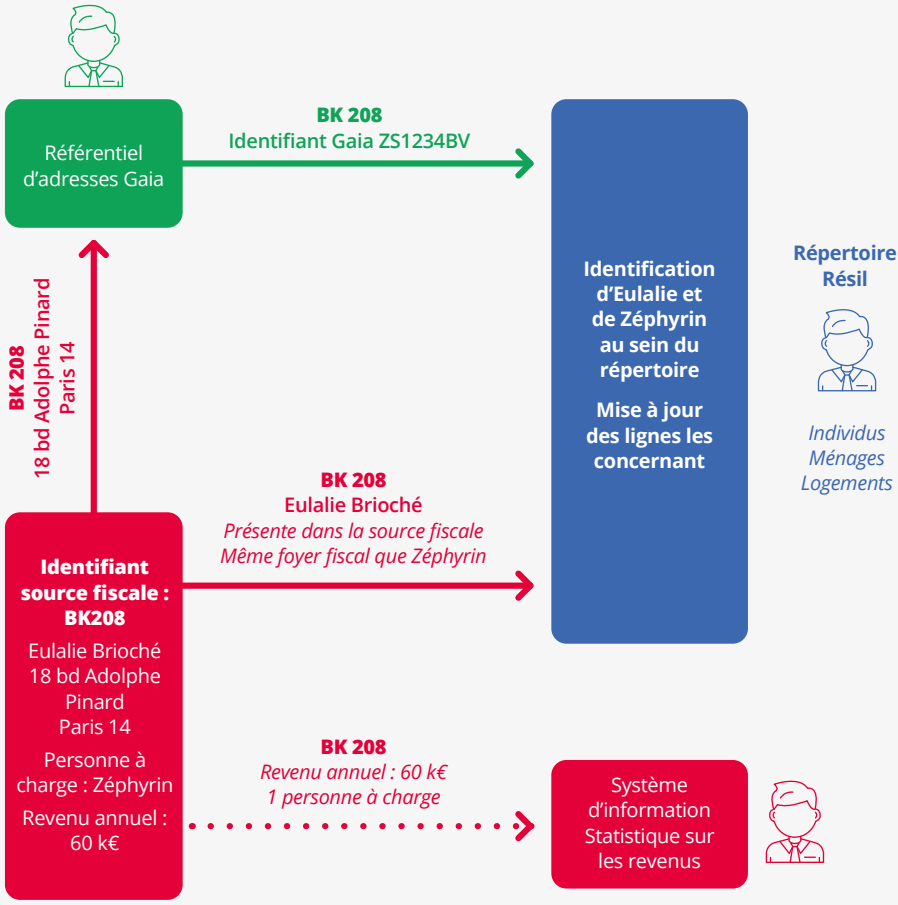
Source	Justification
Répertoire national d'identification des personnes physiques (RNIPP)	• Mise à jour de la liste des individus avec les naissances et décès. • Mise à jour des données d'état civil.
Fichiers fiscaux : • **Fichier d'imposition des personnes (Fip)** • **Fichier permanent des occurrences de traitement des émissions (Pote)** • **Fichier de mise à jour des informations cadastrales (Majic)** • **R-Loc (Référentiel des Locaux)** • **Fichier « Gérer mes biens immobiliers » (GMBI)**	• Connaissance du parc de logements (Référentiel des Locaux, relais progressif des données du Cadastre). • Connaissance de leur usage (résidence principale ou secondaire) et de leurs occupants (via les données de GMBI pour l'occupant déclaré, Fip et Pote pour les autres occupants du logement). • Sont retraitées les données identifiantes et localisantes, à l'exclusion de toute information sur les bases fiscales, revenus fiscaux, impôts dus ou payés.
Déclaration Sociale Nominative (DSN)	La fréquence mensuelle de cette source permet une mise à jour fiable du répertoire d'individus (durée de présence sur le territoire, prise en compte de jeunes actifs en complément des données fiscales, prise en compte plus précoce des changements d'adresse).
Prélèvement à la source pour les revenus autres (PASRAU)	La fréquence mensuelle de cette source permet une mise à jour fiable du répertoire d'individus (durée de présence sur le territoire, prise en compte de bénéficiaires de revenus autres que salariaux, en complément des données fiscales).
Fichier de référence des allocataires Caf (Caisses d'allocations familiales)	Cette source apporte des éléments sur la composition des foyers d'allocataires, éléments indispensables à une information fiable sur les ménages (notamment leur composition).
Fichier d'allocataires de la MSA (Mutualité sociale agricole)	• Cette source apporte des éléments sur la composition des foyers d'allocataires, éléments indispensables à une information fiable sur les ménages (notamment leur composition). • Le champ de la MSA est complémentaire de celui des Caf.
Fichier annuel d'inscriptions dans l'enseignement supérieur	Ce fichier permet d'améliorer la couverture du répertoire sur les jeunes de 17 à 25 ans, parfois absents des sources fiscales ou localisés de manière ambiguë.
Répertoire des communautés constitué pour la réalisation du recensement de la population	Les communautés constituent la résidence principale d'environ 1,3 million de résidents sur le territoire. Il est donc essentiel de compléter le parc de logements ordinaires par celui des communautés. Le répertoire des communautés est un intrant du répertoire des logements de Résil.
Enquêtes statistiques de contrôle de la qualité du répertoire	L'Insee pourrait être amené à réaliser des enquêtes statistiques de contrôle de la couverture du répertoire, voire des enquêtes de complétude sur des territoires mal couverts par les sources administratives.

> ***Pour chacune des sources, le choix des données retenues est sélectif.***

Pour chacune des sources (***voir Tableau des sources***), le choix des données retenues est sélectif. Les données seront sélectionnées et orientées en entrée du système d'information de l'Insee pour n'alimenter que les processus qui en ont besoin. Le dispositif d'accueil fonctionne comme une gare de triage. Lors de cette étape, les NIR présents dans les sources seront remplacés par les codes statistiques non signifiants (CSNS) correspondants : les données d'identité seront orientées vers Résil, soit pour mise à jour, soit pour garantir la qualité de l'identification. Les données d'adresses seront également traitées à la source pour les remplacer par un identifiant non signifiant provenant du répertoire d'adresses de l'Insee.

Les données « métier », accompagnées des identifiants non signifiants de personnes et d'adresses, seront intégrées dans les systèmes d'information pour produire des données statistiques (***figure 3***).

▶ Figure 3 - Alimentation de Résil : sélection et traitement des données



Résil s'inscrit dans un principe de minimisation des données traitées : il comportera très peu de variables et permettra de supprimer les données directement identifiantes des autres systèmes d'information de l'Insee pour les remplacer par des pseudonymes. Par ailleurs, Résil ne contiendra aucune donnée statistique permettant de caractériser les individus et logements, celles-ci étant traitées uniquement par les applications destinées à produire des données statistiques.

Pour alimenter le répertoire, un outil modernisé de mise à disposition des données administratives utilisées à des fins statistiques a été développé.

> ***L'arrivée de Résil est une opportunité pour rationaliser le dispositif d'accueil des données administratives.***

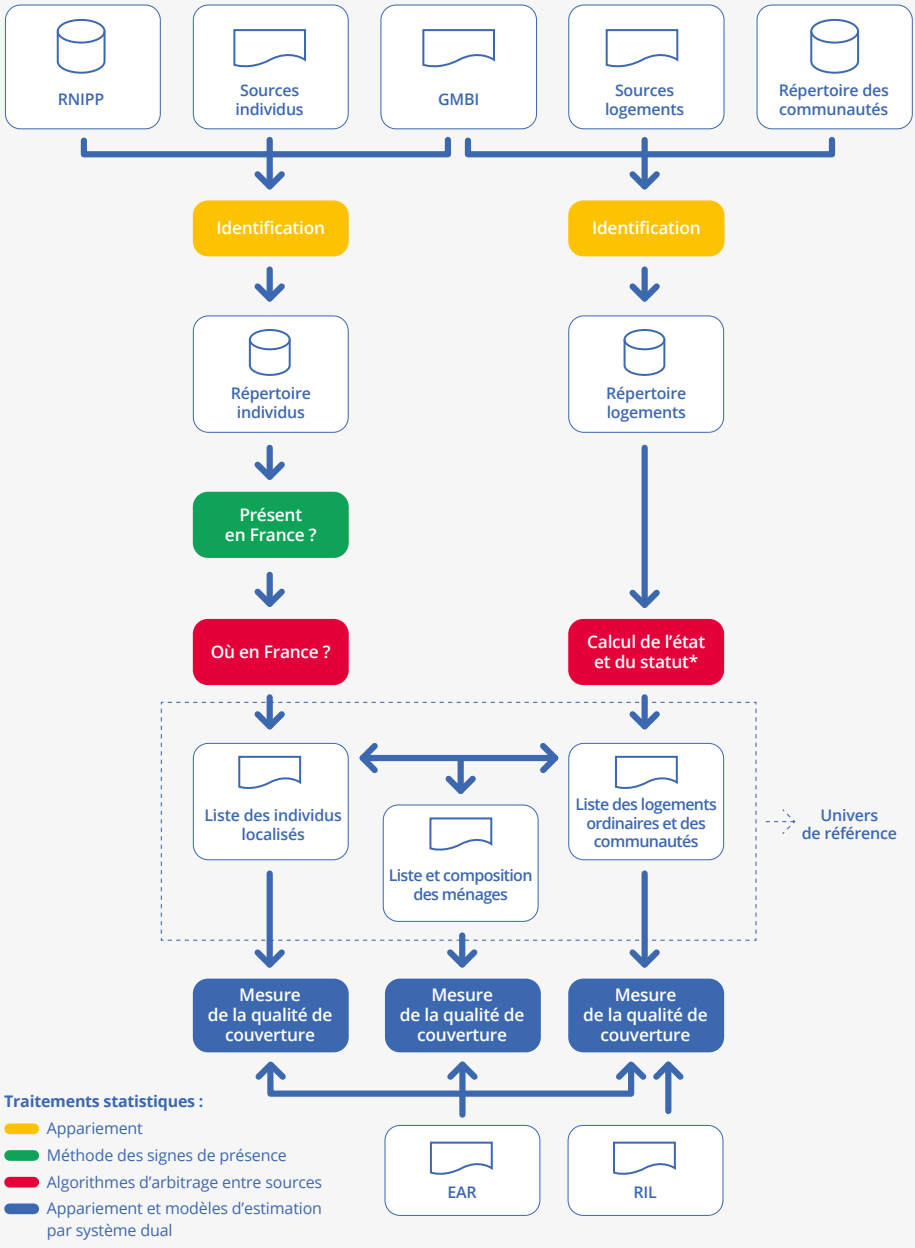
Pour la plupart des sources mobilisées, le répertoire Résil est un nouvel « utilisateur » en complément des producteurs de données statistiques sur l'emploi, les revenus ou le logement. Mais l'arrivée de Résil est une opportunité pour rationaliser le dispositif d'accueil des données administratives, constitué actuellement de plusieurs dispositifs juxtaposés, liés à la fois aux sources et aux usages. Outre le projet de construction des répertoires, le projet Résil s'accompagne d'un projet de modernisation et d'unification du dispositif d'accueil et de structuration des données administratives en données statistiques brutes facilement exploitables pour produire des chiffres relatifs aux unités statistiques (individus, logements et ménages). Un tel dispositif repose sur l'outil Accueil-Réception-Contrôle, dit ARC[18]. Le principe est d'utiliser le même outil pour l'accueil des différentes sources et leur mise à disposition vers les différents utilisateurs au sein du service statistique public. Cette rationalisation permet d'investir sur les performances et l'enrichissement fonctionnel d'un tel outil, sans sacrifier sa sécurité ni sa capacité d'adaptation rapide à des changements dans les sources, voire l'accueil de nouvelles sources. Compte tenu du nombre de sources à accueillir, de leur volume et de leur fréquence (au moins trois sources, parmi les plus volumineuses, sont mensuelles), mais aussi de la nécessité d'une alimentation rapide des systèmes de production d'indicateurs statistiques, la robustesse et les performances de traitement d'un tel outil sont cruciales. Sa capacité d'adaptation rapide aux transformations des sources administratives (dictées par les politiques qu'elles mettent en œuvre et non pas par les statistiques qu'elles permettent de produire) et à l'apparition de nouvelles sources est également essentielle.

Second pilier : pour assurer la qualité du répertoire, des traitements statistiques performants et innovants, inspirés par des pratiques internationales

Disposer des données est indispensable mais ne suffit pas. Il faut également des outils performants pour transformer ces données en un répertoire puis en un univers de référence (liste des individus effectivement résidents à la date du 1er janvier, liste des logements habitables, liste des ménages), qui soit d'une qualité propre à son utilisation statistique.

18 Voir l'article de Lefebvre, Soulier et Tortosa sur l'accueil des données administratives dans ce même numéro.

▶ Figure 4 - Des sources à l'univers de référence : une succession de traitements statistiques



** Calcul de l'état et du statut : il s'agit de déterminer si le logement est habité ou non, s'il s'agit d'une résidence principale, secondaire, ou d'un logement vacant.*

Il s'agit d'abord d'identifier les individus, pour « mettre à jour les bonnes lignes du fichier », sur la base d'informations parfois incomplètes ou entachées d'erreurs ; c'est tout l'enjeu des moteurs d'identification (***figure 4***). Résil s'appuiera sur un processus plus discriminant que celui utilisé pour le CSNS, dans la mesure où il pourra utiliser, pour les cas les plus douteux, des informations complémentaires telles que la composition du ménage ou l'adresse de la personne.

Ensuite, il faut distinguer, parmi les individus présents dans le répertoire, ceux qui résident effectivement sur le territoire national et pour chacun d'eux déterminer leur résidence principale. Cela permet de construire l'univers de référence.

La personne vit-elle toujours sur le territoire national ? La méthode des « signes de présence » (ou « signes de vie » dans la littérature académique et professionnelle) consiste à mobiliser, en complément des données d'état civil incontestables (naissances, décès), l'information relative à la présence des individus dans telle ou telle source administrative, avec une adresse située sur le territoire. Si un individu non décédé dans le RNIPP est absent dans toutes les sources administratives, il y a une forte probabilité qu'il ne réside plus sur le territoire. S'il ne figure que dans une partie des sources dans lesquelles on devrait normalement le retrouver et possède une adresse à l'étranger dans au moins une source, il y a une probabilité non négligeable qu'il ait quitté le territoire. Chacun des signes de présence dans les sources peut être pondéré par la qualité de l'identification de l'individu[19] et sa pertinence au regard des individus concernés. Par exemple, si on utilise le fichier des étudiants, il sera pertinent pour les 18-25 ans. Cette méthode est pratiquée dans plusieurs pays, tels que l'Estonie, l'Irlande, l'Italie ou l'Australie. Elle est encouragée par Eurostat[20], en lien avec le développement de l'usage des données administratives.

Des règles de décision en cas d'adresses multiples

Quand une personne a des adresses différentes d'un fichier à l'autre, il convient également de déterminer quelle est l'adresse de sa résidence principale. La divergence peut résulter d'un décalage dans la mise à jour des fichiers administratifs (la personne a déménagé mais l'information n'a pas encore été prise en compte) ou d'une multiple résidence (les « célibataires géographiques », les étudiants logés sur leur lieu d'études mais encore attachés au foyer fiscal de leurs parents, les enfants en garde alternée, etc.). Les règles de décision à retenir doivent permettre de localiser les personnes à leur résidence principale, selon les concepts prescrits pour les comparaisons internationales et mis en œuvre pour le recensement de la population.

Pour travailler sur des données stables, il est important d'avoir une photo donnant la situation au 1er janvier.

L'univers de référence provenant de Résil doit correspondre à une situation stable pour permettre de partager son utilisation. Or le répertoire est vivant, ses mises à jour sont régulières. Il faut donc « prendre une photo » du répertoire reflétant la situation à une date donnée, celle du 1er janvier par convention.

19 Si les traits d'identité de la source ne permettent pas d'identifier de façon sûre un individu du répertoire, on donnera un poids moins important au signe de présence dans cette source, car il pourrait s'agir d'une erreur d'appariement.
20 Eurostat est l'Office statistique de l'Union européenne.

> ***Pour travailler sur des données stables, il est important d'avoir une photo donnant la situation au 1er janvier.***

Résil sera mis à jour à partir de sources diverses qui n'arriveront pas toutes au même moment. Plutôt que d'attendre la dernière source pour développer la photo, il a été décidé de produire trois versions de chaque photo, au fur et à mesure de l'arrivée des données.

Par exemple, l'univers de référence provisoire au 1er janvier 2025 sera produit à l'été 2025, la version semi-définitive en janvier ou février 2026, la version définitive au début de l'été 2026.

Chacun des utilisateurs pourra donc procéder à son propre arbitrage entre fraîcheur et exhaustivité en fonction de ses utilisations de l'univers de référence.

Il est également nécessaire de mesurer la qualité de couverture, ce que tout répertoire doit pouvoir faire. Cela se fait en confrontant avec les enquêtes annuelles de recensement et en se fondant sur la méthode d'estimation par système dual.

Le principe de cette estimation est de confronter les deux collectes, Résil d'une part et l'enquête de recensement d'autre part, de décompter les personnes présentes dans les deux sources et celles présentes dans l'une ou l'autre des sources afin d'en déduire le nombre de personnes absentes des deux sources, donc la taille de la population totale, sous l'hypothèse notamment d'indépendance des deux collectes et d'absence de personnes comptées à tort dans chacune des deux collectes.

Cette méthode est utilisée dans plusieurs pays pour estimer la couverture des recensements exhaustifs. Elle se développe actuellement pour mesurer la couverture des répertoires, par exemple en Italie pour le répertoire rassemblant les registres municipaux de population.

Troisième pilier : un fondement juridique clair et solide, qui autorise le traitement et protège les données qui en sont issues

Résil s'inscrit dans le cadre juridique national et européen relatif à la production de statistiques publiques et à la protection des données individuelles (loi du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques, Règlement Général pour la Protection des Données (RGPD) et Loi Informatique et Libertés[21]). Aucune disposition de nature législative n'est nécessaire ; l'environnement cité ci-dessus garantit l'accès aux données nécessaires et les obligations relatives à leur protection.

L'Insee a considéré que le niveau juridique approprié pour fonder un traitement de cette nature est celui d'un décret en Conseil d'État. Un arrêté du responsable de traitement, en conformité avec le RGPD, aurait pu suffire s'il s'était agi d'un texte purement technique, mais les interrogations de nature plus politique ou sociétale que peut susciter le dispositif justifiaient un texte de ce niveau juridique, en offrant

21 Voir les références juridiques en fin d'article.

un examen critique préalable, par la Cnil[22] et par le Conseil d'État, renforçant ainsi la légitimité du dispositif.

Le décret en Conseil d'État créant Résil a été publié au Journal Officiel le 7 janvier 2024. Il comporte des dispositions prises dans la plupart des textes créant des traitements de données personnelles : la création du traitement et sa finalité, la liste des variables et leur durée de conservation, les utilisateurs du répertoire, les destinataires des données qu'il gère et des fichiers qu'il produira, et les dispositions relatives à la sécurité du système d'information. Il comprend également plusieurs dispositions plus spécifiques à Résil : la définition des appariements, les conditions d'évolution de la liste des sources, la référence expresse aux exigences déontologiques propres au métier de statisticien. Un arrêté[23] établissant la liste des sources utilisées pour construire et mettre à jour Résil, pris en application de ce décret, a également été publié au Journal Officiel le 7 janvier 2024.

Quatrième pilier : un mandat social à conforter en permanence

Tout ce qui précède vise à conférer à l'Insee la capacité, technique ou juridique, à construire et gérer le répertoire. Cette capacité doit s'accompagner d'une légitimité, ou d'un mandat social. Les corps constitués et autorités compétentes confèrent une partie de cette légitimité, à travers les textes qui encadrent le fonctionnement de la statistique publique et le traitement Résil en particulier. Cependant, cette légitimité ne serait pas suffisante sans un « mandat social », preuve que ce traitement est reconnu et accepté par la population, qui fait confiance à l'Insee pour le mettre en œuvre.

> ***L'Insee a souhaité associer des représentants de la société civile en amont du projet sous la forme d'une concertation approfondie avec différents acteurs.***

L'Insee a souhaité associer des représentants de la société civile en amont du projet sous la forme d'une concertation approfondie avec différents acteurs afin de tenir compte des points de vue exprimés dans la construction du répertoire statistique tout comme dans ses usages.

La concertation[24] menée en 2022 a reposé sur deux grandes actions : d'une part une rencontre[25] du Conseil national de l'information statistique (Cnis)[26] le 28 janvier 2022, d'autre part la création d'un groupe de concertation, placé sous l'égide du Cnis, qui a fonctionné de mai à septembre et dont le rapport[27] est publié sur son site.

22 Cnil : Commission nationale de l'informatique et des libertés.
23 Voir les références juridiques en fin d'article.
24 Voir l'article de Dupont, Dussart et Guillaumat-Tailliet sur les enjeux éthiques de Résil dans ce même numéro.
25 *https://www.cnis.fr/evenements/appariements-de-donnees-individuelles-entre-richesse-de-linformation-statistique-et-respect-de-la-vie-privee/.*
26 Le Conseil national de l'information statistique (Cnis) assure la concertation entre les producteurs et les utilisateurs de la statistique publique.
27 *https://www.cnis.fr/wp-content/uploads/2022/11/rapport-version-dfinitive.pdf.*

La rencontre a permis à l'Insee et à divers représentants de la statistique publique de présenter les pratiques, les usages et les techniques d'appariement ainsi que de montrer les apports pour la connaissance et l'action publique (mesurer l'insertion professionnelle des jeunes, étudier le devenir de bénéficiaires de minima sociaux, comprendre les écarts entre deux sources de données, etc.). Elle a permis l'expression d'interrogations, voire de craintes sur ce qu'un dispositif comme Résil pourrait permettre s'il était mal utilisé, mais elle a aussi fait émerger le souhait que l'effort de communication et de transparence mené à l'occasion de cette rencontre se poursuive par une concertation plus approfondie sur le projet.

Le groupe de concertation a permis de rassembler des expertises très variées (protection des libertés fondamentales, protection des données sur les plans juridique et informatique, transformation numérique, éthique, recherche, etc.) pour dresser la liste des interrogations suscitées par le projet Résil et échanger sur les réponses apportées par l'Insee. Ce groupe a conclu que le projet était légitime et conforme au principe de proportionnalité du traitement, à condition de ne pas utiliser certaines des sources initialement envisagées par l'Insee. Il a considéré également qu'il était nécessaire de bénéficier de regards extérieurs (Conseil d'État, Cnil, Cnis, Autorité de la Statistique Publique, Agence Nationale pour la Sécurité des Systèmes d'Information) lors de la construction et de l'utilisation de Résil. Il insiste également sur la nécessaire transparence sur le répertoire et ses usages, permettant un autre regard extérieur : celui des personnes concernées par le traitement.

L'Insee suit cette voie, tant au niveau du contenu du décret[28] que des modalités de construction du système d'information, mais aussi dans la communication sur le répertoire et ses usages. Une page consacrée à Résil[29] a ainsi été mise en place sur le site internet de l'Insee. Le Cnis jouera un rôle important, selon des modalités à construire, quant à la poursuite de la concertation sur Résil et sur les services qu'il rendra, notamment la construction de fichiers enrichis par appariements.

Dans les deux années qui viennent, les développements du système d'information seront finalisés, le répertoire sera initialisé et les traitements statistiques destinés à en assurer la qualité seront réalisés. Les premiers univers de référence (au 1er janvier 2025) seront produits progressivement entre mi-2025 (le provisoire) et mi-2026 (le définitif) et les premiers services d'appariements seront rendus début 2026.

28 *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000048866207.*
29 *https://www.insee.fr/fr/information/7748883.*

▶ Fondements juridiques

- 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 5 mars 2024]. 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 21 février 2024. [Consulté le 30 avril 2024]. 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]. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

- Décret n° 2024-12 du 5 janvier 2024 portant création d'un traitement automatisé de données à caractère personnel dénommé « répertoire statistique des individus et des logements » (Résil). In : *site de Légifrance*. [en ligne]. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000048866207*.

- Arrêté du 5 janvier 2024 pris en application de l'article 2 du décret n° 2024-12 du 5 janvier 2024 portant création d'un traitement automatisé de données à caractère personnel dénommé « répertoire statistique des individus et des logements » (Résil). In : *site de Légifrance*. [en ligne]. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000048866233*.

- Délibération n° 2023-080 du 20 juillet 2023 portant avis sur un projet de décret en Conseil d'État relatif à la création du traitement automatisé de données à caractère personnel permettant la gestion du répertoire statistique d'individus et de logements, et sur l'arrêté y afférent. In : *site de Légifrance*. [en ligne]. [Consulté le 4 avril 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000048934586?page=1&pageSize=100&searchField=ALL&searchType=ALL&sortValue=DATE_DECISION_DESC&tab_selection=cnil&timeInterval=01%2F07%2F2023+%3E+31%2F08%2F2023&typePagination=DEFAULT*.

▶ Bibliographie

- BAKKER, Bart F.M., VAN ROOIJEN, Johan, VAN TOOR, Leo, 2014. The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. In : *Statistical Journal of the IAOS*. Volume 30, n° 4, pp. 411-424. [en ligne]. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://www.cbs.nl/-/media/imported/documents/2016/53/system-of-social-statistical-datasets.pdf*.

- BÉNICHOU, Yves-Laurent, ESPINASSE, Lionel et GILLES, Séverine, 2023. Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 64-85. [Consulté le 5 mars 2024]. Disponible à l'adresse : https://www.insee.fr/fr/information/7635825?sommaire=7635842.

- BYCROFT, Christine, EATHERLEY, Clara, PAGE, Mathew et TA'ALA, Shane, 2022. A statistical person register in New Zealand: Progress and challenges. In : *Statistical Journal of the IAOS*. [en ligne]. 21 mars 2022. Volume 38, n° 1, pp. 225-230. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210922*.

- DEVILLE, Jean-Claude et SÄRNDAL, Carl-Erik, 1992, Calibration Estimators in Survey Sampling. In : *Journal of the American Statistical Association*. [en ligne]. Juin 1992. Vol. 87, N° 418, pp. 376-382. [Consulté le 23 mai 2024]. Disponible à l'adresse : *https://www.jstor.org/stable/2290268*.

- DEMOTES-MAINARD Magali, 2019. Elire, un projet ambitieux au service du Répertoire électoral unique. In : *Courrier des statistiques*. [en ligne]. Juin 2019. Insee. N° N2, pp. 58-71. [Consulté le 23 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168399?sommaire=4168411*.

- DUPONT, Françoise, 2023. Quels types de sources l'Insee utilise-t-il pour construire ses statistiques ? In : *Le blog de l'Insee*. [en ligne]. 16 mai 2023. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://blog.insee.fr/quels-types-de-sources-l-insee-utilise-t-il/*.

- DUPONT, Françoise, 2023. Les appariements de données de la statistique publique : des analyses enrichies, un cadre juridique protecteur. In : *Le blog de l'Insee*. [en ligne]. 1er septembre 2023. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://blog.insee.fr/appariements-de-donnees-de-la-statistique-publique/*.

- 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]. 29 novembre 2022. Insee. N° N8, pp. 72-92. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665188?sommaire=6665196*.

- HACHID, Ali et LECLAIR, Marie, 2022. Sirus, le répertoire d'entreprises au service du statisticien. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 115-130. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665192?sommaire=6665196*.

- 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 5 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- RIVIÈRE, Pascal, 2022. Qu'est-ce qu'un répertoire ? De multiples exigences pour un système complexe. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 52-71. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665186?sommaire=6665196*

- ZHANG, Li-Chun et DUNNE, John, 2017. Trimmed Dual System Estimation. In : *Capture-Recapture Methods for the Social and Medical Sciences*, pp. 239-259. Éditions Chapman and Hall/CRC. ISBN 978-1-49-874531-4.

La concertation : une étape essentielle pour le projet Résil

Françoise Dupont[*], *Josy Dussart*[**] *et François Guillaumat-Tailliet*[***]

Construire un répertoire statistique des individus et des logements (Résil) est un défi qui présente des enjeux techniques, juridiques, mais aussi éthiques. Les enjeux éthiques s'inscrivent dans un contexte où croissent à la fois le besoin d'informations statistiques fiables et détaillées, la circulation des données personnelles, et la vigilance quant à la bonne utilisation de ces dernières.

La concertation engagée par l'Insee a permis de partager la réflexion sur ce que pouvait être Résil et comment il devait être utilisé, mais aussi comment l'expliquer au plus grand nombre. Menée sous l'égide du Conseil national de l'information statistique (Cnis), cette démarche exigeante a mobilisé des compétences diverses, pour la plupart éloignées du monde de la statistique, pour prendre en compte les considérations éthiques (de protection des libertés publiques, de transparence, etc.) et parvenir à une évaluation partagée des principes de nécessité, minimisation et proportionnalité.

L'Insee a intégré la plupart des recommandations du groupe de concertation dans sa réflexion. Il les a traduites en dispositions juridiques, techniques (dans la conception et le contenu du répertoire), organisationnelles et dans la communication menée autour du répertoire. La concertation ne s'arrête pas là : la démarche de communication et d'écoute doit se poursuivre et un enjeu spécifique reste à traiter concernant l'appréciation des appariements réalisés grâce à Résil au regard des principes de nécessité et de proportionnalité.

Setting up a Statistical Register of Individuals and Dwellings (Résil) is a challenge that involves technical, legal and ethical issues. The ethical issues arise in a context where the need for reliable and detailed statistical information, the circulation of personal data and vigilance over the proper use of such data are all on the increase.

The consultation process initiated by INSEE provided an opportunity to share ideas on what Résil could be and how it should be used, as well as how to explain it to the general public. Conducted under the aegis of the National Council for Statistical Information (CNIS, Conseil national de l'information statistique) this demanding process has mobilised a wide range of skills, most of them outside the world of statistics, to consider ethical considerations (protection of civil liberties, transparency, etc.) and to achieve a shared assessment of the principles of necessity, minimisation and proportionality.

INSEE has incorporated most of the consultation group's recommendations into its reflections. It has translated them into legal, technical (in the design and content of the register) and organisational provisions, and in the communication surrounding the register. The consultation does not end there: the communication and listening process must continue, and a specific issue remains to be addressed concerning the assessment of the matchings made possible by Résil with respect to the principles of necessity and proportionality.

* Au moment de la rédaction de cet article, chargée de mission « Questions juridiques et communication » sur le projet Résil.

** Chargée de mission « Questions juridiques et communication » sur le projet Résil.
josy.dussart@insee.fr

*** Secrétaire général adjoint du Cnis.
francois.guillaumat-tailliet@insee.fr

L'Insee s'est engagé dans la construction d'un répertoire statistique des individus et des logements (Résil)[1] à l'horizon 2025. Ce projet s'inscrit dans un contexte national et international marqué à la fois par :

- une forte attente de statistiques détaillées au regard de l'accroissement du volume des données disponibles, et des innovations en matière d'exploitations statistiques pour des finalités de préparation et d'évaluation de politiques publiques et d'alimentation du débat public. Ce besoin se manifeste en particulier par une demande de plus de réactivité dans la fourniture des statistiques[2] dont le Conseil national de l'information statistique (Cnis)[3] se fait régulièrement l'écho.

- une attention croissante des citoyens et de leurs représentants aux enjeux de protection des données individuelles, face à l'accroissement du volume de ces données et des capacités de traitement. Elle se traduit du côté des producteurs de données par le renforcement des règles relatives à la transparence sur le traitement des données, permettant aux citoyens d'avoir des garanties sur leur bon usage.

> ***L'Insee a considéré que les enjeux juridiques et éthiques étaient à la hauteur des enjeux techniques.***

Le projet de répertoire Résil intègre ces deux attentes. Dès le début du projet, l'Insee a considéré que les enjeux juridiques et éthiques étaient à la hauteur des enjeux techniques et nécessitaient une attention particulière compte tenu du caractère exhaustif du répertoire et de l'utilisation de plusieurs sources appariées pour créer ce répertoire.

Aussi, l'Insee a souhaité associer à la réflexion des représentants de la société, au-delà de ses interlocuteurs habituels, pour confronter son point de vue à d'autres approches et ce, très en amont du projet. Il l'a fait grâce à un groupe de concertation qui s'est réuni de mai à septembre 2022 sous l'égide du Cnis. Cette démarche originale a permis de rassembler des spécialistes sur des sujets très variés (protection des libertés fondamentales, protection des données sur les plans légal et informatique, transformation numérique, éthique, recherche socioéconomique, etc.) pour dresser la liste des interrogations que suscite le projet Résil, échanger sur les réponses apportées par l'Insee à ces interrogations, proposer des modalités concrètes pour poursuivre les réflexions tout au long de la construction du projet jusqu'en 2025 et au-delà. Cette expérience a été riche d'enseignements pour la conduite du projet Résil et plus largement pour la statistique publique.

1 Voir l'article d'Olivier Lefebvre sur Résil dans ce même numéro et plus généralement le dossier dont le présent article fait partie.
2 Il s'agit notamment d'une demande récurrente d'Eurostat (*https://www.insee.fr/fr/metadonnees/definition/c1292*). Lors de la pandémie de Covid-19, l'intérêt de fournir rapidement des informations a été mis en évidence.
3 Voir l'article du Courrier des statistiques N6 : « Le Conseil national de l'information statistique : la qualité des statistiques passe aussi par la concertation », Isabelle Anxionnaz et Françoise Maurel.

▶ Des enjeux éthiques méritant une réflexion partagée

Les données personnelles : une sensibilité croissante au fil des années, qui diffère selon les contextes culturels des pays

Le sujet des données personnelles est un sujet croissant de préoccupation dans un contexte de transformation numérique de la société, avec une masse de données potentiellement mobilisables en hausse continue alors que les outils d'analyse sont de plus en plus puissants, sous l'effet des progrès techniques de stockage et de calcul (Chaire Valeurs et politiques des informations personnelles, 2019). De nouvelles opportunités sont offertes et une forte demande pour réguler ces nouveaux usages émerge dans la société pour profiter des bénéfices tout en limitant les risques liés à leur mauvais usage.

Depuis une dizaine d'années, l'encadrement des nouveaux outils et des nouveaux usages fait l'objet de très nombreuses discussions sous l'angle de la protection des données et celui de la protection des libertés fondamentales aux niveaux national, européen et international dans un champ très vaste, beaucoup plus large que celui qui s'applique aux travaux du service statistique public.

Ainsi, par exemple, l'entrée en vigueur en 2018 du Règlement général sur la protection des données (RGPD, CNIL, 2016) qui concerne toutes les données personnelles a constitué un pas important vers de nouveaux droits pour les personnes et a induit une responsabilisation des pilotes de traitements de données assortie de nouvelles obligations ainsi que la mise en place d'une instance de coordination européenne : le Comité européen de la protection des données[4].

Face à ces enjeux, la statistique publique s'appuie sur un cadre juridique et déontologique à la fois exigeant et protecteur

La statistique publique s'appuie depuis longtemps, sur un cadre légal au niveau national[5], avec une loi spécifique datant de 1951 complétée par la loi générale sur la protection des données personnelles de 1978. Ce cadre est régulièrement mis à jour pour tenir compte de l'évolution des besoins, des opportunités et des pratiques, il est également mis en cohérence avec le droit européen : le règlement 223 sur la production de statistiques européennes et le règlement général sur la protection des données[6].

> ***Un cadre éthique s'est développé au fil des années au niveau international.***

En parallèle, s'est développé au fil des années au niveau international un cadre éthique :

- les Principes fondamentaux de la statistique officielle[7] adoptés en 1994 au niveau de l'ONU ;
- la Déclaration d'éthique professionnelle de l'Institut international de statistique[8] qui vaut pour les

4 Le comité européen de la protection des données est une instance de coordination des autorités de protection des données comme la CNIL des 27 États membres. *https://www.edpb.europa.eu/about-edpb_fr.*
5 Voir *https://www.insee.fr/fr/information/1300616* et les références juridiques en fin d'article.
6 Voir les références juridiques en fin d'article.
7 *https://unstats.un.org/unsd/dnss/hb/F-fundamental%20principles_A4-WEB.pdf.*
8 *https://isi-web.org/sites/default/files/2023-08/isi-declaration-on-professional-ethics_french.pdf.*

statisticiens privés comme les statisticiens publics a été adoptée en 1985 et révisée en 2023 ;

- le Code de bonnes pratiques de la statistique européenne[9] a été adopté en 2005.

Plus récemment, l'UNECE a adopté six valeurs essentielles pour la statistique publique, ainsi qu'un corpus de bonnes pratiques pour les mettre en œuvre concrètement. Ces valeurs inspirent le cadre juridique du service statistique public français (***figure 1***).

Figure 1 - Les six valeurs adoptées par la conférence des statisticiens européens de l'UNECE

Source : UNECE : United Nations Economic Commission for Europe. *La Commission Économique des Nations Unies pour l'Europe a été mise en place en 1947 par le Conseil économique et social des Nations Unies. C'est l'une des cinq commissions régionales des Nations Unies. Elle se compose de 56 États membres : les pays européens, ainsi que les États-Unis, le Canada, Israël, la Turquie, et les ex républiques soviétiques du Caucase et d'Asie centrale.*

L'enjeu de la confiance

Les enjeux liés à la protection des données personnelles sont de plus en plus importants et la confiance que placent les citoyens dans les institutions s'érode, dans un contexte de défiance croissante vis-à-vis des institutions ou de toutes formes d'expertise (Agacinski, 2018 ; Rouban, 2022). L'argument d'autorité ne suffit plus : il faut expliquer, écouter et s'adapter si nécessaire.

Un besoin de transparence sur l'utilisation de donnée administratives à des fins statistiques...

Contrairement à la situation d'enquête où le consentement éclairé est recueilli, lorsqu'on utilise des données administratives, qui sont désormais la base d'une partie importante des

9 *https://www.insee.fr/fr/information/4140105.*

> ***L'information sur tous les traitements de la statistique publique est actuellement assurée conformément aux principes éthiques de la profession de statisticien public et sur la base légale du Règlement général sur la protection des données.***

statistiques publiques, la personne qui fournit des informations personnelles ne donne son consentement explicite que pour la finalité première (déclarer ses revenus, percevoir une prestation, s'inscrire dans l'enseignement supérieur, s'abonner auprès d'un distributeur d'énergie, etc.) et n'a pas conscience des traitements ultérieurs de réutilisation et encore moins des appariements de données. L'information sur tous les traitements de la statistique publique est actuellement assurée conformément aux principes éthiques de la profession de statisticien public et sur la base légale du Règlement général sur la protection des données (RGPD, CNIL, 2016) par une information publique sur l'utilisation des données, dans la rubrique sur la protection des données des sites internet du responsable de traitement[10].

Le Cnis joue également un rôle en faveur de cette transparence, avec en particulier l'examen des programmes de travail de la statistique publique, laquelle développe de plus en plus les enjeux liés aux nouvelles sources de données et aux appariements. Sur ces sujets complexes, les producteurs de statistiques publiques et le Cnis augmentent progressivement la visibilité de ces travaux dans une forme aussi accessible que possible au grand public.

… mais aussi une attention croissante au principe de nécessité et de proportionnalité

Deux principes ont pris plus d'importance récemment dans le dialogue entre les statisticiens et la société civile : la nécessité et la proportionnalité. Ils expriment une demande d'attention croissante sur une utilisation parcimonieuse des données. Il s'agit de vérifier dans un premier temps la **nécessité** de recourir à ces données eu égard à l'objectif poursuivi, puis dans un deuxième temps, que les données sélectionnées et les traitements requis sont bien **proportionnés** par rapport aux finalités poursuivies au titre de l'intérêt général.

Les collègues canadiens ont donné récemment une visibilité particulière à ces deux principes en mettant en place en 2018 un cadre spécifique portant sur la nécessité et la proportionnalité qui s'applique à tous les traitements de données (Principes de nécessité et de proportionnalité de Statistique Canada, 2019 ; Rancourt, 2019). Cette initiative fait suite à des discussions dans le cadre de la modification de la loi statistique canadienne en 2017 et des interpellations de la société et d'élus sur des projets de traitements de données bancaires. Chez nos collègues canadiens, des efforts constants et croissants sont réalisés depuis pour communiquer de façon très pédagogique sur tous les travaux de la statistique publique et sur l'attention accordée à la nécessité et à la proportionnalité des travaux. L'utilisation des données administratives à des fins statistiques et des appariements de données sont bien documentés sur le site de Statistique Canada[11].

10 *https://www.insee.fr/fr/information/3719162* pour l'Insee.
11 Données administratives *https://www.statcan.gc.ca/fr/nos-donnees?MM=1* ; appariements *https://www.statcan.gc.ca/fr/nos-donnees/ou/microdonnees* et *https://www.statcan.gc.ca/fr/enregistrement/somm*.

En France, une telle démarche d'examen de nécessité et de proportionnalité n'est pas nouvelle. Elle est pratiquée sous l'égide du Cnis et du Comité du label de la statistique publique[12] depuis 1994 pour toutes les enquêtes de la statistique publique. Il s'agit de vérifier que la charge de réponse pour les enquêtés n'est pas excessive et que les informations demandées ne sont pas trop intrusives au regard des finalités. Des personnes extérieures à la statistique publique sont sollicitées en fonction des thèmes (syndicats professionnels et de salariés, CNIL, chercheurs). Cette démarche, qui est installée pour les enquêtes, reste à développer en ce qui concerne les sources administratives et les appariements de données.

Dans ce cadre, une démarche de concertation s'est imposée pour le projet Résil

> ***Il a été décidé d'engager une démarche de concertation qui soit à la fois élargie, par la variété des compétences mobilisées, et approfondie grâce au temps consacré aux échanges.***

Concernant le projet Résil, l'Insee a eu très tôt la conviction que ces différents enjeux méritaient une attention particulière, au regard du caractère exhaustif de Résil et des appariements qu'il allait faciliter – critères qui requièrent une analyse d'impact au sens du RGPD – , mais surtout qu'il était essentiel de ne pas y réfléchir seul. Il fallait ainsi lancer la réflexion très en amont du projet, pour s'assurer de sa légitimité, examiner à quelles conditions les principes de nécessité et de proportionnalité pouvaient être respectés, et enrichir l'analyse de risques associés à ce projet. Il a été décidé d'engager une démarche de concertation (***encadré 1***) qui soit à la fois élargie, par la variété des compétences mobilisées, et approfondie grâce au temps consacré aux échanges.

Comment organiser la concertation sur le projet Résil et ses usages ?

Compte tenu du caractère très technique du projet, la solution de la plateforme en ligne a été exclue. Après des échanges dans le cadre du Cnis le 28 janvier 2022 et des échanges avec la CNIL, il semblait également difficile de se tourner vers un panel citoyen. En effet, pour évaluer le projet Résil, il faut pouvoir, le temps de la concertation, s'approprier suffisamment les enjeux des travaux d'appariement et d'analyse des bases de données de la statistique publique, leur contexte juridique et éthique pour être à même d'évaluer les risques et les bénéfices et ainsi intervenir dans les discussions et les alternatives possibles. Les échanges sur le projet ont permis de montrer que le coût d'entrée était trop élevé et le sujet trop abstrait, sans impact direct et facile à s'approprier par des citoyens, pour mettre à niveau un panel de citoyens qui n'aurait pas déjà réfléchi à ces sujets (contrairement aux sujets généralement présentés dans une concertation).

12 Le Comité du label de la statistique publique évalue les modalités de mise en œuvre des enquêtes, notamment en prenant en compte la qualité statistique du projet, la charge qu'implique l'enquête pour les personnes physiques ou morales, le degré de concertation avec les utilisateurs et le respect des termes de l'avis d'opportunité délivré par le Cnis. Il intervient aussi à la demande de l'Autorité de la statistique publique (ASP), lors de la labellisation ou de la reconnaissance de la qualification de statistiques d'intérêt général de statistiques administratives, produites par des organismes chargés d'une mission de service public ne faisant pas partie du Service statistique public (SSP).

▶ Encadré 1. Qu'est-ce qu'une concertation ?

Dans un de ses avis, le Cese* (Conseil économique, social et environnemental), définit la concertation comme « un dialogue structuré autour d'un projet entre parties prenantes, visant à sa réalisation dans les meilleures conditions possibles à partir de la prise en compte des différents points de vue. Elle peut être envisagée comme une aspiration à trouver collectivement des formes d'intérêt commun ». Il indique également que « Cette démarche est associée à un processus de prise de décision clair et bien identifié ». Il ajoute : « On entend par parties prenantes, d'une part les personnes, d'autre part tout groupe ou organisation directement ou indirectement concerné ou affecté par les activités, les objectifs du projet ».

La concertation est différente de la consultation et de la négociation. La consultation est un recueil d'avis ponctuel auprès des parties prenantes sur un projet ou une question qui permet d'alimenter un processus de décision avec une seule interaction avec les parties prenantes. La négociation est un échange visant à trouver un accord à partir de positions différentes.

La concertation vise à travailler collectivement avec des acteurs représentant une grande diversité de points de vue pour trouver ensemble une réponse à un problème ou pour aboutir à l'enrichissement d'un projet. Elle s'appuie sur la confiance en la sincérité des échanges et en la prise en compte future des conclusions de la concertation dans les décisions. L'objectif n'est pas forcément de parvenir à un consensus, mais que les décisions soient les plus pertinentes possibles, en intégrant le maximum de points de vue. Cet exercice de démocratie participative répond à une attente forte des citoyens et plus généralement de l'ensemble de la société (Dictionnaire critique et interdisciplinaire de la participation).

Elle se pratique selon des formes diverses, en fonction des sujets et des attentes des porteurs de projet. Dans tous les cas, l'organisation concrète des interventions d'experts, la fourniture de supports écrits et/ou vidéos qui permettent d'avoir une information éclairée et équilibrée, représentent un défi pour assurer en même temps la précision, la concision et l'accessibilité (intelligibilité) pour tous.

Quelques formes de concertation

La consultation en ligne vise à mettre à disposition de tous, sur une plateforme dédiée, les informations sur un projet, puis à recueillir les réactions publiques des citoyens, entreprises, organisations représentatives et associations, via un site internet. Le bilan est rendu public et le projet intègre autant que possible les demandes qui ont été exprimées. Recourir à cette méthode suppose que le projet se prête à une présentation très synthétique, autoportée et qu'il soit possible d'assurer à la démarche une publicité suffisante et non biaisée pour assurer la représentativité de toutes les parties prenantes concernées dans les réponses.

Le panel citoyen est un dispositif qui rassemble une trentaine de citoyennes et citoyens bénévoles tirée au sort en assurant des critères de représentativité, durant plusieurs sessions de travail réparties sur une ou plusieurs semaines. L' objectif est qu'ils répondent à une question précise, grâce à une information éclairée par de nombreux intervenants représentant les différents points de vue sur le sujet traité et aux délibérations menées tout au long des travaux.

La convention citoyenne est basée sur le même principe et diffère par l'ampleur du dispositif qui est beaucoup plus large et adapté à une question plus générale comme la fin de vie ou la transition écologique (rassemblant plutôt 150 à 200 personnes et impliquant un travail sur plusieurs mois). Une des difficultés est de trouver les personnes disponibles et disposées à prendre du temps, souvent le week-end ou le soir, pour s'approprier le sujet et donner leur avis soit en présentiel soit à distance avec une petite compensation financière.

Le panel d'experts est adapté à des sujets très techniques : il est composé d'experts bénévoles ou non. Son rôle est de croiser les regards de différentes expertises, de synthétiser des contributions diverses et de les partager.

Le focus group est une discussion dirigée par un animateur au sein d'un petit groupe de personnes représentatif de la société ou d'une sous-partie (une douzaine de personnes maximum qui reçoivent une petite compensation financière). Le focus group peut permettre de tester un concept en amont d'une démarche « concertative » plus large à laquelle il ne peut se substituer.

Il existe également d'autres outils (CNDP**, 2023).

* *La loi organique du 15 janvier 2021 fait du Cese le carrefour de la participation citoyenne pour nourrir ses travaux sous plusieurs formes : l'intégration de citoyennes et citoyens tirés au sort à ses formations de travail, les consultations en ligne, les pétitions citoyennes (le Cese peut être saisi par voie de pétition, dès lors qu'une pétition qui lui est adressée sur la plateforme https://petitions.lecese.fr/ recueille 150 000 signatures). Le Cese peut également s'auto-saisir d'une thématique soulevée par une pétition ou encore recevoir les auteurs de la pétition dans le cadre de ses travaux.*

** *Commission nationale du débat public.*

Après des discussions au sein du Cnis, en s'inspirant de l'expérience acquise par les praticiens de la concertation, le choix a été fait de réunir un groupe de spécialistes aux compétences diverses pour évaluer toutes les facettes du projet, ses opportunités et ses risques, et la proportionnalité de la solution technique en cours de conception du côté de l'Insee. Les domaines couverts étaient vastes : libertés publiques, protection des données, usages des données pour les sciences sociales, éthique, communication, usage des outils numériques, etc. L'Insee a estimé que des spécialistes seraient plus à même de se saisir du sujet et de formuler des critiques précises pour l'aider à faire évoluer le projet vers une solution équilibrée et respectueuse des libertés fondamentales, qui permet d'éloigner les craintes que peut susciter ce projet de répertoire. Ces spécialistes ont l'habitude de porter les préoccupations des citoyens dans des débats, ils peuvent plus facilement apporter la contradiction si nécessaire. Le caractère abstrait du sujet et les efforts nécessaires pour se l'approprier ont été confirmés par le groupe qui a fourni un travail d'appropriation remarquable à partir des interventions et supports fournis, rendus au préalable aussi simples et accessibles que possible. Le groupe a réuni dix-huit personnes, lors de sept réunions et dix auditions.

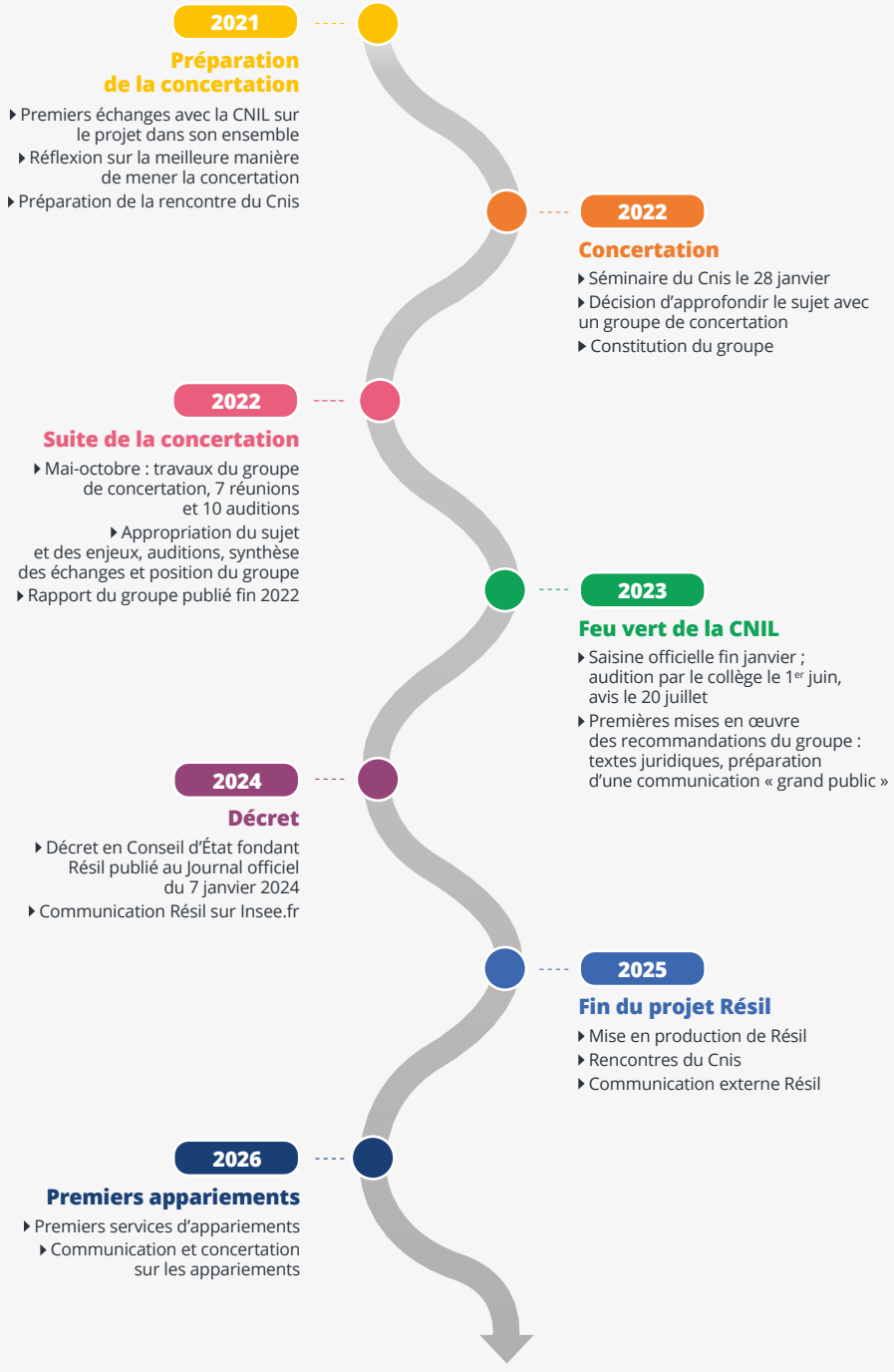
▶ Le déroulement de la concertation

Le groupe de concertation sur le projet de **Répertoire Statistique des Individus et des Logements** (***figure 2***), placé sous l'égide du Cnis, s'inscrit dans le prolongement de la rencontre du Cnis du 28 janvier 2022 (Rencontres du Cnis, 2022) sur les appariements. En effet, celle-ci avait permis de présenter pour la première fois le projet Résil et ses finalités ainsi que le contexte statistique national et international en matière d'appariements de données personnelles (Bénichou et alii, 2023 ; Dupont et alii, 2023) afin de recueillir les points de vue des représentants de la société dans sa diversité (administration publique, chercheurs, associations, syndicats, collectivités locales) en amont du projet. Des intervenants de la table ronde sur la transparence et l'information du public, qui découvraient le projet Résil et le cadre légal et déontologique des travaux de la statistique publique (Anxionnaz et alii, 2021 ; Bureau, 2020 ; Christine et Roth, 2020 ; Redor, 2023), avaient souligné que cette réflexion sur Résil et les appariements devait à leurs yeux être approfondie sur une plus longue période avec un panel d'expertise adapté aux enjeux éthiques et juridiques.

La mission du groupe de concertation était de dresser la liste des interrogations que suscite le projet Résil, échanger sur les réponses apportées par l'Insee à ces interrogations, proposer des modalités concrètes pour poursuivre les réflexions tout au long de la construction du projet jusqu'en 2025 et, le cas échéant, au-delà. Le Bureau du Cnis du 12 mars 2022 en a approuvé la création et le mandat.

Le Cnis a proposé à Jean-Marie Delarue, conseiller d'état honoraire, de présider ce groupe de concertation compte tenu de son expertise et de son intérêt pour le droit des libertés publiques et le respect de la vie privée. Outre son président, le groupe était constitué du maître d'ouvrage de Résil et de quatorze personnes (***encadré 2***) ayant une spécialité susceptible d'exprimer les interrogations de la société sur le projet Résil (protection des libertés fondamentales, protection des données sur le plan juridique et informatique, travaux statistiques, travaux de la recherche économique et sociale, transformation numérique, communication auprès du grand public sur les données, éthique).

Figure 2 - Calendrier de la concertation et de l'encadrement juridique de Résil



▶ Encadré 2. Les participants au groupe de concertation

Président :
Jean-Marie Delarue
(Conseiller d'État honoraire)

Rapporteuse :
Françoise Dupont (Insee)

Accompagnement du groupe :
- **François Guillaumat-Tailliet** (secrétaire général adjoint du Cnis)
- **Cédric Afsa** (inspecteur général honoraire de l'Insee)

Membres du groupe :
- **Maryse Artiguelong**, Ligue des droits de l'Homme
- **Jennyfer Chrétien**, Renaissance numérique
- **Valérie Gayte**, CFE-CGC (Confédération française de l'encadrement - Confédération générale des cadres), Cese
- **Mark Hunyadi**, Université de Louvain, philosophe
- **Alexandre Léchenet**, data journaliste, la Gazette des communes
- **Olivier Lefebvre**, Insee
- **Georges Louis**, CFE-CGC
- **Michelle Meunier**, sénatrice de Loire-Atlantique
- **Benjamin Nguyen**, INSA (Institut national des sciences appliquées) Centre - Val de Loire, Laboratoire d'Informatique Fondamentale d'Orléans
- **Emmanuelle Roux**, Chaudron numérique
- **Marion Selz**, Société française de statistique
- **Roxane Silberman**, Centre d'accès sécurisé aux données (CASD)
- **Bernard Sujobert**, CGT (Confédération générale du travail)
- **Cécile de Terwangne**, Université de Namur, juriste
- **Youssr Youssef**, data journaliste, le Figaro

Le champ, le dispositif, les acteurs

Le choix du groupe de spécialistes placé sous l'égide du Cnis s'est imposé assez vite. Les conditions matérielles et financières ont été bien cadrées au départ. Le président du groupe, reconnu pour son expertise et son indépendance, a joué un rôle fondamental dans la crédibilité de la démarche.

Une partie importante du travail a consisté à repérer les expertises existantes non encore familières du Cnis et de l'Insee qui seraient utiles pour constituer le groupe et disponibles pour assister à une concertation gratuitement. Or les spécialistes (issus du monde de la recherche, d'une entreprise, d'une institution publique ou d'une association) sont déjà fortement sollicités et les moyens des associations sont tout particulièrement limités. La participation au groupe représente un investissement substantiel en termes de temps pour s'approprier le sujet et participer activement aux échanges. Il est donc difficile de trouver des personnes qui sont disponibles pour plusieurs demi-journées de réunions et la participation à titre gracieux n'est pas non plus une évidence. Rappelons qu'il est d'usage dans les panels citoyens de couvrir les frais de transports et donner une petite rétribution (certes très modeste) pour la participation. Pour cette concertation sous l'égide du Cnis, la règle habituelle de la gratuité (hors frais de transport) pour la participation a été appliquée.

> ***La concertation portait sur le répertoire Résil, mais elle s'est naturellement emparée des appariements que Résil permet.***

Pour réunir les meilleures conditions du débat, il a fallu clarifier le mandat (ce à quoi s'engagent les participants et ce à quoi ils ne s'engagent pas), les marges de manœuvre dans le projet (ce qui peut être remis en question et ce qui est déjà décidé ou fortement contraint). Ceci a été abordé en amont lors du recrutement puis complètement explicité lors de la première séance. Le périmètre, le champ

qui est l'objet de la discussion doit aussi être clair pour tous pour ensuite lister les aspects qu'il faudra absolument éclairer pour une discussion sans angle mort. Dans le cas présent, la concertation portait sur le répertoire Résil, mais elle s'est naturellement emparée des appariements que Résil permet. Les questions à éclairer ont été assez facilement déduites de la rencontre du Cnis du 28 janvier 2022.

Des auditions ont permis d'élargir le périmètre de la réflexion en associant des interlocuteurs variés, français comme étrangers :

- un juriste spécialiste des libertés fondamentales ;
- la directrice d'un institut anglais spécialisé dans les démarches de concertation avec les citoyens sur les sujets numériques (Peppin, 2022) ;
- un expert en matière de médiation avec le grand public ;
- l'Anssi (Agence nationale de la sécurité des systèmes d'information) pour évoquer les risques en matière de cybersécurité ;
- un statisticien connaissant le projet Safari (Poulain, 2022 ; Espinasse et Roux, 2022) permettant ainsi de commenter les différences majeures avec le projet Résil ;
- un chercheur en économie pour évoquer les besoins d'informations et comparer à l'international les possibilités en matière d'appariement pour ce public ;
- Eurostat pour les besoins statistiques et les pratiques des autres pays ;
- un représentant de Statistique Canada pour évoquer la démarche canadienne d'analyse de la nécessité et la proportionnalité ;
- un représentant du Cese mais son audition n'a pas pu se tenir.

Les collègues étrangers (Eurostat, Statistique Canada) ont expliqué que les préoccupations étaient largement partagées, et que chaque pays avait mis en œuvre des modalités qui paraissaient appropriées, compte tenu du contexte institutionnel ou culturel propre à chacun.

Un état d'esprit ouvert dans les échanges afin d'accueillir les critiques, questions, propositions

Lors des premières réunions du groupe de concertation, il est apparu essentiel de brosser un tableau d'ensemble permettant à tous les participants, aussi éloignés de la statistique soient-ils, de s'approprier le cadre général (juridique, pratique, déontologique). De ce point de vue, le travail effectué pour le groupe, puis, ensuite, pour préparer la communication grand public a permis à l'Insee de progresser vers plus de simplicité pour des présentations futures. Pour toutes les présentations, les efforts de simplification déjà réalisés pour la rencontre du Cnis, intensifiés pour la première séance et ajustés au fur et à mesure aux besoins des participants au fil des séances, ont été bénéfiques.

Lors des réunions, les choses ont été rendues les plus simples possibles pour les participants, qu'il s'agisse des modalités de participation, de la prise de connaissance des spécificités du sujet et de ses enjeux ou encore de la mise à disposition de supports informatifs et pédagogiques.

> ***Pendant les réunions du groupe, le choix a été fait de donner la priorité à l'échange, donc de réduire les temps de présentation.***

Pendant les réunions du groupe, le choix a été fait de donner la priorité à l'échange, donc de réduire les temps de présentation[13], en étant aussi concis et simple que possible pour expliquer le projet, les contraintes et opportunités, les enjeux. Un point d'attention dans tous les débats : éviter l'implicite, car le risque est fort que les interprétations de chacun, conditionnées par son propre référentiel, ne soient pas partagées et que des quiproquos perturbent la réflexion. La clarification du vocabulaire était une étape cruciale. L'animation a permis aux intervenants d'être également précis et concis sans brider l'expression.

Enfin, dans cette concertation, on s'est exposé à des questionnements et des propositions parfois très éloignées de la réflexion déjà réalisée en interne sur le projet, voire en contradiction. Le partage des connaissances, des enjeux et des idées (solutions) participe à créer une dynamique de confiance. Toutefois, pour maintenir la confiance, les échanges ont été retranscrits à chaque séance puis dans le rapport avec le plus de fidélité et de précision possible. La rédaction a fait l'objet d'une validation minutieuse avec les participants pour être fidèle aux débats.

L'apport du groupe du Cnis aux réflexions de l'Insee quant à la construction du projet et son accompagnement est incontestable, mais il est important de souligner que le sujet n'est pas épuisé. D'abord car le temps était limité (sept réunions en cinq mois dont deux pendant l'été) et le sujet vaste, technique et peu débattu jusqu'alors ; ensuite parce que si l'on a cherché à mobiliser des spécialistes en dehors de la statistique publique, ils ne se considéraient pas comme investis de la mission de représenter toute la société dans ce groupe, leur expertise rendant leur approche un peu différente de celle du grand public, même s'ils ont contribué à l'appréhender ; enfin parce qu'il s'agit d'un projet en construction à horizon 2025, dont l'évolution et les usages ne peuvent pas être figés une fois pour toutes.

Continuer à bénéficier de regards extérieurs

La démarche de concertation doit se poursuivre : c'est une préconisation forte du groupe.

La concertation sur Résil n'est pas achevée. Elle se poursuivra, sous des formes adaptées et encore à définir, d'abord durant la phase de mise en œuvre de Résil, puis tout au long de son utilisation. Prolonger la concertation sur les principes de nécessité et de proportionnalité, intervenir en cas de risque de mésusage, conseiller l'Insee dans la mise en œuvre de dispositions spécifiques, auditer le résultat, tester les supports de communication : dans tous ces cas, le recours à des instances extérieures à l'Insee est pour les participants du groupe de concertation une garantie visible par le citoyen que les engagements seront suivis donc tenus.

Pour la suite du projet, des regards extérieurs enrichis (parfois déjà sollicités en amont), permettront de vérifier différentes facettes du projet ou de sa mise en œuvre :

- le Cnis pour intégrer les besoins mais aussi, si son rôle évolue, pour prendre en compte le point de vue des citoyens contributeurs, au regard des questions éthiques ;

13 La proportion conseillée par la Commission nationale du débat public est 2/3 d'échange pour 1/3 de présentation.

• la CNIL qui veille au respect de la protection des données personnelles et qui sera consultée pour l'ajout éventuel de sources alimentant Résil ;

• l'Anssi, pour la sécurité du système d'information. Le recours à l'Anssi est de nature à rassurer les citoyens car le répertoire bénéficiera des protections informatiques les plus fortes (haute protection) et d'une revue de conformité des pratiques à jour ;

• l'Autorité de la statistique publique (ASP), garante des bons usages à finalités uniquement statistiques et dans le respect du Code de bonnes pratiques de la statistique européenne et « lanceur d'alerte » dans le cas contraire ;

• des « focus groups » de citoyens pour recueillir leur point de vue sur le dispositif et la manière dont il est présenté.

Au-delà de ces regards, recommandés par le groupe, il sera important de continuer à associer aux échanges des interlocuteurs porteurs de regards spécifiques, sur l'éthique, la protection des libertés publiques, les usages du numérique, etc. Les modalités de cette réflexion commune restent à inventer, mais il est indispensable de poursuivre cette démarche de confrontation des points de vue. Il sera important pour l'Insee de continuer à s'inscrire dans une démarche de communication sur ses travaux et ses valeurs, associée à une écoute renforcée des publics et une analyse régulière de la confiance portée aux informations[14] qu'il délivre.

> ***Il sera important de continuer à associer aux échanges des interlocuteurs porteurs de regards spécifiques, sur l'éthique, la protection des libertés publiques, les usages du numérique, etc.***

▶ Les recommandations du groupe de concertation

En préalable, le groupe a fait part du risque que l'accroissement des appariements encourage une tendance du politique à la « gouvernance par le nombre » : en d'autres termes, la mise à disposition de données plus nombreuses pourrait encourager les gouvernants à ne décider qu'en fonction de données chiffrées. Ce risque existe, mais le groupe a considéré qu'il tenait davantage à l'usage des données plutôt qu'à leur production. Le groupe a également exprimé des craintes sur les risques de mésusages des données de Résil ou des appariements de données qui en résultent en cas de fuite de données consécutive à une cyberattaque, ou lors d'un détournement de finalité à la suite d'une pression politique par exemple.

Les préoccupations du groupe sont centrées sur trois ordres d'idées qui transparaissent dans les recommandations : le principe d'un répertoire de cette nature ; les atteintes à la vie privée qui pouvaient en résulter ; la limitation des accès internes et la protection contre les risques de cyberattaque.

Concernant le principe du répertoire, le groupe a admis que le critère de **nécessité** était satisfait dans la mesure où les besoins de connaissance avec une précision accrue sont réels et qu'il n'y a pas d'alternative convaincante. L'utilisation de répertoires lui

14 Perception de l'image de l'Insee et de ses indicateurs socioéconomiques auprès du « grand public », enquête menée annuellement : *https://www.insee.fr/fr/information/3669009*.

est également apparue conforme au principe de **proportionnalité** exigé des textes en vigueur. Le groupe a également noté que les services proposés par Résil seraient accessibles uniquement aux agents du Service statistique public, pour des traitements à finalité statistique exclusivement, et que la maîtrise de l'outil demeurerait du seul ressort de l'Insee. Ces éléments ont été précisés dans les textes juridiques fondant Résil.

Un avis défavorable pour certaines sources

L'absence d'usage du NIR[15] et un usage approprié du Code statistique non signifiant (CSNS[16]) ont été notés ainsi que l'engagement de l'Insee qu'aucune donnée « sensible » au sens du RGPD ne figurerait dans le répertoire.

Face aux sources de données envisagées pour Résil, le groupe[17] a rendu un avis défavorable sur l'utilisation du fichier des titres de séjour des étrangers (AGDREF[18]), en raison de la sensibilité de la question de la possession des titres de séjour par les étrangers. Il a également émis un avis défavorable en raison d'un risque d'image et d'une non-proportionnalité pour le fichier de la carte Vitale[19] et pour le fichier RNCPS[20] (lui-même conglomérat de fichiers de diverses caisses de protection sociale).

Un cadre juridique à enrichir

L'encadrement juridique de Résil par un décret en Conseil d'état pris après avis de la CNIL a paru souhaitable au groupe. Il a recommandé que les sources de données constitutives des répertoires des individus et des logements figurent dans un texte séparé pris après avis du Cnis et de la CNIL, afin de s'assurer qu'un examen de la nécessité et de la proportionnalité soit réalisé pour chaque source utilisée. Ces textes devront être mis à jour, après avis de la CNIL et du Cnis, à chaque modification de la liste des sources (voir les références juridiques en fin d'article).

Le groupe a estimé que le fondement juridique des appariements pouvait être mieux assuré et a recommandé qu'il soit clarifié. Il a par ailleurs jugé que la vigilance sur le risque de stigmatisation de groupes sociaux du fait de l'accroissement du nombre des appariements, bien que déjà présente dans les travaux statistiques actuels, devait être renforcée. À cette fin, le groupe recommande de publier la liste des appariements réalisés grâce à Résil en mentionnant le responsable de traitement, les sources utilisées, les finalités des traitements, les populations concernées, de manière à rendre compte du respect des principes de nécessité et de proportionnalité. Il a préconisé que le Cnis donne un avis d'opportunité sur les appariements réalisés par la statistique publique en prenant en compte les dimensions éthiques et celles liées aux droits et aux libertés des personnes et a recommandé pour cela d'élargir les missions du Cnis et de modifier sa composition, en y nommant des membres compétents en matière de libertés et de protection des données.

15 NIR : Numéro d'inscription au Répertoire national d'identification des personnes physiques ou numéro de sécurité sociale.
16 Voir l'article du Courrier des statistiques N9 : « Le code statistique non signifiant : un enjeu majeur pour le service statistique public », Yves-Laurent Bénichou, Lionel Espinasse et Séverine Gilles
17 Le détail est donné en partie 5.3.7 du rapport du groupe de concertation (Rapport du groupe de concertation du Cnis, 2022).
18 AGDREF : Application de gestion des dossiers des ressortissants étrangers en France.
19 La carte Vitale est la carte de l'assurance maladie en France.
20 Répertoire National Commun de la Protection Sociale. *https://www.securite-sociale.fr/la-secu-en-detail/gestion-financement-et-performance/rncps*.

Prévenir les mésusages

La sécurité informatique de Résil doit être garantie par l'Insee qui porte la responsabilité du dispositif et des données qu'il traite. Les précautions que l'Institut prendra à cet égard (en conformité avec la politique de sécurité de l'État et des recommandations de l'Anssi) ont paru appropriées au groupe qui demande à ce que la sécurité du système d'information fasse l'objet d'audits réguliers par un intervenant externe.

Le groupe a recommandé également de s'appuyer sur l'Autorité de la statistique publique (ASP) pour prévenir les risques de mésusages qui seraient en contradiction avec les règles fixées : l'ASP doit pouvoir intervenir à titre préventif en cas de pression sur les services et dénoncer tout mésusage ; il est important qu'elle soit régulièrement informée de l'avancement du projet, puis de ses usages.

Développer la communication vers le grand public

> ***Le groupe a suggéré de développer une communication à destination du grand public sur l'utilisation des sources administratives et les appariements de données.***

Plus largement, le groupe a suggéré de développer une communication à destination du grand public sur l'utilisation des sources administratives et les appariements de données en s'inspirant des bonnes pratiques des instituts nationaux statistiques étrangers, en particulier celui du Canada. Il a recommandé de poursuivre la concertation sur le projet Résil : il a proposé de tenir une nouvelle rencontre du Cnis sur les appariements et la mise en œuvre de Résil en 2025 et a souhaité que le bureau du Cnis et les commissions compétentes soient régulièrement informés des avancées du projet Résil. En ce qui concerne la communication sur le projet et sur ses usages, la transparence doit rester la règle. Pour cela, il a recommandé d'ouvrir une rubrique internet grand public très complète sur le site de l'Insee pour présenter le projet Résil, ses finalités, les apports attendus, les modalités de protection des données, les textes juridiques et la délibération de la CNIL.

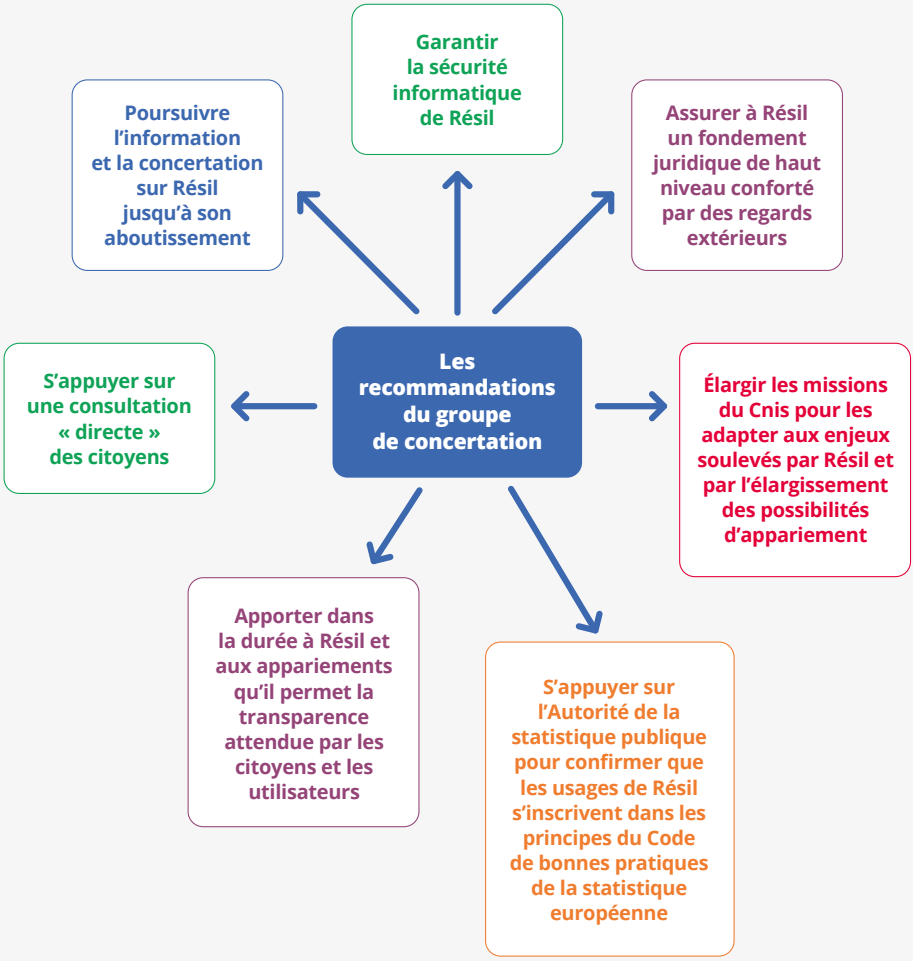
▶ Les enseignements de cette concertation

Pour l'Insee, l'exercice a été exigeant, mais très stimulant et utile. Les échanges ont permis une réelle amélioration du projet et une appréciation partagée des enjeux de proportionnalité, permettant de questionner certaines évidences (qui ne l'étaient pas tant que ça, finalement) et de positionner de manière pertinente le curseur. Il en tire des enseignements directs concernant Résil, mais aussi des enseignements plus généraux sur la démarche de concertation et la communication sur la production des statistiques.

Recommandations en lien direct avec le projet Résil

Concernant le projet, l'Insee a décidé de prendre en compte la quasi-totalité des recommandations (***figure 3***), ce qui était possible grâce à une concertation suffisamment en amont du projet.

▶ Figure 3 - Les recommandations du groupe de concertation



Source : Rapport du Cnis n° 161, novembre 2022.

L'Insee a ainsi intégré dans le décret fondant Résil les dispositions recommandées par le groupe. Il a également mis en œuvre, conformément aux recommandations du groupe, une communication spécifique au projet Résil : celle-ci se présente sous la forme de pages internet donnant accès à des infographies et vidéos, avec un travail important pour les rendre les plus accessibles possibles[21]. Ces supports ont fait l'objet d'un examen critique par un « focus group » composé de personnes représentatives de la société (aux profils variés) qui connaissaient peu ou pas nos travaux. Les échanges très riches d'enseignements ont permis de simplifier au maximum le message délivré. Cette communication vient compléter une communication plus générale sur nos travaux, réalisée grâce à deux billets

21 Le projet Résil : un outil pour mieux connaître la société française, Présentation grand public de Résil sur le site insee.fr : *https://www.insee.fr/fr/information/7748883*.

de blog : l'un « Quels types de sources l'Insee utilise-t-il pour construire ses statistiques ? », le second « Les appariements de données de la statistique publique : des analyses enrichies, un cadre juridique protecteur » (Dupont, mai et septembre 2023).

Dans sa délibération, la CNIL a fait de multiples références aux travaux du groupe de concertation, d'abord pour saluer la démarche engagée par l'Insee, ensuite pour en intégrer les conclusions en complément des informations transmises par l'Insee. Elle appelle de ses vœux la poursuite de la démarche de recherche de regards extérieurs, de prise en compte des questions éthiques, de transparence sur Résil et ce qu'il permettra de faire.

Au-delà du projet Résil, cette démarche de concertation a permis aux membres du groupe de découvrir les travaux et les pratiques des statisticiens publics, notamment l'ampleur des données traitées, ainsi que leurs valeurs. Elle a, de plus, fortement incité l'Insee à communiquer davantage sur ses travaux et ses valeurs.

Enseignements pour mener les concertations sur des projets à fort impact

Par rapport aux pratiques de concertation plus habituelles du Cnis, le format de la concertation retenu pour Résil est différent : il associe des personnalités plus éloignées de la statistique publique et de ses utilisations. La mise au point s'est appuyée sur diverses pratiques et expertises de concertation (CNDP[22], chercheurs, Cese, etc.). À l'issue de cet exercice, il est possible de tirer des enseignements qui ne sont pas spécifiques à Résil.

Afin que la concertation soit génératrice d'impacts sur le projet et donc d'intérêt à participer pour les parties prenantes, il est indispensable de la positionner assez tôt dans le calendrier du projet, dès que ses grandes lignes sont connues, avec peut-être l'apport d'un prototype de ce que l'on souhaite réaliser et des expérimentations sur certaines parties du projet. Le tempo est important, car si cela intervient trop tôt, le risque est de discuter sur des principes très abstraits ; au contraire, si le projet est trop avancé, la marge de manœuvre de co-construction est trop limitée.

Il ne faut pas sous-estimer le temps nécessaire à la préparation du cadre de concertation : toute concertation implique de déterminer son champ, son organisation opérationnelle, les règles du dialogue, puis de mettre en place la structure (infrastructure de concertation, recherche de « concertants », création des supports pédagogiques pour éclairer de façon pédagogique le sujet de concertation).

Suivant en cela les recommandations du groupe de concertation, la concertation sur Résil devra se poursuivre au sein du Cnis, instrument de gouvernance central en matière de concertation sur la statistique publique en France. Ce dernier devra être le garant d'un équilibre constant entre le besoin de connaissance et la protection des libertés individuelles, dans un contexte marqué par un recours croissant aux appariements de données individuelles.

Plus généralement, Résil ouvre des perspectives de développement des appariements dans un cadre protecteur sur le plan informatique, juridique et éthique. Cela suppose de faire évoluer les modalités de concertation en accordant plus de place aux réflexions sur la protection des libertés individuelles en matière d'appariements. Cette réflexion est en cours.

22 Commission nationale du débat public.

▶ 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). In : *Journal officiel de l'Union européenne*. [en ligne]. Mise à jour le 4 mai 2016. [en ligne]. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A32016R0679*.

- Règlement (CE) No 223/2009 du Parlement européen et du Conseil du 11 mars 2009 relatif aux statistiques européennes et abrogeant le règlement (CE, Euratom) nº 1101/2008 relatif à la transmission à l'Office statistique des Communautés européennes d'informations statistiques couvertes par le secret, le règlement (CE) nº 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]. Version consolidée 2015. [en ligne]. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:02009R0223-20150608*.

- Loi nº 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. [en ligne]. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

- Loi nº 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 23 mai 2024. [en ligne]. [Consulté le 18 juin 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460*.

- Loi nº 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

- Loi organique nº 2021-27 du 15 janvier 2021 relative au Conseil économique, social et environnmental. In : *site de Légifrance*. [en ligne]. Mise à jour le 18 janvier 2021. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000042093735/*.

- Décret nº 2024-12 du 5 janvier 2024 portant création d'un traitement automatisé de données à caractère personnel dénommé « répertoire statistique des individus et des logements » (Résil). In : *site de Légifrance*. [en ligne]. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000048866207*.

- Arrêté du 5 janvier 2024 pris en application de l'article 2 du décret no 2024-12 du 5 janvier 2024 portant création d'un traitement automatisé de données à caractère personnel dénommé « répertoire statistique des individus et des logements » (Résil). In : *site de Légifrance*. [en ligne]. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000048866233*.

• Délibération no 2023-080 du 20 juillet 2023 portant avis sur un projet de décret en Conseil d'État relatif à la création du traitement automatisé de données à caractère personnel permettant la gestion du répertoire statistique d'individus et de logements, et sur l'arrêté y afférent. In : *site de Légifrance*. [en ligne]. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000048934586?page=1&pageSize=100&searchField=ALL&searchType=ALL&sortValue=DATE_DECISION_DESC&tab_selection=cnil&timeInterval=01%2F07%2F2023+%3E+31%2F08%2F2023&typePagination=DEFAULT*.

▶ Bibliographie

- AGACINSKI, Daniel, 2018. Expertise et démocratie. Faire avec la défiance. In : *Rapport de France Stratégie*. [en ligne]. Décembre 2018. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.strategie.gouv.fr/publications/expertise-democratie-faire-defiance*.

- ANXIONNAZ, Isabelle et MAUREL, Françoise, 2021. Le Conseil national de l'information statistique – La qualité des statistiques passe aussi par la concertation. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 123-142. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398693?sommaire=5398695*.

- BÉNICHOU, Yves-Laurent, ESPINASSE, Lionel et GILLES, Séverine, 2023. Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N°N9, pp 64-85. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635825?sommaire=7635842*.

- BUREAU, Dominique, 2020. L'Autorité de la statistique publique. Dix ans d'activité, pour une statistique indépendante et de qualité. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 21-38. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5008696?sommaire=5008710*.

- Chaire valeurs et politiques des informations personnelles, 2019. Données personnelles et confiance : évolution des perceptions et usages post-RGPD. In : *site de chaire VP-IP*. [en ligne]. 31 octobre 2019. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://cvpip.wp.imt.fr/31-10-2019-donnees-personnelles-et-confiance-evolution-des-perceptions-et-usages-post-rgpd/*.

- CHRISTINE, Marc et ROTH, Nicole, 2020. Le Comité du label. Un acteur de la gouvernance au service de la qualité des statistiques publiques. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 39-52. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5008698?sommaire=5008710*.

- CNDP, 2023. Modalités d'information et de participation du public. In : *Commission nationale du débat public, catalogue d'outils*. [en ligne]. Version d'octobre 2023. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.debatpublic.fr/sites/default/files/2023-10/catalogue-outils.pdf*.

- CNIL, 2016. Le règlement général sur la protection des données – RGPD. In : *site de la CNIL*. [en ligne]. 24 mai 2016. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.cnil.fr/fr/reglement-europeen-protection-donnees*.

- CNIS, 2022. Le projet de répertoire statistique d'individus et de logements « Résil ». In : *Rapport du groupe de concertation du Cnis*. [en ligne]. Novembre 2022. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2022/11/rapport-version-dfinitive.pdf*.

- CNIS, 2022. Rencontre – Appariements de données individuelles : entre richesse de l'information statistique et respect de la vie privée. In : *Site du Cnis*. [en ligne]. 28 janvier 2022. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.cnis.fr/evenements/appariements-de-donnees-individuelles-entre-richesse-de-linformation-statistique-et-respect-de-la-vie-privee/?category=1067*.

- Dictionnaire critique et interdisciplinaire de la Participation, DicoPart. Gis Démocratie et Participation. In : *site de DicoPart*. [en ligne] [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.dicopart.fr/*.

- DUPONT, Françoise, 2023. Quels types de sources l'Insee utilise-t-il pour construire ses statistiques ? In : *Le blog de l'Insee*. [en ligne]. 16 mai 2023. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://blog.insee.fr/quels-types-de-sources-l-insee-utilise-t-il/*.

- DUPONT, Françoise, 2023. Les appariements de données de la statistique publique : des analyses enrichies, un cadre juridique protecteur. In : *Le blog de l'Insee*. [en ligne]. 1er septembre 2023. [Consulté le 5 mars 2024]. Disponible à l'adresse : *https://blog.insee.fr/appariements-de-donnees-de-la-statistique-publique/*.

- DUPONT, Françoise, GUILLAUMAT-TAILLIET, François et D'ALESSANDRO, Cristina, 2023. Appariements de données individuelles : vers une meilleure qualité et plus de transparence. In : *Cnis – Chroniques n°32*. [en ligne]. Avril 2023. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2023/04/cnis-32-v5.pdf*.

- 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]. 29 novembre 2022. Insee. N° N8, pp. 72-92. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665188?sommaire=6665196*.

- PEPPIN, Aidan, 2022. Who cares what the public think ? UK public attitudes to regulating data and data-driven technologies. In : *site de Ada Lovelace Institute*. [en ligne]. 5 mai 2022. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.adalovelaceinstitute.org/evidence-review/public-attitudes-data-regulation/*.

- POULAIN, Claude, 2022. Le projet SAFARI (1970-1974). In : *Terminal*. [en ligne]. 12 octobre 2022. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://journals.openedition.org/terminal/8787*.

- RANCOURT, Éric, 2019. The Scientific Approach as a Transparency Enabler Throughout the Data Life-cycle. In : *Statistical Journal of the IAOS*. 10 décembre 2019. Vol. 35, n°4, pp. 549-558.

- REDOR, Patrick, 2023. Confidentialité des données statistiques : un enjeu majeur pour le service statistique public. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N°N9, pp 46-63. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635823?sommaire=7635842*.

- ROUBAN, Luc, 2022. *Les raisons de la défiance*. 6 janvier 2022. Presses de Sciences Po. ISBN 9782724638202.

- Statistique Canada, 2019. Principes de nécessité et de proportionnalité. In : *site de Statistique Canada*. [en ligne]. Mis à jour le 27 mai 2022. [Consulté le 7 mai 2024]. Disponible à l'adresse : *https://www.statcan.gc.ca/fr/confiance/reponse*.

Les appariements : finalités, pratiques et enjeux de qualité

Heidi Koumarianos[*], *Olivier Lefebvre*[**] *et Lucas Malherbe*[***]

Les appariements rapprochent les données individuelles issues de fichiers différents. Dans un contexte de disponibilité croissante de sources, notamment administratives, ils sont de plus en plus fréquemment utilisés dans la statistique publique, à des fins d'analyse, pour éclairer des questions nouvelles, ou pour améliorer des processus de production. Ces traitements statistiques posent souvent des difficultés, liées aux imperfections des données utilisées et à leur volume.

Un cadre juridique approprié est nécessaire pour les mettre en œuvre, compte tenu des enjeux de respect des secrets, en particulier la protection des données à caractère personnel. Une bonne connaissance des données utilisées et une attention particulière pour déterminer le paramétrage sont également indispensables pour assurer la meilleure qualité possible du résultat au regard de l'usage attendu, un appariement n'étant jamais fiable à 100 %.

La qualité de ces appariements est donc un enjeu majeur pour la statistique publique et passe par des évaluations directes du processus, nécessairement complétées par l'étude des impacts statistiques des appariements sur les données produites.

Record linkage reconciles individual data that are taken from various data files. It is used more and more frequently in official statistics, whether for analytical purposes, to investigate new subjects, or to improve production processes. Such statistical processing often raises issues relating to the imperfections of the data used and their volume.

Given the issues related to the compliance with data confidentiality, especially regarding personal data protection, an appropriate legal framework is required to implement them. A good knowledge of the data used and particular care in determining the parameters are also necessary to ensure the best possible quality of the results, since record linkage is never 100% reliable.

The quality of these matches is therefore a major issue for official statistics, and requires a direct assessment of the process, necessarily supplemented by a study of the statistical impact of the linkage on the data produced.

* Méthodologue, DMCSI, Insee.
heidi.koumarianos@insee.fr

** Maître d'ouvrage du programme Résil, DSDS, Insee.
olivier.lefebvre@insee.fr

*** À la date de rédaction, data scientist, DMCSI, Insee.

La mesure des revenus des ménages ou le suivi de l'insertion professionnelle des diplômés ont un point commun : ces deux opérations reposent sur l'utilisation de plusieurs sources de données, qu'il faut combiner entre elles. Au niveau le plus fin, il faut réunir, pour chaque individu ou ménage, les données qui le concernent dans chacune des sources. Cela permet de couvrir l'ensemble des sources de revenus des ménages, qu'ils soient imposables ou non imposables, ou encore de suivre le parcours professionnel des diplômés, et notamment leurs conditions d'entrée sur le marché du travail.

Combiner différentes sources permet une observation plus riche et plus efficace. Utilisable dans bien des domaines, l'appariement (l'opération qui permet d'associer et combiner les sources) constitue en quelque sorte une technique de collecte, avec ses contraintes techniques, ses défis méthodologiques, son encadrement juridique, ses enjeux déontologiques. La plupart des instituts de statistique utilisent cette technique pour la production de données statistiques, en lien avec l'utilisation croissante de données administratives, souvent très précises mais très ciblées quant à leur contenu, et qui demandent donc à être complétées.

L'Insee et plus généralement la statistique publique procèdent à des appariements depuis de longues années ; cette pratique s'est progressivement étendue grâce au développement de techniques performantes de traitement des données et à un accès facilité aux fichiers source. On peut ainsi citer Fidelimmo (André et Meslin, 2022), qui permet de mieux analyser le patrimoine immobilier des ménages et les effets redistributifs ou non de la taxe foncière, InserJeunes (Midy, 2021) pour la mesure de l'insertion professionnelle des apprentis, ou encore Sirus[1] (Hachid et Leclair, 2022), colonne vertébrale de la statistique d'entreprises.

▶ Apparier pour enrichir ou mieux comprendre des sources…

L'appariement, lorsqu'il est effectué pour un usage statistique, permet en première approche d'apporter un complément d'information à un fichier statistique existant. Plus généralement, on peut distinguer plusieurs usages de l'appariement :

- **Compléter le champ d'analyse** : le dispositif Filosofi[2] par exemple utilise différentes sources pour reconstituer les revenus des ménages, qu'il s'agisse de revenus du travail ou de prestations sociales. Il permet d'avoir une vue plus complète des revenus des ménages, à une échelle géographique fine pouvant aller jusqu'aux quartiers d'une ville, si ceux-ci sont d'une taille suffisante.
- **Éclairer certains phénomènes** : par exemple, apparier des fichiers de diplômés de l'enseignement supérieur avec des fichiers d'emploi permet de décrire l'insertion professionnelle des jeunes diplômés.

1 Sirus : Système d'immatriculation au répertoire des unités statistiques.
2 Filosofi : ensemble d'indicateurs sur les revenus localisés sociaux et fiscaux.

• **Connaître l'impact d'une aide sociale ou d'une aide à destination des entreprises** : apparier le fichier des bénéficiaires à un fichier décrivant la situation (emploi, réussite dans l'enseignement supérieur, résultats financiers) permet de savoir si l'aide a été suivie d'effets et mettre en place une évaluation de l'impact de l'aide.

• **Mieux comprendre le contenu des sources analysées** : par exemple, l'appariement du fichier des demandeurs d'emploi avec celui de l'enquête Emploi a permis de mieux comprendre les évolutions différentes du chômage au sens du Bureau International du Travail et du nombre de demandeurs d'emploi inscrits à France Travail[3].

▶ … ou encore améliorer des processus de production

Au-delà de la construction de données nouvelles, combiner les sources permet d'améliorer notablement des processus de production statistique. On modifie la phase de collecte des informations ou de contrôle ou d'évaluation de la qualité des sources (cohérence avec les concepts que l'on souhaite mesurer, mesure de la couverture). Plus précisément, un appariement conduit à :

• **Alléger les questionnaires d'enquête** : le principe est de ne pas demander à un ménage (ou à une entreprise) une information qu'il (ou elle) a déjà transmise à une administration, selon le principe « Dites-le nous une fois ». Par exemple, l'appariement de l'enquête Emploi avec les données fiscales permet de réduire le nombre de questions portant sur le revenu.

• **Mettre à jour un répertoire ou un référentiel** (RNIPP[4], Sirene[5], REU[6], Résil[7], Sirus, etc.) : on ajoute de nouvelles entités dans le répertoire (auquel cas il est essentiel de vérifier qu'elles n'y figurent pas déjà, pour éviter les doublons) ou on met à jour certaines caractéristiques (auquel cas il est essentiel de vérifier qu'on met à jour la bonne observation). La qualité des répertoires est indispensable à celle des processus statistiques (Espinasse et Roux, 2022 ; Demotes-Mainard, 2019).

• **Analyser la couverture d'une source en l'appariant à un référentiel** : c'est un progrès important dans l'analyse des évolutions ou dans le traitement des « trous de collecte », comme pour les non-répondants d'une enquête (ce qui est actuellement possible en statistique d'entreprises avec Sirus et qui pourra l'être en statistique démographique et sociale avec Résil).

▶ De quoi parle-t-on ?

En pratique, l'appariement désigne l'opération de rapprochement, au niveau de chacune des unités d'observation, de deux fichiers de données A et B, soit pour enrichir l'un des fichiers avec des variables supplémentaires ou mises à jour, soit pour créer un nouveau fichier contenant tout ou partie des variables de chacun des fichiers ***(Figure 1)***.

3 France Travail : depuis 2023, France Travail succède à Pôle emploi.
4 RNIPP : Répertoire national d'identification des personnes physiques.
5 Sirene : Système national d'identification et du répertoire des entreprises et de leurs établissements.
6 REU : Répertoire électoral unique.
7 Résil : Répertoire Statistique des Individus et des Logements. Voir l'article d'Olivier Lefebvre sur Résil dans ce même numéro.

Figure 1 - Un exemple d'appariement pour étudier le lien entre profession et niveau de diplôme

Table A

Identifiant | Traits d'identité (variables d'appariement) | Variable statistique d'intérêt

ID_A	Nom	Prénom	Date de naissance	Profession
A1	RAFIK	Kélian	03/02/2001	Enseignant
A2	GIROUT	Claudia	05/08/1993	Esthéticienne
A3	REGARD	Esteban	23/11/1997	Cariste
A4	MOREAU	Lya	12/11/2002	Agent de police
A5	PEREZ	Maria	27/06/1987	Cadre commercial
A6	LEFEBVRE	Olivier	27/06/1999	Statisticien

?

Table B

Identifiant | Traits d'identité (variables d'appariement) | Variable statistique d'intérêt

ID_B	Nom	Prénom	Date de naissance	Dipôme le plus élevé
B1	LEFEVRE	Oscar	02/04/1975	DUT
B2	RAFIK	Kélian	03/02/2001	Master
B3	REGAR	Esteban	23/11/1997	Bac
B4	GIROUT	Claudia	05/08/1993	Brevet d'études professionnelles
B5	PEREZ	Marina	27/06/1987	Doctorat

Table de résultats

Id	Date de naissance	Profession	Dipôme le plus élevé
R1	03/02/2001	Enseignant	Master
R2	05/08/1993	Esthéticienne	BEP
R3	23/11/1997	Cariste	Bac
R4	27/06/1987	Cadre commercial	Doctorat

Note de lecture : À partir des deux fichiers A et B, contenant respectivement la profession et le diplôme, on veut construire un fichier de résultat contenant, pour les individus figurant dans les deux fichiers, leur profession et leur diplôme. Le processus de construction de ce fichier est détaillé dans la suite de l'article.

> ***Si les fichiers ont un identifiant commun de bonne qualité, l'opération technique est triviale… Elle se résume alors à une « jointure » sur cet identifiant.***

Un appariement peut être destiné à un usage administratif (par exemple une vérification de droits), opérationnel (par exemple fusionner des fichiers de clients) ou à un usage statistique (***encadré 1***).

La suite de cet article porte sur la manière de réaliser les appariements et sur leurs usages statistiques.

Pour que ce rapprochement soit pertinent, il faut savoir à quelle ligne[8] du fichier B correspond chaque ligne du fichier A, en s'assurant donc qu'elles portent toutes deux sur la même unité d'observation. Si les fichiers ont un identifiant commun de

8 Une ligne correspond à un enregistrement représentant un individu.

bonne qualité, l'opération technique est triviale : deux lignes ayant le même identifiant portent naturellement sur la même unité d'observation. Elle se résume alors à une « jointure » sur cet identifiant (il convient néanmoins d'en traiter tous les autres aspects : encadrement juridique et déontologique, analyse de la qualité statistique du résultat, etc.). Dans le cas contraire, deux approches sont possibles :

- l'identification individuelle : on trouve un identifiant commun en comparant ces deux fichiers A et B à un troisième fichier, C, souvent de taille plus importante, qui joue le rôle de référentiel. Cette étape d'identification vise donc à chercher successivement, pour les observations des fichiers A et B, à quelle ligne du fichier C elles correspondent, et à faire la « jointure » sur l'identifiant correspondant[9] ;

- la confrontation de paires : on confronte directement les deux fichiers, en cherchant parmi toutes les paires d'observations possibles, lesquelles correspondent à la même entité ; on utilise pour cela des techniques spécifiques, fondées soit sur l'application de règles de décisions successives, soit sur des modèles probabilistes.

Il s'agit d'appariements de micro-données. Il existe également des appariements statistiques, fondés sur l'appartenance à des strates, que l'on n'évoquera pas ici. Les anglophones désignent les premiers par « *record linkage* » et les seconds par « *propensity score matching* » (Rosenbaum et Rubin, 1983).

▶ Encadré 1. Appariement, enrichissement, interconnexion, couplage : tous synonymes ?

Un **enrichissement** de données d'un fichier par un autre fichier consiste à trouver, pour un individu donné, les informations qui le concernent dans deux fichiers différents, puis à constituer un troisième fichier avec les données ainsi rassemblées.

Même si les appariements ne désignent que la première phase de cette opération (celle qui consiste à relier entre elles deux observations relatives à la même entité), les statisticiens désignent souvent cette technique sous le nom d'« **appariement** » de fichiers (ou de données). C'est ce terme qui figure dans la loi de 1951 et le décret d'application de la loi pour une République numérique*.

Les rédacteurs de la loi relative à l'Informatique, aux fichiers et aux libertés de 1978 et du Règlement général pour la protection des données ont employé les termes « **interconnexion** », « **rapprochement** », ou « **mise en relation** » de fichiers. L'interconnexion de fichiers et leur rapprochement constituent deux formes de mises en relation de fichiers ; le terme d'interconnexion est plus souvent employé pour des mises en relation présentant un fort degré d'automatisation, voire totalement automatisées.

D'autres termes peuvent être utilisés par les statisticiens pour désigner les appariements : **combinaison ou couplage** de fichiers (le dernier terme étant celui utilisé au Canada à la fois par Statistique Canada et dans la directive** qui régit ces opérations).

Pour clarifier le sujet et mieux asseoir la notion d'appariement en droit français, le décret fondant Résil propose une définition de l'appariement*** :

« Ces appariements constituent des mises en relation, au sens du 3° du I de l'article 33 de la loi [Informatique et Libertés], entre les données à caractère personnel enregistrées sur le « répertoire statistique des individus et des logements » et des sources de données statistiques tierces. Ils donnent lieu à la création de nouveaux fichiers, lesquels constituent des traitements de données à caractère personnel au sens du [RGPD]. »

* *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).*

** *Statistique Canada. 2017. Directive sur le couplage de microdonnées. https://www.statcan.gc.ca/fr/enregistrement/politique4-1.*

*** *Décret n° 2024-12 du 5 janvier 2024 portant création d'un traitement automatisé de données à caractère personnel dénommé « répertoire statistique des individus et des logements » (Résil) (voir fondements juridiques).*

9 Si l'un des deux fichiers est lui-même un référentiel, par exemple si on cherche à mettre à jour ce dernier, on ne réalise l'opération d'identification que sur l'autre fichier.

▶ Un traitement de données nécessite un cadre juridique adapté

L'appariement de données personnelles est un traitement de données au sens juridique qu'il faut traiter comme tel. Un responsable du traitement doit être identifié, charge à lui de s'acquitter des obligations imposées par le Règlement général pour la protection des données (RGPD) et la loi Informatique et Libertés[10] : vérification du respect des principes de nécessité, minimisation et proportionnalité[11], inscription au registre des traitements de son administration, réalisation d'une étude d'impact si ce traitement présente certaines caractéristiques (par exemple s'il porte sur une population très nombreuse ou mobilise des variables sensibles, etc.). Quand l'appariement porte sur des données statistiques ou est réalisé à des fins statistiques, les données sont placées sous la protection de la loi de 1951 sur l'obligation, la coordination et le secret en matière de statistiques[12] (*Redor, 2023*).

▶ Une pratique souvent délicate

Quand on ne dispose pas d'identifiant commun aux deux fichiers, l'appariement doit être effectué sur des variables permettant d'identifier sans ambiguïté les personnes.

Un fichier comporte plusieurs types d'informations jouant un rôle différent dans un processus d'appariement. On peut classer ces informations en trois catégories :

- les informations identifiantes primaires : il s'agit de traits d'identité associés à un individu de manière unique et très stable dans le temps. Pour une personne, il s'agit de ses nom(s), prénom(s), lieu et date de naissance[13] ;
- les informations identifiantes secondaires : ce sont des informations qui ne sont pas associées de manière unique et permanente à un individu, mais peuvent permettre d'améliorer le processus d'appariement. Pour une personne, il pourra s'agir par exemple de sa commune et de son adresse de résidence ;
- les autres informations ne sont généralement pas utilisées dans un processus d'appariement, mais sont des variables d'intérêt pour le fichier statistique produit. Elles peuvent toutefois être mobilisées lors de l'évaluation de la qualité, en détectant par exemple des incohérences dans les enregistrements appariés.

Cela pose plusieurs difficultés : ces informations ne sont pas toujours présentes dans les fichiers et peuvent être entachées d'erreurs ou d'omissions. Leur comparaison est très coûteuse en ressources informatiques, et ce coût croît rapidement avec la taille des données.

10 Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (voir les références juridiques en fin d'article).
11 Selon le RGPD, « les données à caractère personnel doivent être [...] adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées (minimisation des données) ».
12 Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques (voir les références juridiques en fin d'article).
13 Pour un établissement, la raison sociale et l'adresse constituent des informations primaires.

▶ Quels critères utiliser quand on n'a pas d'identifiant commun ?

Une première condition est bien sûr de disposer des informations identifiantes dans les deux fichiers, et que ces informations soient comparables, c'est-à-dire que leur contenu sémantique et leur représentation soient identiques. Par exemple, on dispose de la commune de résidence dans les deux fichiers et cette dernière figure dans les deux cas sous forme de libellé, ou de code selon une nomenclature identique.

Outre ce critère nécessaire de comparabilité des données, trois conditions principales sont requises pour réaliser avec succès un appariement :

- la richesse des informations ;
- la qualité des informations ;
- un processus efficace pour traiter un important volume de données.

Les informations dont on dispose doivent être suffisamment nombreuses, discriminantes et précises pour permettre de distinguer les individus les uns des autres. Connaître le mois de naissance par exemple est beaucoup moins informatif que connaître le prénom (un douzième de la population partage la même valeur pour le mois de naissance). Plus l'information est précise, plus elle permettra de distinguer un individu d'un autre. Il sera ainsi plus intéressant d'utiliser une date de naissance complète plutôt qu'une année, une commune de naissance plutôt qu'un département, etc.

On peut aussi chercher à mobiliser un plus grand nombre d'informations, en utilisant une adresse de résidence, en plus des traits d'identité. Il peut toutefois subsister malgré tout, des cas non univoques (en cas d'homonymie par exemple), qui sont d'autant plus nombreux que les données sont peu précises ou comportent des erreurs.

▶ Les données identifiantes ne sont jamais parfaites : comment s'en accommoder ?

La deuxième difficulté est liée à la qualité des données. Toute information manquante, incomplète ou erronée est susceptible de nuire à la qualité de l'appariement, en entraînant des appariements à tort, ou à l'inverse, en « ratant » de nombreuses paires. Lors d'une enquête statistique, les traits d'identité des personnes interrogées ne font pas partie des variables d'intérêt, et peuvent alors souffrir de défauts de qualité. Or, ces informations sont indispensables, dès lors que l'on envisage un processus d'appariement.

Pour pallier ces difficultés, le statisticien dispose d'outils permettant de préparer les données utilisées pour l'identification en vue d'améliorer le résultat de son appariement.

Ces outils ne sont pas magiques et ne peuvent créer une bonne information là où elle manque ou est erronée. On peut toutefois recourir à un processus de normalisation des données, notamment pour améliorer la comparabilité :

- on norme les données dans un format identique de part et d'autre : transformer

décembre en 12, ou un nom de commune en son code commune ou encore utiliser une casse similaire (suppression des caractères accentués, passage en minuscules, etc. (Cotton et Haag, 2023)) ;

- on structure ensuite les informations : normaliser un libellé d'adresse (type de voie, nom de voie, numéro dans la voie, indice de répétition, commune) ou encore identifier un nom dans un champ comportant aussi la civilité ; ce second point est toutefois moins évident, car il nécessite d'analyser les libellés.

Parfois le sujet est plus délicat, lorsqu'un fichier administratif comporte dans un même champ le nom marital et le nom de naissance ou encore le nom de plusieurs titulaires d'une carte grise par exemple. Ce processus est efficace dans la mesure où les traitements entrepris sont assez déterministes (comme rassembler les Bd, boul, Boulevard sous une seule dénomination). Mais attention à ne pas aller trop loin : la volonté de supprimer des données erronées peut conduire à supprimer de l'information, et finalement dégrader le processus (Koumarianos, 2022). Ce reproche est parfois adressé aux algorithmes phonétiques : ils ont pour objectif de neutraliser les différences orthographiques, mais ils peuvent alors considérer à tort que Lefebvre et Lefèvre, ou Schmidt et Schmitt sont deux modalités identiques (Randall et alii, 2013).

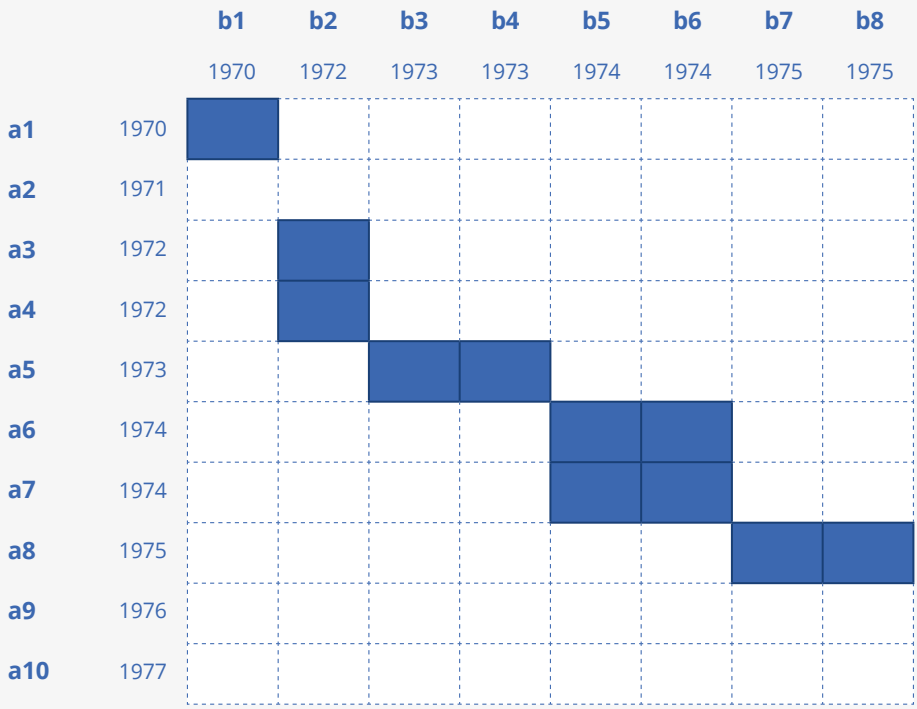
Lorsqu'il subsiste des erreurs dans les informations, de type faute de frappe ou faute d'orthographe, il sera souvent plus efficace de gérer ces problèmes ultérieurement dans le processus d'appariement, en mobilisant des mesures de similarité tenant compte de ces potentielles coquilles, plutôt que d'établir une comparaison s'appuyant sur une égalité stricte.

▶ L'appariement est une opération gourmande, comment la rendre frugale ?

Enfin un troisième enjeu est celui de la performance informatique du processus. Un processus d'appariement revient à sélectionner, au sein de deux fichiers de tailles N_1 et N_2, les paires d'individus identiques au sein de l'ensemble des paires possibles. Cet ensemble est très grand, de taille $N_1 \times N_2$, alors que le nombre de paires d'individus identiques est inférieur ou égal au minimum de (N_1,N_2). En dépit des progrès techniques, un appariement reste un problème de grande taille. Apparier deux fichiers de 60 000 lignes signifie raisonner dans un ensemble de 3,6 milliards de paires potentielles. Parmi l'ensemble des paires, un grand nombre rassemble deux individus qui ne se ressemblent pas du tout. Il n'est donc pas efficace de constituer l'ensemble théorique de toutes les paires. On cherche souvent à le réduire à un sous-ensemble de paires plus vraisemblables. C'est ce qu'on appelle le blocage : réduire la dimension du problème en ne comparant, par exemple, que des individus nés la même année (***figure 2***). Les informations utilisées pour le blocage doivent être d'excellente qualité. Dans le cas contraire, cela entraînerait de nombreux appariements manqués.

> ***C'est ce qu'on appelle le blocage : réduire la dimension du problème.***

▶ Figure 2 - Exemple de blocage sur l'année de naissance



Lecture : Dans cet exemple fictif, on cherche à apparier le fichier A qui comporte 10 individus et le fichier B qui en comporte 8.
On bloque sur l'année de naissance, ce qui permet de travailler sur les paires d'individus nées la même année.
Au lieu de constituer 80 paires à des fins de comparaison (l'ensemble des carrés de la matrice), seules 11 paires sont étudiées (les carrés bleus).

▶ Réussir son appariement : un équilibre délicat entre théorie et technicité, connaissance des données et empirisme

Pour réaliser un appariement de données, il est fréquent de recourir à un outil dédié. On peut mobiliser des packages d'outils statistiques « au cas par cas », mais dans un contexte de production répétée, il est courant de disposer d'outils dédiés à cette opération. Ces outils peuvent être génériques et permettre l'appariement de n'importe quel jeu de données : ils comportent alors généralement un ensemble de paramètres à choisir avec soin pour disposer d'un résultat de bonne qualité.

Au sein des Instituts nationaux statistique (INS), il existe des outils généralistes d'appariement, développés en interne ou plus largement par une autre administration et qui proposent un ensemble d'outils (fonctions de comparaison, de classification, choix

de méthodes de blocage) comme l'outil Relais développé par Istat[14] (Cibella et alii, 2012), G-link développé par StatCan[15] (Chevrette, 2011) ou SPLink utilisé par l'ONS[16] britannique (Cleaton et alii, 2022).

D'autres outils sont plus spécifiques et sont conçus pour répondre à un besoin plus ciblé : par exemple, les outils récemment développés au sein du service statistique public français comme l'outil Rapsodie[17], spécialisé dans l'appariement avec les données fiscales (Jabot et Treyens, 2018) ou l'outil Inserjeunes (Midy, 2021).

Les méthodes utilisées ne sont pas spécifiques, mais l'utilisation régulière sur un certain type de données conduit à des sélections de règles et de paramètres particulièrement adaptés à un jeu de données spécifique : bien coder des informations manquantes au sein d'un fichier, par exemple la modalité SNP (Sans Nom Prénom) dans un fichier administratif, ou prendre en compte les valeurs refuge pour les dates de naissance au 01/01 ou tous les 15 chaque mois[18].

▶ Encadré 2. Les principales étapes d'un processus d'appariement

De nombreux auteurs s'accordent à formaliser les différentes étapes d'un processus d'appariement (*Christen, 2012*), en distinguant :

– une phase de préparation des données (qui comprend l'analyse de la qualité, la normalisation des variables) ;

– une phase de constitution des paires qui tient compte des problématiques de volume et optimise le sous-ensemble constitué ;

– une phase de comparaison des individus au sein des paires qui utilise des fonctions plus ou moins complexes de calcul de similarité ou de distance ;

– une phase de classification qui sélectionne les paires retenues et écarte les paires rejetées ;

– une phase d'évaluation toujours nécessaire qui conduit parfois à modifier les étapes précédentes.

À chaque étape d'un appariement (***encadré 2***), l'expertise et la connaissance des données sont nécessaires et permettent souvent d'améliorer les résultats du processus. Il n'existe pas de solution toute prête, adaptée à tous les appariements. Il est nécessaire de tenir compte de la qualité des données, de leurs caractéristiques pour sélectionner la méthode pertinente et choisir les bons paramètres pour les fonctions de comparaison et de classification.

> ***L'appariement est un processus comportant une dimension de réglages fins d'un ensemble de paramètres.***

Quel que soit l'outil choisi, l'appariement est un processus comportant une dimension de réglages fins d'un ensemble de paramètres, souvent approchés de manière itérative et empirique.

14 Relais : *Record Linkage At Istat* ; Istat : Institut national de statistique italien.
15 G-Link : *Generalized system for record linkage* ; StatCan : Institut national de statistique canadien.
16 SPLink : *probabilistic record linkage at scale* ; ONS : Office for National Statistics, Institut national de statistique du Royaume-Uni.
17 Rapprochement des données sociales, des enquêtes et des impôts.
18 Il s'agit de valeurs du domaine de définition de la variable concernée, attribuées parfois en cas de non-réponse. Ainsi la date du 01/01 est très souvent attribuée lorsqu'un jour de naissance est inconnu.

Comparer les informations, puis sélectionner les paires : le cœur d'un processus d'appariement

La suite de l'article n'est pas une description détaillée des étapes de comparaison et de classification mais donne les grandes lignes des étapes centrales d'un processus d'appariement (Christen, 2012, et Malherbe, 2023 pour une présentation plus complète).

Après avoir identifié un ensemble de paires potentielles, par exemple après une étape de blocage, on procède à la classification de ces paires. Pour chacune d'elles, on compare les deux enregistrements liés. Ceci permet de déterminer s'il s'agit d'une paire d'individus vraiment identiques, vraiment différents ou si un doute subsiste. Il existe une grande variété de méthodes de classification. Elles diffèrent sur un ensemble d'éléments, en particulier le caractère plus ou moins automatique de la définition des paramètres et le recours ou non à un ensemble de paires annotées[19]. L'approche dite probabiliste se caractérise par un degré d'automatisation relativement élevé tout comme l'utilisation du machine learning[20], là où les autres approches nécessitent de définir certains paramètres clés de façon manuelle, grâce à l'expertise et la connaissance des données du statisticien.

Les méthodes déterministes : système de règles…

Cette méthode consiste à apparier les deux fichiers au cours de plusieurs étapes en commençant par des règles strictes et en relâchant progressivement les contraintes. Les individus appariés à une étape ne sont plus considérés pour les étapes suivantes.

> ***Plusieurs étapes en commençant par des règles strictes et en relâchant progressivement les contraintes.***

La première étape est en général un appariement exact : si toutes les variables d'appariement d'une ligne du fichier A sont identiques à celles d'une ligne du fichier B (par exemple, mêmes nom,prénom, date et lieu de naissance), on apparie les 2 lignes. Les étapes suivantes autorisent des légères différences et deviennent de moins en moins strictes. Les autoriser peut se faire soit en excluant simplement un champ de la comparaison, soit en imposant une contrainte plus souple qu'une correspondance exacte. L'idée étant de relâcher progressivement les contraintes, ce sont plutôt les champs les moins discriminants ou ceux contenant le plus d'erreurs qui sont relâchés en premier, par exemple le jour de naissance plutôt que le nom de famille : on perd moins d'information et on retire des données erronées pouvant conduire à un appariement à tort.

19 Ce sont des paires pour lesquelles, après observation humaine le plus souvent, on apporte et parfois commente l'information suivante : individus identiques, individus différents, impossible de décider.

20 L'apprentissage automatique (*machine learning* en anglais) est un champ d'étude de l'intelligence artificielle qui vise à donner aux machines la capacité d'« apprendre » à partir de données, via des modèles mathématiques. Cette approche n'a pas fait ses preuves dans le cadre des appariements en raison du caractère asymétrique du problème de classification d'une part (on cherche n paires parmi un ensemble de taille n^2), et du faible nombre de variables d'autre part (Malherbe, 2023).

▶ ... ou somme pondérée de mesures de similarité...

Une autre approche consiste à calculer des mesures de similarité pour chaque champ identifiant puis à les agréger pour obtenir une mesure globale de similarité pour chaque paire. Une mesure de similarité textuelle est un nombre qui représente la proximité de deux mots ou textes. Les mesures de similarité classiques pour les appariements reposent sur la distance de Levenshtein[21] ou celle de Jaro-Winkler[22] (Herzog et alii, 2007). La similarité globale au niveau de la paire s'obtient ensuite par une somme pondérée des similarités de chaque champ. Les poids associés aux différentes variables sont définis de façon empirique par le statisticien, sur la base de leur caractère plus ou moins discriminant ainsi que de leur qualité.

Cette méthode s'accompagne presque systématiquement de la sélection d'un seuil. Dans ce cas, une paire n'est liée que si sa similarité globale dépasse le seuil.

▶ ... ou appariement par moteur de recherche : un outil efficace pour gérer d'importants volumes de données

Une troisième approche, beaucoup moins conventionnelle, consiste à utiliser un moteur de recherche textuelle, comme Elasticsearch[23] ou Solr. Ce type d'outils est plutôt conçu initialement pour rechercher efficacement de l'information dans un ensemble de textes très vaste, comme l'ensemble des produits sur un site d'e-commerce par exemple. Il peut pourtant se révéler très utile pour réaliser des appariements, en particulier lorsque les fichiers sont très volumineux.

D'un point de vue pratique, un appariement avec un moteur de recherche se déroule de façon assez différente des approches précédentes. La première étape consiste à indexer l'un des deux fichiers, généralement le plus volumineux. Cette opération consiste à stocker les données de ce fichier de façon à ce que ce soit très efficace d'y rechercher de l'information. La structure de données utilisée dans ce cadre s'appelle un index inversé[24]. La seconde étape consiste à effectuer des requêtes, c'est-à-dire rechercher dans cet index une correspondance pour chacun des individus de l'autre fichier. L'outil fournit en sortie l'ensemble des individus correspondant à la requête et les classe grâce à un score de pertinence. Les moteurs de recherche sont très flexibles dans la définition des requêtes, laissant à l'utilisateur une grande marge de manœuvre sur les filtres et les éléments qui entrent dans le score de pertinence. Cette approche peut intervenir en complément d'une première phase d'appariement exact, ce qui permet de réduire la taille des fichiers à traiter.

21 La distance de Levenshtein est une distance, au sens mathématique du terme, donnant une mesure de la différence entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.

22 La distance de Jaro-Winkler mesure la similarité entre deux chaînes de caractères. Il s'agit d'une variante proposée en 1999 par William E. Winkler, découlant de la distance de Jaro qui est principalement utilisée dans la détection de doublons.

23 Pour des précisions sur Elasticsearch, voir l'article « Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers », (Bénichou et alii, 2023).

24 Structure qui donne, pour chaque mot trouvé dans un corpus, la liste des documents où il se trouve.

Cette approche a fait ses preuves puisqu'elle est employée notamment dans le cadre du code statistique non signifiant (Bénichou et alii, 2023). Elle le sera également pour Résil. Il est important de noter qu'elle est plus adaptée à des tâches d'identification, c'est-à-dire lorsqu'on recherche des individus dans un répertoire ou dans un fichier quasi exhaustif sur la population d'intérêt. Il s'agit alors plutôt d'effectuer un grand nombre de recherches « individuelles », indépendantes les unes des autres, lorsqu'un processus d'appariement plus classique raisonnera sur deux ensembles de données pris dans leur totalité.

▶ L'approche probabiliste

L'approche probabiliste provient d'un cadre mathématique établi (Fellegi et Sunter, 1969). Le principe est d'attribuer à chaque paire une probabilité de correspondre à un seul et même individu, calculée grâce à un ensemble de paramètres. Ces paramètres ne sont pas choisis manuellement, mais estimés directement à partir des paires d'individus à apparier (Winkler, 2000). Ils représentent la capacité des différentes variables identifiantes à discriminer les paires. Une correspondance sur le nom de famille constitue ainsi un meilleur indice pour lier une paire d'individus qu'une correspondance sur le genre. Les paramètres du modèle d'appariement probabiliste captent cette information via une probabilité conditionnelle notée u. Ce paramètre représente la probabilité d'observer la même valeur sur un champ donné, sachant que les deux individus de la paire sont différents. Par exemple, la valeur de cette probabilité pour le mois de naissance serait approximativement 1/12.

Par ailleurs, la qualité des différentes variables identifiantes est prise en compte dans les paramètres via la probabilité m. Celle-ci est définie comme la probabilité d'observer la même valeur sur un champ donné au sein d'une paire d'individus identiques. Si les données étaient de parfaite qualité, cette valeur vaudrait toujours 1, mais c'est rarement le cas. On peut alors interpréter la grandeur 1-m comme le taux d'erreur sur un champ donné.

Une fois estimés les paramètres u et m pour chaque variable, il est possible d'obtenir pour chaque paire une probabilité de correspondre au même individu. La règle de décision consiste alors à comparer cette probabilité à un seuil défini par le statisticien pour choisir de lier ou non la paire. La valeur du seuil est à fixer en fonction de l'objectif poursuivi et du type d'erreur acceptable. Si le seuil est élevé, on prend peu de risques sur les paires qui sont liées, mais on risque d'en manquer. À l'inverse, si le seuil est bas, le taux d'appariement est important, mais on a un risque de lier des paires à tort (voir plus bas le paragraphe sur le statut des paires).

La méthode d'appariement probabiliste s'appuie sur les données elles-mêmes pour l'estimation des paramètres u et m. Cela permet de tenir compte du caractère informatif de chaque variable, sans nécessiter une connaissance fine des données. Cependant, cette méthode est très coûteuse en ressources de calcul, ce qui la rend difficile à mettre en œuvre sur les volumes de données habituels dans les processus d'appariement. En matière de qualité, la méthode probabiliste fait jeu égal avec des outils déterministes adaptés aux données (Haag et alii, 2022), mais ne peut rivaliser, du point de vue des ressources informatiques et du temps de traitement sur des données individuelles volumineuses.

Quelle que soit la méthode, un appariement est un processus imparfait. Il convient d'évaluer sa qualité, tant lors de l'élaboration du processus que lors de sa mise en œuvre.

▶ Qualité des appariements et enjeux statistiques

Dans un premier temps, évaluer la qualité permet d'identifier d'éventuelles pistes d'amélioration pour l'appariement. Par exemple, si une sous-population spécifique est mal appariée, il convient d'y porter une attention particulière, par exemple en améliorant le nettoyage et la normalisation des données sur cette sous-population. L'examen manuel de paires permet également de repérer des erreurs fréquentes (telles que des interversions de noms et de prénoms, ou l'utilisation d'anciens noms de communes) et d'adapter l'appariement pour les éviter.

> ***Il est important de s'assurer de la qualité des données à l'issue du processus d'appariement et d'évaluer l'impact de l'appariement sur les résultats de l'étude.***

De nombreux appariements servent à des études statistiques. Il est important de s'assurer de la qualité des données à l'issue du processus d'appariement et d'évaluer l'impact de l'appariement sur les résultats de l'étude. En effet, un défaut de qualité de l'appariement entraîne des incohérences sur les données individuelles (appariements erronés) ou un défaut de représentativité (lorsque les appariements manqués portent plus spécifiquement sur certaines populations).

La qualité des données appariées peut être évaluée de différentes manières, complémentaires. Il est parfois possible de mesurer la qualité du processus lui-même, en s'appuyant sur la proportion de la population appariée ou l'étude des paires retenues ou rejetées. Il est également souhaitable de compléter l'analyse dans une perspective plus statistique, en comparant les populations étudiées avant et après appariement. A-t-on, par exemple, la même structure par âge, la même répartition sur le territoire ?

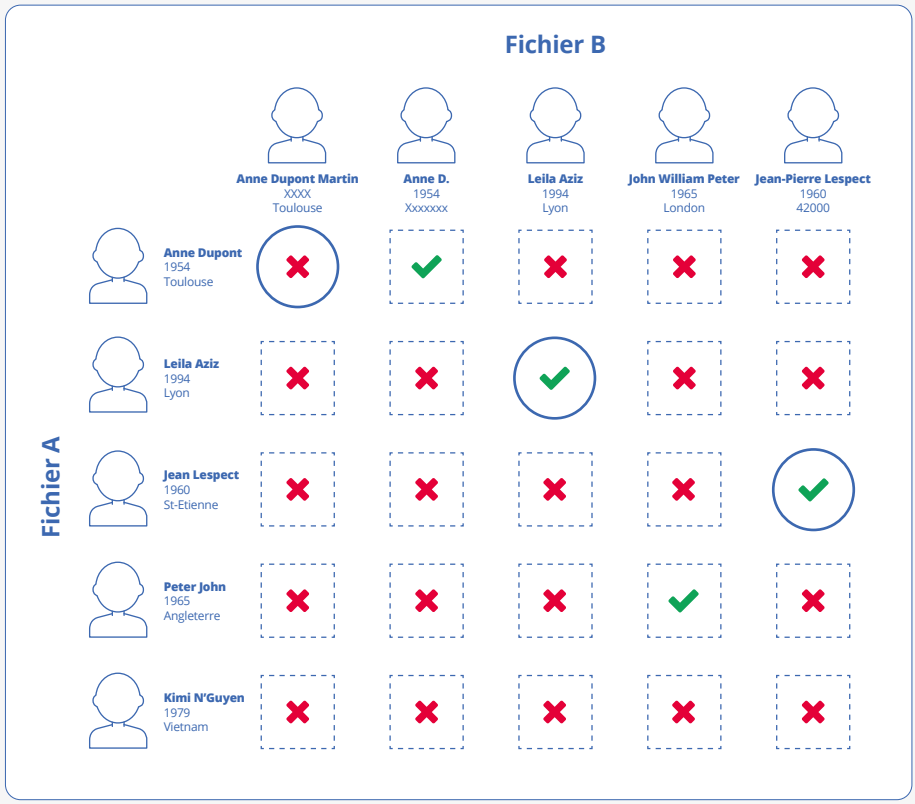
▶ Des mesures d'évaluation fondée sur le statut des paires

Lorsqu'on réalise un appariement, on peut se tromper de deux manières : lier à tort deux enregistrements, c'est-à-dire considérer deux enregistrements comme représentant la même personne à tort (les faux positifs), ou ne pas lier deux enregistrements représentant la même personne (les faux négatifs).

Lorsqu'on dispose du « vrai » résultat, on peut alors caractériser les paires en quatre groupes, en fonction de leur statut réel et du statut prédit à l'issue du processus d'appariement (***figure 3***), tel que :

- Les « bonnes décisions » :
  - les vrais positifs (VP) sont les paires d'individus identiques (concordantes) qui ont été liés par le processus ;

▶ Figure 3 - Statut des paires



Statut prédit	Statut réel		
		Individus identiques (paires concordantes) ✔	Individus différents (paires non concordantes) ✖
	Paire liée ○	Vrais positifs ✔	Faux positifs ✖
	Paire non liée ▯	Faux négatifs ✔	Vrais négatifs ✖

Lecture : 3 paires sont liées, dont 2 sont des vraies positives et 1 est une fausse positive. 22 paires sont rejetées, dont 2 sont des fausses négatives et 20 des vraies négatives.

• les vrais négatifs (VN) sont les paires d'individus différents (non concordantes) qui n'ont pas été liés par le processus ;

- Les « mauvaises décisions » :
  • les faux positifs (FP) sont les paires d'individus différents (non concordantes) qui ont été liés à tort par le processus ;
  • les faux négatifs (FN) sont les paires d'individus identiques (concordantes) qui n'ont pas été liés à tort (en quelque sorte oubliées ou non trouvées) par le processus.

Cette caractérisation est un outil intéressant, mais elle ne constitue pas une évaluation quantitative de la performance. Il est cependant possible de définir de telles mesures à partir des effectifs de chacune de ces catégories.

▶ Deux indicateurs usuels pour décrire la qualité d'un appariement

Lors d'un appariement, les effectifs des classes sont extrêmement déséquilibrés : pour deux fichiers de taille n, le nombre de paires d'individus identiques est proche de n tandis que le nombre de paires d'individus différents est approximativement de n^2. On choisit généralement des mesures comme la précision et le rappel, qui ne font pas appel au nombre de paires négatives (***figure 4***).

La précision se définit de la façon suivante :

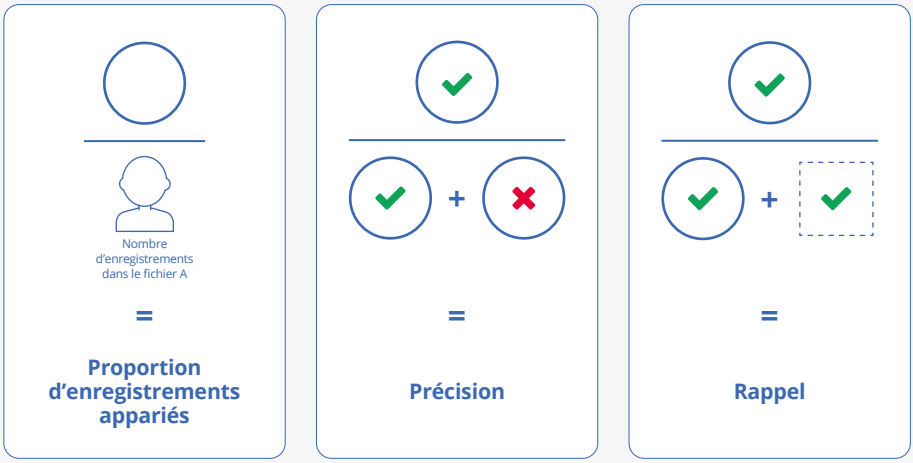
$$\text{Précision} \begin{cases} = \dfrac{\text{Vrais positifs}}{\text{Vrais positifs + Faux positifs}} \\ = \dfrac{\text{Nombre de paires liées ET concordantes}}{\text{Nombre de paires liées}} \\ = \text{Taux de réussite sur les paires liées} \end{cases}$$

Une précision élevée signifie que l'erreur sur ce modèle, lorsqu'il lie une paire, est rare. Cependant, cela ne donne pas d'information sur sa capacité à identifier un grand nombre de paires. Dans le cas extrême, un modèle liant une seule paire à juste titre aurait une précision parfaite de 1. Un tel modèle n'est pourtant pas satisfaisant. C'est pourquoi le rappel intervient souvent en complément de la précision. Le rappel, aussi appelé sensibilité, correspond à la proportion de cas positifs identifiés comme tels par le modèle.

Il se définit comme suit :

$$\text{Rappel} \begin{cases} = \dfrac{\text{Vrais positifs}}{\text{Vrais positifs + Faux négatifs}} \\ = \dfrac{\text{Nombre de paires liées ET concordantes}}{\text{Nombre de paires concordantes}} \\ = \text{Taux d'identification des paires concordantes} \end{cases}$$

▶ Figure 4 - Indicateurs usuels s'appuyant sur le statut des paires



Lecture : Dans la figure 3, la proportion d'enregistrements appariés est de 60 % (3/5), la précision de 66 % (2/3) et le rappel de 50 % (2/4).

Un rappel élevé signifie que les paires sont facilement identifiées par le modèle. Là aussi, dans le cas extrême, toutes les paires possibles seraient sélectionnées, le rappel vaudrait 1, mais le résultat comporterait beaucoup de paires appariées à tort.

▶ L'usage final des données appariées conduit à arbitrer entre précision et rappel

Pour évaluer la qualité d'un appariement, il est nécessaire de définir des objectifs et de réaliser un arbitrage entre les faux négatifs (ou paires concordantes oubliées) et les faux positifs (paires non concordantes acceptées). La notion de qualité est toujours liée à l'usage envisagé. Ainsi, lorsque les techniques d'appariement sont utilisées à des fins opérationnelles (dans le cadre d'opérations de gestion administrative par exemple), on porte une grande attention à chaque résultat individuel et on cherche le plus souvent à éviter les faux positifs (meilleure précision possible). Dans un contexte statistique, la précision est souhaitable, mais on souhaite également éviter un biais de représentativité induit par un défaut de rappel. Il est impossible d'être parfait sur les deux aspects : si on augmente le rappel et le taux d'appariement, alors la précision se dégrade et inversement. Selon l'usage envisagé des données appariées (enrichissement de données, évaluation de la couverture), on effectue un choix sur le niveau de précision attendue.

Les outils nécessaires pour évaluer la qualité d'un appariement par la qualité des paires

Si la proportion d'enregistrements appariés se calcule très facilement pour n'importe quel appariement, ce n'est pas le cas de la plupart des autres mesures de qualité. Celles-ci nécessitent en effet des informations supplémentaires sur les fichiers appariés, le plus souvent un échantillon de paires annotées. Dans le meilleur des cas, on dispose d'un *gold standard* ou fichier étalon. Il s'agit d'un échantillon de paires, représentatif des fichiers à apparier et dont le vrai statut est connu. L'obtention d'un tel échantillon est différent selon chaque situation.

Cependant, dans la majorité des cas, il n'existe pas de *gold standard* et il faut donc ajouter une étape d'annotation manuelle pour qualifier un ensemble de paires, en faisant intervenir un observateur humain : la paire proposée par le processus est-elle une paire « valide » ?, ou le processus a-t-il apparié à tort ?, ou bien encore est-il impossible de trancher ? Cette étape d'annotation manuelle prend du temps et peut donc s'avérer très coûteuse.

D'autres moyens de mesurer au moins partiellement ces grandeurs existent, en utilisant par exemple une sous-population dont on connaît a priori le statut d'appariement attendu, ou en observant la cohérence des données appariées (Doidge et alii, 2020). Lors d'un processus d'appariement, il est fréquent d'annoter manuellement un échantillon de paires, généralement choisies dans un sous ensemble de paires au statut « incertain » ; c'est le cas des paires pour lesquelles la similarité est proche du seuil de rejet, afin d'évaluer le taux de vrais et faux positifs de part et d'autre de ce seuil. Si les paires rejetées sont très majoritairement des faux négatifs, on modifie le seuil afin d'accepter ces paires rejetées à tort, au détriment d'un petit nombre de faux positifs supplémentaires.

Évaluer la qualité d'un appariement par son impact sur les données

Si les indicateurs précédents permettent d'évaluer le niveau de qualité d'un appariement, ils ne sont pas toujours simples à mesurer par l'utilisateur final, d'autant que le processus est parfois exécuté par un service tiers. C'est souvent le cas, à des fins de protection des données individuelles notamment ou en raison de la technicité de certaines opérations (préparation des données, paramétrage de l'outil d'appariement).

Lorsque le service tiers effectue l'appariement, il a connaissance de l'ensemble des variables d'appariement et il est donc en mesure d'évaluer la qualité du processus grâce aux méthodes mentionnées précédemment. Ce n'est en général pas le cas de l'utilisateur final, qui ne dispose pas des variables d'appariement (notamment les traits d'identité). Aussi, il est souhaitable que le service tiers réalisant l'appariement, produise et lui transmette des évaluations de son processus et des indicateurs de qualité associés au résultat de l'appariement. Ces mesures sont importantes, mais elles ne sont pas suffisantes pour évaluer l'impact statistique des erreurs d'appariement.

> ***Évaluer la qualité d'un appariement ne relève donc pas de la seule responsabilité de l'entité qui réalise l'appariement, mais s'appuie sur les travaux complémentaires du ou des services qui exploitent les données appariées.***

En effet, l'utilisateur dispose d'un plus grand nombre de variables, les variables d'intérêt, qu'il utilise pour produire des statistiques (par exemple, diplôme, profession, niveau de revenu, etc.). Ces informations permettent d'évaluer l'impact de l'appariement de façon statistique, par son impact sur la population d'intérêt notamment via des distributions ou statistiques des variables d'intérêt. L'utilisateur peut ainsi vérifier la représentativité de la population appariée par rapport au fichier source : par exemple, a-t-on déformé la structure par âge de la population ?

Ce deuxième niveau d'analyse, plus statistique et tourné vers l'usage est indispensable pour évaluer les éventuels biais induits par le processus d'appariement. Dès lors, si le statisticien a connaissance d'un défaut de représentativité dans la population appariée, il peut mobiliser des traitements statistiques adéquats, comme c'est le cas lors du traitement de toute source statistique.

Évaluer la qualité d'un appariement ne relève donc pas de la seule responsabilité de l'entité qui réalise l'appariement, mais s'appuie sur les travaux complémentaires du ou des services qui exploitent les données appariées.

Conclusion

Les appariements de données se développent ces dernières années au sein de la statistique publique, portés à la fois par une demande croissante de données enrichies et par la disponibilité croissante des données administratives ainsi que l'augmentation des ressources computationnelles.

Ils sont essentiels à différents processus statistiques et seront au cœur du programme Résil[25].

La qualité de ceux-ci est un enjeu important à évaluer tant lors de la réalisation de ces appariements que lors des utilisations faites ultérieurement des données appariées.

S'il existe des outils et des méthodes identifiées pour réaliser des appariements, ils doivent nécessairement être accompagnés de travaux d'analyse des données et d'expertise du statisticien pour sélectionner les paramètres les plus adéquats pour les jeux de données concernés.

25 Voir l'article d'Olivier Lefebvre sur Résil dans ce même numéro.

▶ Fondements juridiques

- 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 22 février 2024]. 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 21 février 2024. [Consulté le 22 février 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460*.

- 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]. Version initiale. [Consulté le 22 février 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033735139/*.

- Décret n° 2024-12 du 5 janvier 2024 portant création d'un traitement automatisé de données à caractère personnel dénommé « répertoire statistique des individus et des logements » (Résil). In : *Site de Légifrance*. [en ligne]. Version initiale. [Consulté le 22 février 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000048866207*.

▶ Bibliographie

- ANDRÉ, Mathias et MESLIN, Olivier, 2022. Patrimoine immobilier des ménages : enseignements d'une exploitation de sources administratives exhaustives. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 107-125. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035944?sommaire=6035950*.

- BÉNICHOU, Yves-Laurent, ESPINASSE, Lionel et GILLES, Séverine, 2023. Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp 64-85. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635825?sommaire=7635842*.

- CHEVRETTE, Antoine, 2011. G-Link: A Probabilistic Record Linkage System. In : *NORC Conference Proceedings*, [en ligne]. Mai 2011. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.norc.org/content/dam/norc-org/pdfs/G-Link_Probabilistic%20Record%20Linkage%20paper_PVERConf_May2011.pdf*.

- CHRISTEN, Peter, 2012. *Data Matching–Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. 5 juillet 2012. Springer. ISBN 978-3642311635.

- CIBELLA, Nicoletta, SCANNAPIECO, Monica, TOSCO, Laura, TUOTO, Tiziana et VALENTINO, Luca, 2012. Record Linkage with RELAIS: Experiences and Challenges. In : *Site de Istat*. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/relais*.

- CLEATON, Mary, HALL, Johanna, SHIPSEY, Rachel, WHITE, Zoe et XHAFERAJ, Kristina, 2022. A case study of using Splink: Census duplicate matching. Proceedings of Statistics Canada Symposium 2022. In : *Plateforme open source GitHub de l'Office for National Statistics*. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://github.com/Data-Linkage/Splink-census-linkage/blob/main/SplinkCaseStudy.pdf*.

- COTTON, Franck et HAAG, Olivier, 2023. L'intégration des données administratives dans un processus statistique – Industrialiser une phase essentielle. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp 104-125. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635829?sommaire=7635842*.

- DEMOTES-MAINARD, Magali, 2019. Élire, un projet ambitieux au service du Répertoire électoral unique. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 58-71. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168399?sommaire=4168411*.

- DOIDGE, James, CHRISTEN, Peter et HARRON, Katie, 2020. Quality assessment in data linkage. National Statistician's Quality Review. In : Site de *UK government*. Mis à jour le 16 juillet 2021. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage*.

- 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]. 29 novembre 2022. Insee. N° N8, pp. 72-92. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665188?sommaire=6665196*.

- FELLEGI, Ivan P. et SUNTER, Alan B., 1969. A theory for record linkage. In : *Journal of the American Statistical Association*. Vol. 64, No 328, pp. 1183-1210. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf*.

- HAAG, Olivier, KOUMARIANOS, Heidi et MALHERBE, Lucas, 2022. Probabilistes ou déterministes, des méthodes d'appariements au banc d'essai du programme Résil. In : *Site des JMS de l'Insee*. [en ligne]. JMS 2022. [Consulté le 20 février 2024]. Disponible à l'adresse : *http://jms-insee.fr/jms2022s07_3/*.

- HACHID, Ali et LECLAIR, Marie, 2022. Sirus, le répertoire d'entreprises au service du statisticien. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 115-130. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665192?sommaire=6665196*.

- HERZOG, Thomas N., SCHEUREN, Fritz J. et WINKLER, William E., 2007. Data Quality and Record Linkage. In : *Researchgate*. [en ligne]. Janvier 2007. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.researchgate.net/publication/220695391_Data_Quality_and_Record_Linkage*.

- JABOT, Patrick et TREYENS, Pierre-Eric, 2018. Proposition d'un nouveau processus d'appariement au Pôle Revenus Fiscaux et Sociaux (RFS). Une application à l'enquête CARE. In : *Actes des journées de méthodologie statistique 2018*. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *http://www.jms-insee.fr/2018/S20_1_PPT_TREYENS_JMS2018.pdf*.

- KOUMARIANOS, Heidi, 2022. Impact du nettoyage des données sur la qualité d'un appariement. In : *Site des JMS de l'Insee*. [en ligne]. JMS 2022. [Consulté le 20 février 2024]. Disponible à l'adresse : *http://jms-insee.fr/jms2022s07_2/*.

- MALHERBE, Lucas, 2023. Appariements de données individuelles : concepts, méthodes, conseils. In : *Documents de travail n° M2023/03*. [en ligne]. 3 juillet 2023. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/7644535*.

- MIDY, Loïc, 2021. Un outil d'appariement sur identifiants indirects : l'exemple du système d'information des jeunes. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 82-99. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398689?sommaire=5398695*.

- RANDALL, Sean M., FERRANTE, Anna M., BOYD, James H. et SEMMENS, James B., 2013. The effect of data cleaning on record linkage quality. In : *BMC Medical Informatics and Decision Making*. Vol. 13, n° 1, pp. 64. [en ligne]. 5 juin 2013. [Consulté le 20 février 2024]. Disponible à l'adresse : *DOI 10.1186/1472-6947-13-64*.

- REDOR, Patrick, 2023. Confidentialité des données statistiques : un enjeu majeur pour le service statistique public. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 46-63. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635823?sommaire=7635842*.

- ROSENBAUM, Paul R. et RUBIN, Donald B., 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. In : *Biometrika*. Vol. 70, No. 1, pp. 41-55. [en ligne]. Avril 1983. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.jstor.org/stable/2335942*.

- Statistique Canada. 2017. Directive sur le couplage de microdonnées. In : *site de Statistique Canada*. [en ligne]. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://www.statcan.gc.ca/fr/enregistrement/politique4-1*.

- WINKLER, William E., 2000. Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. In : *Statistical Research Report Series*. RR2000/05, US Bureau of the Census. [en ligne]. 4 octobre 2000. [Consulté le 20 février 2024]. Disponible à l'adresse : *https://courses.cs.washington.edu/courses/cse590q/04au/papers/WinklerEM.pdf*.

L'accueil des données administratives : un processus structurant

Olivier Lefebvre[*], *Manuel Soulier*[**] *et Thomas Tortosa*[***]

Compte tenu de l'importance croissante des données administratives dans le processus d'élaboration des statistiques, rationaliser leur traitement devient un enjeu essentiel.
Et cela dès l'accueil de ces données ! On les reçoit telles que configurées en fonction de leurs usages administratifs et il faut les transformer en données statistiques, c'est-à-dire en données organisées selon les unités statistiques d'intérêt (individus, ménages, établissements, employeurs, etc.) et les concepts statistiques. Cette phase peut être mutualisée et « découplée » des processus statistiques en aval, offrant ainsi davantage de possibilités d'adaptation, mais aussi de partage d'information. L'enjeu est de construire un dispositif alliant adaptabilité, performance, sécurité et traçabilité.
L'Insee a engagé cette démarche avec l'outil ARC (Accueil-Réception-Contrôle) à partir d'un cas d'usage exigeant : le traitement mensuel d'environ deux millions et demi de fichiers de déclarations sociales nominatives (DSN). En étendant progressivement ses fonctionnalités et ses performances pour l'adapter à de nouvelles données et à de nouvelles contraintes, ARC constitue désormais un composant essentiel du dispositif de production statistique de l'Insee.

Given the growing importance of administrative data in the statistical production process, rationalising the way it is processed is becoming a major challenge.
The work starts as soon as the data arrive! They are received as configured according to their administrative uses, thus requiring to transform them into statistical data, i.e. data organised according to the statistical units of interest (individuals, households, establishments, employers, etc.) and statistical concepts. This phase can be pooled and "decoupled" from downstream statistical processes, which provides greater scope for adaptation, but also for sharing information. The challenge is to set up a system that combines adaptability, performance, security and traceability.
INSEE has adopted this approach with the ARC (Accueil-Réception-Contrôle – receipt, acceptance, control) tool, based on a demanding use case: the monthly processing of around 2.5 million nominative social declarations. By gradually extending its functions and performance to adapt to new data and new constraints, ARC is now an essential part of INSEE's statistical production system.

* Maître d'ouvrage du programme Résil, DSDS, Insee.
olivier.lefebvre@insee.fr

** Chef de projet Informatique sur l'application ARC (Accueil Réception Contrôle), Direction régionale du Centre – Val de Loire, Insee.
manuel.soulier@insee.fr

*** Chef de projet statistique pour le projet « accueil des sources » au sein du programme Résil, DSDS, Insee.
thomas.tortosa@insee.fr

Industrialiser l'intégration des données administratives dans nos systèmes d'information est essentiel, compte tenu de l'importance croissante de ce type de données dans nos processus de production statistique (Cotton et Haag, 2023). Pour relever ce défi, la solution adoptée par l'Insee est une structure d'accueil des sources offerte aux producteurs de statistiques, reposant sur un outil générique moderne et mutualisé.

Au début de l'utilisation de données administratives, ce travail d'intégration se faisait isolément d'une source à l'autre : chacun développait son propre processus, adapté à la source traitée et aux traitements en aval. Ce modèle a fonctionné pendant des décennies. Cependant, dans le courant des années 2010, ces données administratives sont devenues plus nombreuses, plus fréquentes, plus évolutives, et de surcroît susceptibles d'alimenter plusieurs chaînes de production de données. Par ailleurs, des besoins nouveaux ont émergé dans le processus d'accueil qui devait être plus « adaptable », performant, traçable, ouvert, tout en continuant d'assurer un niveau élevé de sécurité.

Les qualités attendues d'un service d'accueil des fichiers administratifs dans un univers statistique : adaptabilité, performance, traçabilité et sécurité

Le service doit être **adaptable**. Il doit prendre en compte au plus vite des changements de contenu ou de format de l'information transmise. De telles transformations sont inévitables, car à l'image des politiques publiques ou de leurs processus de mise en œuvre, la donnée administrative n'est pas figée et évolue en fonction des impératifs de la politique publique qu'elle accompagne.

Ces évolutions doivent pouvoir être appliquées rapidement, sans pour autant être propagées simultanément à tous les fichiers si ceux-ci proviennent d'organismes différents. Les fichiers à accueillir peuvent coexister dans plusieurs versions avec des contenus modifiés selon la date de conception.

Pour répondre à ces besoins d'adaptabilité et de réactivité, le système doit d'une part gérer et accueillir simultanément plusieurs versions de fichiers, et d'autre part proposer des traitements dits « génériques », c'est-à-dire fonctionnant sur tous les fichiers.

> ***Les statisticiens doivent disposer dans la fonction d'accueil, d'outils pour paramétrer les traitements et mesurer l'impact des changements de paramètres sur les données.***

Aussi, les statisticiens doivent disposer dans la fonction d'accueil, d'outils pour paramétrer les traitements et mesurer l'impact des changements de paramètres sur les données.

En outre, cette fonction doit tout à la fois accueillir des fichiers dont le format et le contenu peuvent être amenés à évoluer rapidement et alimenter des chaînes statistiques utilisatrices de données stables sur le long terme. Étendre le principe de généricité de l'outil servant d'interface entre l'accueil et les applications clientes est une solution pour répondre à cette problématique.

La **performance** informatique devient cruciale. En effet, les données sont transmises de plus en plus souvent et doivent être valorisées rapidement. Désormais, des sources extrêmement lourdes sont reçues chaque mois par l'Insee et nécessitent d'être traitées rapidement pour assurer la pertinence des statistiques produites dans des délais toujours plus réduits (demande sociale européenne).

La **traçabilité** est une exigence de qualité fondamentale dans un processus de production. Plus les changements sont nombreux, plus il est indispensable d'en conserver la trace. Cela permet, si nécessaire, de reproduire le traitement, de rendre compte des opérations réalisées et ainsi d'analyser plus facilement les évolutions à mener sur les traitements en aval, et enfin de documenter le processus.

Les données administratives constituent une source importante de données. Elles alimentent depuis longtemps des processus de production mais pourraient servir à des tests de nouveaux traitements statistiques. Avec les nouveaux métiers de la science des données et la montée en puissance à l'Insee des métiers de *data scientists*, il est nécessaire de prévoir des mécanismes permettant d'ouvrir, de manière contrôlée, l'accès aux données brutes statistiques[1] afin de les exploiter de façon innovante.

Enfin, la **sécurité** s'avère primordiale pour ce type de données, celles-ci pouvant contenir des données personnelles imposant une stricte confidentialité, tout comme la protection de leur intégrité ; l'exigence de traçabilité évoquée plus haut participe également de la sécurité d'ensemble du processus et des données traitées.

Prévoir un accueil transverse des données implique donc de centraliser les règles de sécurité dans le cadre d'une gouvernance des données administratives. Il s'agit par exemple de mutualiser et réaliser le plus tôt possible des processus transversaux tels que la « pseudonymisation » des données (Cotton et Haag, 2023), mettre en place une politique et des outils de gestion des droits d'accès, tout en laissant la possibilité à chaque propriétaire d'appliquer des règles spécifiques supplémentaires.

> ***La mutualisation d'un outil d'accueil de données administratives s'est imposée.***

Ces exigences nécessitent des investissements significatifs ; ainsi s'est imposée la mutualisation d'un outil d'accueil de données administratives, capable de gérer des sources de natures et d'origines différentes, et d'alimenter des processus d'exploitation divers. Un tel outil repose sur un découplage de la fonction d'accueil de la donnée et de la fonction de traitement ou d'analyse de celle-ci.

1 Les données brutes statistiques sont des données administratives existant sous un format exploitable statistiquement. Elles sont « brutes » du point de vue du statisticien, car elles n'ont pas encore été traitées pour un usage statistique.

▶ Découpler la phase d'accueil des données de celle des traitements statistiques…

Constituer le service d'accueil des données, c'est penser la phase d'accueil comme une activité à part entière, qu'il faut dissocier des traitements nécessaires à la création d'un produit statistique.

Traditionnellement, le processus d'élaboration d'un produit statistique s'appuyant sur des données externes est de type itératif. Après une étape d'appropriation, le statisticien intègre son fichier pour obtenir un résultat attendu via différentes étapes. Si ces dernières peuvent varier, on retrouve les suivantes :

- si besoin, structurer et transformer le fichier en base de données ;
- renommer les variables afin de pérenniser le traitement ou expliciter le nom de celles-ci ;
- retraiter les variables afin de corriger certaines imperfections (non-réponse, valeurs aberrantes, etc.) ;
- créer des variables statistiques issues d'une ou plusieurs variables parfois modifiées ou agrégées (toutes celles constituant le salaire, tous les revenus d'une catégorie d'agents) ;
- réaliser un produit de diffusion.

Ces phases sont la plupart du temps implémentées par blocs.

▶ … mutualiser l'accueil des données…

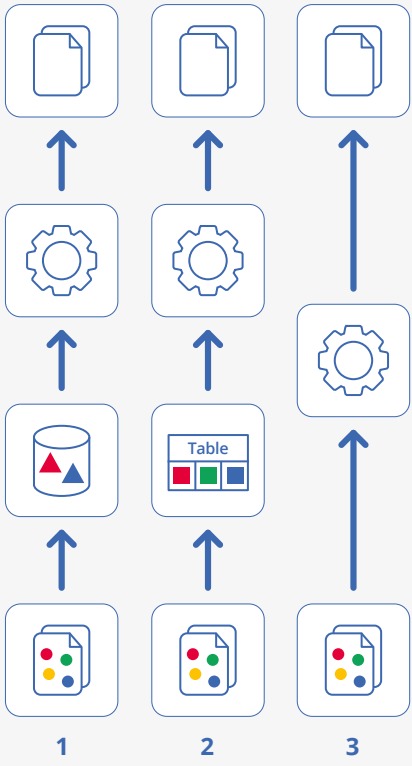
Le statisticien dispose de l'intégralité des données contenues dans le fichier et, par étapes, il va le transformer en un produit statistique. Toutefois, s'il est confortable lorsqu'on le crée, ce processus devient rapidement difficile à maintenir et à exploiter s'il n'est pas développé et implémenté de façon modulaire. En cas de transformations des données, modifier la chaîne de production peut vite s'avérer complexe, si la phase de traitement est adhérente à la phase d'accueil, c'est-à-dire si les traitements s'appuient directement sur les données brutes administratives.

L'idée de découpler l'accueil des données s'est progressivement imposée (***figure 1***). Ainsi, l'Insee a mis en place l'accueil de la Déclaration Sociale Nominative (DSN)[2], et développé un outil dédié (Accueil-Réception-Contrôle, dit ARC), pour alimenter la production de statistiques sur l'emploi et les revenus d'activité (Renne, 2018). C'est également le cas dans le processus Esane (Élaboration de Statistiques Annuelles d'Entreprises) qui s'appuie d'une part sur des données d'enquêtes et d'autre part sur des données fiscales. Plus récemment, le projet de refonte de Fidéli[3] à la suite de la disparition de la taxe d'habitation, a partagé l'application en deux : la première

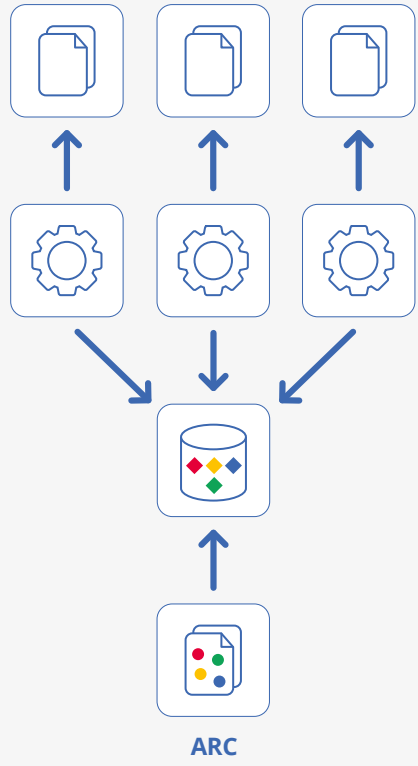
2 La Déclaration Sociale Nominative (DSN) est une déclaration obligatoire, unifiée et en ligne, établie par chaque employeur. La DSN permet d'assurer le recouvrement des cotisations et d'attribuer des droits aux salariés.
3 Fidéli : Fichier Démographique d'origine fiscale sur les Logements et les Individus (Lamarche et Lollivier, 2021).

pour assurer l'accueil des données et la seconde pour effectuer les traitements. Réalisé ex-post au regard des expériences, séparer l'accueil de sources externes des traitements statistiques est un choix d'une utilité majeure pour les statisticiens.

Figure 1 - L'accueil, première étape de la rationalisation des traitements



Lecture : Les données sont accueillies plusieurs fois. Chaque pipeline gère différemment les données en entrée. Le pipeline 1 sélectionne les données utiles et modifie la forme mais aussi le contenu, qu'il stocke en base de données. Le pipeline 2 sélectionne les données et les affine avant de les stocker dans une table exploitable par un logiciel statistique (SAS ou R, par exemple). Quant au pipeline 3, il intègre en bloc le fichier dans ces traitements.



Lecture : Le service d'accueil ARC restructure l'information du fichier, mais sans la modifier. Chaque pipeline peut interroger le fichier pour récupérer les données.

▶ … Pour un usage rationalisé et maîtrisé !

Avec ce découplage, le système d'information est alors plus robuste. L'accueil des sources absorbe en grande partie les chocs exogènes, si ce service a été conçu pour cela[4]. Enlever, ajouter ou modifier des informations peut être traité dans cette phase afin d'alimenter les traitements statistiques en aval de façon quasi identique, en minimisant la maintenance de ce dernier. ARC repose sur ce principe : une phase de conceptualisation des données, c'est-à-dire de transformation des données reçues en un système de données brutes statistiquement exploitables. Cette phase permet de gérer des modifications, comme renommer des variables ou modifier leur contenu lorsque cela est possible.

Par exemple, le fichier des déclarations de revenus POTE[5] de la DGFiP[6] est au format texte ; la lecture n'est donc possible que grâce à un dessin de fichier.

Position		Lg	Numéri-coualpha	Format de lecture	Format d'écriture	Input format	NomPAC	Libellé
247	254	8	9(8)	8.		8.	DADOKZ	SITFAM : DATE DECES DE LA 2042

Ici la variable DADOKZ, qui correspond à la date de décès du référent fiscal du foyer, se lit dans le fichier de la position 247 à 254 (longueur 8). Le format indiqué est un format numérique (longueur 9), alors que la nature de la variable est de type date.

À l'issue du processus d'accueil, la donnée sera mise à disposition dans une variable date_dc (qui explicite le contenu de la variable) au format date (« YYYY-MM-DD »).

Le découplage rend possible et simplifie l'utilisation multiple des données. Lorsque l'accueil des données est intégré dans une chaîne de production d'une application de production statistique, il est difficile, voire impossible, d'ouvrir cet accès aux données à d'autres applications.

Par exemple, les données des déclarations sociales nominatives (DSN) chargées dans ARC étaient initialement destinées à la chaîne structurelle du calcul de l'emploi salarié.

Cette dernière produit les statistiques annuelles de l'emploi en matière de stock ou de répartition par statut ou activité économique. Pour cette utilisation, la chaîne structurelle est « cliente » et consommatrice des données DSN accueillies dans ARC.

Lorsque l'Insee a souhaité utiliser ces mêmes données pour réaliser des estimations conjoncturelles de l'emploi salarié, l'adaptation a été simplifiée par le découplage : à l'instar de la chaîne structurelle, la nouvelle chaîne conjoncturelle a été déclarée « cliente » de ARC. Cela aurait été quasi impossible si l'accueil de la DSN avait été intégré et couplé à la chaîne structurelle.

4 Voir ci-après la partie sur les nouveaux métiers.
5 Fichier « Permanent des Occurences de Traitement des Émissions » : il contient les données relatives aux déclarations des revenus de l'année transmises par les contribuables à la DGFiP au printemps de l'année suivante.
6 DGFiP : la direction générale des Finances publiques est une direction de l'administration publique centrale française qui dépend du ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique.

> ***Isoler le processus d'accueil donne l'opportunité de gérer les droits d'accès aux données par les applications clientes.***

De plus, isoler le processus d'accueil donne l'opportunité de gérer les droits d'accès aux données par les applications clientes, chacune d'elles pouvant sélectionner les données dont elle a besoin parmi celles mises à disposition. Le partage se fait à la source sans avoir besoin de construire une « passerelle » entre les applications.

Enfin, ce découplage permet aussi d'identifier l'accueil des données comme un processus à part entière, avec tous ses avantages. Il est ainsi possible de le découpler des processus clients et donc de le faire évoluer de façon indépendante, ou encore d'ajouter un processus client supplémentaire alimenté à la source. Cela permet aussi de « cibler » les réflexions et investissements pour l'optimiser.

▶ Gérer et « activer » les métadonnées

> ***Sans métadonnées, les données mises à disposition sont inexploitables.***

Sans métadonnées, les données mises à disposition sont inexploitables : en effet, elles sont issues d'une phase de transformation des données administratives en concepts statistiques. Les métadonnées servent à documenter les données produites et sont dans ce cas dites « passives ». Elles sont générées au fil des traitements et permettent de garder en mémoire les opérations sur les données, ainsi que l'utilisation des variables en interne comme en externe. Pour cela, l'Insee dispose d'un référentiel de métadonnées statistiques, RMéS, qui en permet la gestion, le partage et la diffusion (Bonnans, 2019).

Au sein du domaine des enquêtes, les métadonnées sont exploitées en entrée du processus de conception du support de collecte afin de le générer (Cotton et Dubois, 2019). Cette utilisation est possible grâce à la mise en place d'un ensemble d'outils et de services reliés à RMéS. On parle alors de métadonnées « actives », c'est-à-dire qu'elles ont une fonction autre que documentaire dans le processus statistique. Un des enjeux de l'accueil des données administratives est donc de rendre ces méta-données « activables », à l'instar de ce qui se pratique sur les enquêtes ; cela implique de documenter le plus en amont possible les métadonnées issues des fichiers administratifs.

Fournir aux producteurs les métadonnées associées aux données statistiques brutes leur donne les éléments pour documenter leur processus et, in fine, intégrer leurs propres métadonnées au sein de RMèS. Une expérimentation récente sur des données foncières a démontré que la majeure partie des métadonnées saisies dès la phase d'accueil pouvaient être réutilisées sans modifications dans les processus ultérieurs, et ce jusqu'à la réalisation des bases de diffusion.

Par ailleurs, livrer des métadonnées en entrée du processus de traitement permet de mettre en place des métadonnées dites de production : par opposition aux métadonnées dites de diffusion, elles permettent de tracer, configurer voire spécifier les modifications successives des données. On parle alors de lignage des données[7] (Biseul, 2023).

▶ Deux nouveaux métiers : modélisateur de données et intendant des données

Mettre en place un service d'accueil exige de spécialiser cette phase d'accueil des données. Qui dit spécialisation, dit métiers. Au regard des différentes tâches à réaliser, deux nouveaux métiers ont émergé.

Le premier est celui de **modélisateur des données**. Celui-ci conçoit les modèles dans lesquels seront insérées les données administratives pour les utilisateurs. Il transforme un modèle donné, sur lequel il ne peut agir, en un modèle statistiquement exploitable ; ce dernier doit être à la fois robuste aux changements et construit de telle sorte que les utilisateurs puissent utiliser simplement les données pour produire des statistiques. Il lui faut donc être compétent en matière de modélisation mais aussi être à l'écoute des utilisateurs afin que le modèle développé corresponde aux attentes.

Les données étant par nature administratives, il est parfois difficile de les transformer en concept statistique. Le modélisateur peut les segmenter par thématique : on parle alors de partitionnement vertical des données. Une modélisation technique vient compléter la modélisation sémantique.

Par exemple, le fichier des déclarations de revenus (POTE), fourni par la DGFiP, est un fichier très volumineux, tant par son nombre de lignes (45 millions) que par son nombre de variables (600 pour la partie fixe, plus du double pour la partie variable). Dans ce fichier, il est possible d'identifier plusieurs thématiques autour de l'impôt : impôt sur le revenu, contribution sociale généralisée, impôt sur la fortune immobilière, etc. Autant de « thèmes » qui peuvent être isolés. L'existence et les caractéristiques d'un impôt ne sont par nature pas pérennes : la taxe d'habitation en est un exemple. L'avantage de cette modélisation est de construire un modèle autour de thématiques pour chaque impôt. In fine, l'utilisation du fichier sera plus robuste puisque seules les parties affectées par les changements sont modifiées.

Le second métier nouveau est celui d'**intendant des données**. L'intendance des données est un concept pour lequel il existe plusieurs définitions, correspondant à des contours plus ou moins ambitieux. La définition retenue dans cet article est celle de Statistique Canada[8], à savoir : « L'intendance des données, c'est la gouvernance des données en action, c'est-à-dire la déclinaison opérationnelle de la politique en matière de données ». Il s'agit donc de la mise en œuvre effective des règles régissant la collecte, la gestion, la sécurité, la qualité et la diffusion des données au sein d'une organisation.

7 *https://www.journaldunet.fr/web-tech/guide-du-big-data/1516833-data-lineage-definition-principes-et-outils/.*
8 *https://www.statcan.gc.ca/fr/afc/litteratie-donnees/catalogue/892000062020013.*

Ainsi, pour l'intendant, sont requises des compétences en matière d'administration des données puisqu'il assure la réception, le contrôle et la documentation des données qu'il met à disposition.

Il doit aussi posséder des compétences en matière de gestion de données, puisqu'il assure la gestion des droits d'accès aux données et suit les conventions passées avec les fournisseurs (respect des délais et du format de transmission et de conservation des données).

Concernant l'accès aux données, la diversité des utilisations possibles pour une même donnée administrative complexifie la problématique de leur sécurisation. « Pseudonymisées » et transformées en un format statistique, les données sont plus ouvertes et plus « partageables ». Cependant, ce partage doit être organisé et sélectif, selon des finalités précises et en conformité avec le principe de proportionnalité des traitements. L'intendant des données doit appliquer le droit des utilisateurs concernant leur accès aux données.

Enfin, ce spécialiste est en contact direct avec les utilisateurs et les producteurs de données administratives. Il est notamment en première ligne pour régler les problèmes de transmission de données. Des qualités relationnelles sont donc également requises.

L'intendant est légitime pour la collecte de la donnée et sa gestion mais pas sur ses usages. Il faut donc conserver une relation entre le producteur statistique et le fournisseur, apurée des questions de gestion, donc centrée sur le contenu et l'usage des données ainsi que sur leurs évolutions.

▶ Accueil – Réception – Contrôle (ARC), le service informatique d'accueil des données de l'Insee —

L'application informatique Accueil-Réception-Contrôle ARC assure le service d'accueil à l'Insee depuis une dizaine d'années, puisqu'elle existe depuis 2015 avec l'accueil des fichiers de la DSN.

> ***L'application informatique Accueil-Réception-Contrôle ARC assure le service d'accueil à l'Insee depuis une dizaine d'années.***

ARC couvre fonctionnellement certaines des phases du GSBPM[9] (***figure 2***). Ce dernier est un cadre standard des organismes statistiques qui leur permet d'adopter une terminologie commune pour décrire le cycle de vie d'une opération statistique (Erikson, 2020).

9 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.

Figure 2 - Le champ fonctionnel d'ARC au regard du modèle générique de description des processus de production statistique (GSBPM)

Légende :

- Ces phases sont des « prérequis » au bon fonctionnement d'ARC, à gérer en amont et en dehors.
- Ces phases sont couvertes par ARC sans que le statisticien puisse les paramétrer.
- Ces phases sont celles que le statisticien peut paramétrer dans ARC.
- Phase en cours de développement.


Gestion de la qualité / Gestion des métadonnées
Traitement
Analyse
Diffusion
Évaluation
5.1 Intégration des données
5.2 Classification et codage
5.3 Examen et validation
5.4 Édition et imputation des données
5.5 Calcul de nouvelles variables et unités
5.6 Calcul des coefficients de pondération
5.7 Calcul des agrégats
5.8 Finalisation des fichiers de données
6.1 Élaboration du projet de produits
6.2 Validation des produits
6.3 Interprétation et explication des produits
6.4 Mise en place du contrôle de la divulgation
6.5 Finalisation des produits
7.1 Actualisation des systèmes de produits
7.2 Élaboration des produits de diffusion
7.3 Gestion de la publication des produits de diffusion
7.4 Promotion des produits de diffusion
7.5 Gestion de l'assistance aux utilisateurs
8.1 Recueil des produits d'évaluation
8.2 Conduite de l'évaluation
8.3 Adoption d'un plan d'action

Le statisticien s'appuie sur l'opération transverse de gestion des métadonnées pour élaborer la conception des produits finaux de diffusion et la conception de la description des variables utilisées. Pour cela, il utilise directement l'implémentation proposée dans l'application ARC ou lui soumet une modélisation DDI[10] (Dondon et Lamarche, 2023) réalisée via l'outil Colectica Designer[11].

Les phases de conception de la collecte et de conception du cadre et de l'échantillon sont hors du champ de l'application ARC et sont élaborées en accord avec le fournisseur de données.

Les fonctionnalités couvrant les phases de conception du traitement et de l'analyse, de conception de système de production et du déroulement des travaux et les trois phases d'élaboration modélisent la fonction d'accueil dans ARC et posent le cadre des configurations possibles pour le statisticien. Elles ont été construites lors de la conception de l'application et constituent le pipeline d'ARC (***figure 3***).

Le statisticien a toutefois la main pour configurer le déroulement de travaux pour les phases de l'étape « Traitement » ; il peut tester ses configurations sur de petits volumes de données, dans des espaces dédiés, disjoints des espaces de production et appelés « bacs à sable ». Les bacs à sable dans ARC permettent la mise à l'essai du système de production sur un nombre réduit de fichiers sources. Lorsque la mise au point est terminée, le statisticien applique ses configurations sur le flux de fichiers réels et procède à la finalisation du système de production.

Dans le cadre d'un traitement statistique utilisant une fonction d'accueil, ARC couvre les étapes « conception », « élaboration » et « collecte » du GSBPM et partage avec les applications utilisatrices des données, certaines phases de l'étape « Traitement ». Par exemple, ARC intervient sur la phase d'édition et imputation des données[12] pour publier certaines données afin de les rendre exploitables statistiquement (correction de modalités, mise en conformité au format ou aux valeurs attendues), alors que les transformations statistiques au sens métier (imputations de valeurs manquantes ou détection de valeurs aberrantes) sont déportées sur les applications de traitement des données.

Il n'y a pas à ce jour de production automatisée d'un bilan qualité de l'accueil d'une source (nombre d'enregistrements lus, nombre de valeurs erronées, etc.), qui correspond à la phase de recueil des produits d'évaluation du GSBPM. Celle-ci est actuellement prise en charge par le système d'information (SI) client. Le produit ARC devra évoluer afin de proposer une implémentation de ce recueil pour les traitements le concernant (et notamment le respect des normes annoncées), car l'assurance qualité est une composante essentielle de la fonction d'accueil.

10 DDI : *Data Documentation Initiative* 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. Le format DDI repose sur le format XML (voir définition du format XML plus loin).

11 *https://www.colectica.com/software/designer/.*

12 « Data editing and imputation » en anglais.

▶ Figure 3 - Le pipeline de la Déclaration Sociale Nominative (DSN)

▶ Le pipeline de traitement à travers l'exemple de la DSN —

> ***Les déclarations sociales nominatives sont converties en données statistiques par le traitement informatique ARC, via une succession de modules fonctionnels. Chaque module réalise une opération précise.***

La DSN est une déclaration en ligne obligatoire, permettant aux employeurs de transmettre aux organismes de protection sociale les informations relatives aux salariés (Humbert-Bottin, 2018). En décembre 2023, l'Insee a reçu environ 2,5 millions de fichiers d'employeurs avec les données salariales de leurs employés. Il en recevait 1,8 million en 2016.

Les DSN sont converties en données statistiques par le traitement informatique ARC. Ce traitement est architecturé en « pipeline », c'est-à-dire constitué d'une succession de modules fonctionnels, chaque module réalisant une opération précise. Les modules de ARC sont préalablement paramétrés par le statisticien. ARC suit une logique de document : chaque fichier est traité individuellement et indépendamment des autres, et son nom est conservé comme identifiant pour chaque donnée.

Intégrer les données administratives

La phase GSBPM d'intégration des données est couverte dans ARC par les deux premiers modules du pipeline de traitement.

Dans le module **réception**, les documents reçus sont référencés dans ARC. Cette étape est essentielle, d'autant plus pour la DSN pour laquelle le nombre de fichiers reçus au fil de l'eau se compte en millions. Chaque document de la DSN correspond à la déclaration d'une entreprise (au sens large, y compris les employeurs publics désormais). Ce module permet donc de vérifier qu'aucune déclaration n'est en double ou oubliée.

Puis vient le **chargement**. Les fichiers sont d'abord lus, puis leurs données et leur structure définie dans le modèle XML[13] de la DSN sont stockées dans la base de données. C'est lors de cette étape que les fichiers sont associés à leur « norme », laquelle identifie la source et le millésime, selon des règles définies par le statisticien. Ces deux critères de norme et de millésime déterminent quelle sera la suite du traitement.

Les transformer en tables utilisables par le statisticien

Les documents de la DSN sont des fichiers XML dont la structure et l'arborescence sont documentées dans le cahier technique de la DSN[14], de manière très complète. Ce format est adapté aux logiciels de gestion utilisés par les entreprises, mais beaucoup moins à

13 XML : eXtensible Markup Language (XML) est un langage utilisant des balises permettant de représenter des données de manière structurée.

14 Le « cahier technique de la DSN » est un document décrivant de manière détaillée la norme d'échange de la DSN : signification de chaque donnée, domaines de valeurs, contrôles,nomenclatures, structure des messages transmis, etc. *https://www.agirc-arrco.fr/mon-entreprise/specialistes-de-la-paie/declaration-sociale-nominative-dsn/* (Dubrulle et alii, 2023).

la statistique. Le but du pipeline dans ARC est de transformer les données de la DSN en tables exploitables par les statisticiens.

Une fois intégrées, les données des fichiers sont structurées dans le module de **structuration**. Le XML ne fournit que des relations d'ordre hiérarchique entre les différentes données : par exemple, dans la DSN, une entreprise contient un ou plusieurs établissements, desquels dépendent un ou plusieurs individus, qui sont associés à un ou plusieurs contrats de travail. Or certaines règles de gestion ne peuvent pas être représentées par simple relation hiérarchique, comme les lieux de travail. Ces derniers ne dépendent pas hiérarchiquement de l'entreprise, mais il existe une relation entre un contrat de travail et le lieu où il est exercé. Cette relation est documentée formellement dans le cahier technique de la DSN. L'étape de structuration permet donc d'ajouter autour des données, un ensemble de règles de gestion pour créer des liens entre données et faciliter les opérations statistiques.

> ***Le statisticien peut définir des contrôles de conformité, des redressements de forme, des filtres sur les données.***

L'étape suivante est le **contrôle**. Ce module implémente les traitements relevant des phases d'examen et validation et d'édition et imputation. Le statisticien peut définir des contrôles de conformité, des redressements de forme, des filtres sur les données. Ainsi, il est possible de contrôler le format d'un champ, compléter des dates non renseignées, ou filtrer sur les dates de déclaration.

Le module de ***mapping*** vient ensuite mettre au format les données dans le modèle conçu par le statisticien pour l'exploitation statistique : cela correspond au calcul de nouvelles variables et unités.

Les fichiers de la DSN évoluent d'un millésime à l'autre. Le statisticien peut modifier les règles utilisées dans le *mapping* pour gérer ces changements, mais indépendamment du modèle statistique qui peut rester inchangé. Pour les applications en aval, cette constance permet de comparer les données d'une année sur l'autre sans maintenance. Le modèle statistique lui-même peut évoluer, mais marginalement et de façon maîtrisée par le statisticien.

Cette étape de mapping requiert la construction préalable du modèle statistique appelé dans ARC « famille de norme ». Ce modèle permet de mettre en relation des entités statistiques et peut être soit défini directement dans l'application ou importé d'une spécification DDI. Le modèle de la DSN va contenir des tables, comme la table Employeur, avec les champs d'adresse de l'employeur, d'activité principale de l'entreprise ou le numéro d'immatriculation au répertoire Sirene[15] (Siret) ou encore la table Individu avec les champs du prénom ou du pays de résidence.

Tous ces modules de transformation peuvent s'appuyer sur des données externes intégrées par le statisticien, telles que des tables de nomenclatures, des référentiels ou des tables de correspondance. Par exemple, pour la DSN, les codes de pays de naissance des salariés sont recodés selon le code officiel géographique. ARC implémente ainsi en partie la phase de classification et codage.

15 Sirene : Système informatisé du répertoire national des entreprises et des établissements.

Mettre les fichiers de données à disposition pour les traitements ultérieurs

À l'issue du *mapping*, ARC finalise les fichiers de données. Les fichiers de la DSN, fraîchement transformés en tables de données exploitables, sont **mis à disposition** pour être **récupérés par les applications clientes**. L'application permet de gérer les clients de chaque source de données. Chaque livraison de données est horodatée et une seule récupération des données est autorisée par client, ceci afin d'éviter de doublonner des données. Une fois les fichiers téléchargés par les clients déclarés, ARC les supprime après un laps de temps défini par le statisticien.

Les données mises à disposition subissent d'autres transformations par les applications clientes de ARC : restructuration des données par unité statistique, calcul de variables dérivées, etc.

L'intégration des données administratives (Cotton et Haag, 2023) de la DSN passe par toutes ces étapes, de l'accueil réalisé par ARC jusqu'à leur transformation par les applications clientes. Le Répertoire Statistique des Individus et des Logements (Résil[16]) utilise le même schéma d'intégration et ARC a été choisi pour mettre en œuvre la fonction d'accueil.

▶ L'accueil des DSN et la volonté de réutilisation dans le système d'information de l'Insee

L'application informatique ARC a été conçue dans le cadre de la construction d'un Système d'Information sur l'Emploi et les Revenus d'Activité (SIERA) alimenté par diverses sources de données administratives. Il coordonne plusieurs applications et prend en charge l'ensemble des traitements produisant des indicateurs statistiques structurels et conjoncturels sur l'emploi et les salaires. Plus précisément, ARC avait pour objectif, au sein de ce SI, d'accueillir dès 2015 les fichiers mensuels de la Déclaration Sociale Nominative envoyés par la Cnav[17]. Ce fut un challenge pour la maîtrise d'ouvrage et l'équipe de développement : le flux des données à traiter mensuellement était massif et concentré (à l'origine, 1,8 million de fichiers par mois, reçus entre le 18 et le 22 de chaque mois). Les dessins de fichiers n'étaient pas stabilisés et les délais de mise en œuvre courts !

Deux besoins fondamentaux : performance et adaptabilité…

En matière de fonctionnalités, le produit devait répondre à deux besoins a priori orthogonaux : être suffisamment performant pour traiter tous les mois l'ensemble des fichiers en une semaine, (la contrainte telle qu'elle s'exprimait au début) et pouvoir s'adapter rapidement à des changements de ces fichiers. Pour cela, il doit être optimisé en permanence, tant sur le plan de la pertinence que de la rapidité des traitements. Cela implique souplesse et réactivité dans la mise au point des changements liés au contenu ou au format des données source, tout en minimisant l'impact sur les performances.

16 Voir l'article d'Olivier Lefebvre sur Résil dans ce même numéro.
17 Les données DSN sont produites par le GIP MDS (Groupement d'intérêt public pour la modernisation des déclarations sociales) *https://www.net-entreprises.fr/*.

> ***ARC doit être optimisé en permanence, tant sur le plan de la pertinence que de la rapidité des traitements.***

L'option prise est de laisser la main aux statisticiens pour programmer, tester et reprogrammer, dans un « bac à sable » les traitements d'accueil en fonction de l'évolution des sources et des attentes des traitements statistiques en aval. Par ailleurs, l'application est conçue de manière à ce que les réglages mis au point par les statisticiens ne s'appuient que sur des paramétrages des différentes phases de l'accueil, sans influer sur les traitements, et donc sans risque d'altérer les performances du système. Le statisticien peut ainsi se concentrer sur les aspects « métier » et son collègue en charge de l'exploitation informatique fait l'économie d'une phase d'optimisation.

… qui rendent l'application plus pérenne ?

Développer rapidement cette fonction d'accueil a permis la conception du projet ARC en mode agile[18]. D'un point de vue métier, ce besoin de souplesse était également très fort. En effet, les maintenances adaptatives pour absorber les changements de normes des fichiers source réalisées sur d'autres chaînes de traitement du SIERA comme le traitement de la N4DS[19], s'avéraient très coûteuses. ARC devait répondre à ce problème de façon générique pour être réutilisable et pouvoir accueillir d'autres sources que la DSN.

ARC fut ainsi déployé en 2015 pour l'accueil des fichiers XML de la DSN puis réutilisé dans le SIERA pour accueillir les fichiers des Déclarations Annuelles de Données Sociales (DADS), qui devaient « coexister » jusqu'en 2021 avec la DSN. L'application a également été utilisée pour l'accueil de fichiers déjà produits par l'Insee, dans une optique de mise dans un format commun. Dans cette première version, le produit était déjà capable de prendre en charge l'accueil de fichiers XML, CSV[20] et clé-valeur[21].

Les contraintes auxquelles il a fallu faire face pour développer un outil d'accueil de la DSN ont conduit à lui conférer les « bonnes propriétés » pour le désigner comme l'outil central d'un processus d'accueil des sources administratives découplé des traitements en aval. L'usage d'ARC s'est alors progressivement développé.

▶ Premières utilisations hors du cadre initial

Face à un changement de norme pour les liasses fiscales utilisées dans le processus Esane (Élaboration de statistiques annuelles sur les entreprises), il a été envisagé d'utiliser ARC plutôt que de développer un nouveau système spécifique à ce processus.

18 L'agilité a pour objectif d'orienter les efforts vers ce qui a le plus de valeur pour l'utilisateur, en s'adaptant aux changements à moindre coût.
19 N4DS : Norme pour les Déclarations Dématérialisées Des Données Sociales, utilisée par les DADS.
20 CSV désigne un format de fichiers dont le rôle est de présenter des données séparées par des virgules. Il s'agit d'une manière simplifiée d'afficher des données afin de les rendre transmissibles d'un programme à un autre.
21 Le format de stockage clé-valeur fait correspondre des clés (par exemple des rubriques métiers) avec des valeurs.

L'instruction technique menée a mis plusieurs points en évidence : la nécessité d'une évolution fonctionnelle d'ARC, puis l'intérêt de cet outil pour les utilisateurs et enfin le fait que l'utilisation de l'application représentait l'option la moins coûteuse pour prendre en charge cette évolution.

La couverture fonctionnelle d'ARC n'était cependant pas complètement suffisante pour prendre en charge les fichiers hiérarchiques complexes que sont les fichiers fiscaux. La normalisation du processus d'accueil proposé par l'application a néanmoins permis de réaliser les évolutions rapidement et d'enrichir l'offre du produit. ARC a donc pour la première fois été réutilisé hors SIERA en 2019 pour l'accueil des fichiers fiscaux dans Esane.

En 2020, dans le cadre d'un groupe de travail du système statistique européen sur le partage d'outils statistiques, ses fonctionnalités ont une nouvelle fois été étendues et cela dans deux directions. La première permettait de déployer l'application sur des infrastructures conteneurisées[22], permettant notamment une scalabilité[23] accrue. La seconde rendait possible l'appel des étapes d'accueil de fichier en mode web-service. ARC devenait alors capable de traiter des invocations à la demande, unitaires, en complément des traitements de masse initialement développés. SIRENE4 a ainsi intégré l'application dans ce mode d'utilisation de machine à machine pour contrôler de façon automatique la conformité des liasses d'immatriculation provenant du Guichet Unique[24] (Alviset, 2020).

ARC a donc évolué progressivement, pour devenir une application robuste et performante (***encadré***).

▶ Encadré. Caractéristiques techniques et performances.

ARC est un ETL opensource (*Extract Transform Load*) développé par l'Insee. Le code source de l'application est hébergé sur l'espace github inseeFr : *https://github.com/InseeFr/ARC*

ARC propose un module web, un module batch et un module web-service. Chaque module est conteneurisé avec Docker* et déployable de façon autonome selon les besoins métiers. Les conteneurs sont disponibles sur *dockerhub* : *https://hub.docker.com/u/inseefr*

Le module web propose une interface homme machine pour paramétrer, lancer des traitements sur des fichiers et piloter éventuellement les traitements massifs réalisés en batch. Le module batch permet de traiter des flux massifs de fichiers et d'assurer la reprise en cas d'erreur. Le module web-service expose un service de récupération de données et un service de traitement de fichier unitaire.

ARC peut traiter des fichiers XML, clé-valeur et de type texte, aux formats CSV, délimités ou positionnels.

Les performances d'ARC ont considérablement progressé depuis l'origine : les premiers chargements mensuels de la DSN réalisés en 2015 duraient 9 jours contre 60 heures aujourd'hui alors même que le volume de données à traiter est deux fois plus important. Cette amélioration significative a notamment été permise par le découplage entre stockage des données et traitements.

Les traitements dans ARC sont réalisés sur des bases PostgreSQL. L'instance dédiée au chargement de la DSN dispose actuellement d'une seule base de données avec 32 CPU et 32 Go de RAM.

La dernière version d'ARC est scalable horizontalement. Plutôt que de disposer d'une unique base de données avec beaucoup de ressources, l'application peut utiliser plusieurs petites bases de données en parallèle. Cette architecture permet d'éviter les problèmes de capacité de traitement inhérents à l'utilisation d'une seule machine : le temps de traitement décroît proportionnellement avec le nombre de bases de données dédiées à l'application.

* *Docker est un système permettant de créer, partager et exécuter des conteneurs.*

22 Une infrastructure de conteneurs permet d'automatiser le déploiement, la mise à l'échelle et la gestion des conteneurs. Un conteneur est un environnement d'exécution contenant tous les composants nécessaires (code, dépendances et bibliothèques) pour exécuter le code de l'application sans utiliser les dépendances de la machine hôte.
23 Capacité à s'adapter à des évolutions importantes du volume de données à traiter.
24 Le Guichet électronique des formalités d'entreprises (Guichet unique) est un portail internet sécurisé, auprès duquel toute entreprise est tenue de déclarer sa création ainsi que différents événements de vie, depuis le 1er janvier 2023.

Le changement d'échelle : l'organisation de la mutualisation

L'extension progressive des fonctionnalités et des usages d'ARC ainsi que son caractère incontournable dans différentes productions ont conduit à une prise en charge adaptée à ces enjeux. Il s'agissait d'une part de piloter la mise en œuvre et le déploiement d'investissements transverses (amélioration des performances, traitement des métadonnées, maintien en conditions opérationnelles et de sécurité, etc.), et d'autre part d'animer la communauté des utilisateurs (communication sur le produit et ses évolutions, recueil et priorisation des besoins, formations, etc.). Ces actions sont essentielles pour un produit mutualisé comme ARC afin de conserver un champ fonctionnel et de prendre en compte les besoins des divers utilisateurs.

En 2021, le programme Résil est devenu maîtrise d'ouvrage du produit ARC du fait de son positionnement central dans le système d'information, mais aussi en raison de la diversité des sources qu'il accueille.

Conclusion

Le monde de la donnée, comme celui de l'informatique, est en perpétuelle évolution. Pour répondre aux défis informatiques, les entreprises ont développé des stratégies comme la mise en place de l'agilité et du DevOps[25]. Cette idée a été reprise par l'univers de la data avec le développement du DataOps[26], qui vise notamment à concilier automatisation, reproductibilité, interactivité et traçabilité en matière de traitement des données, tout en réunissant différents métiers de la data autour d'outils communs. L'Insee, en proposant dans ses développements l'agilité, intègre déjà beaucoup de recettes du DataOps. Le service d'accueil que rend ARC s'inscrit pleinement dans cette démarche, de par la souplesse et le découplage des traitements qu'il met en œuvre.

ARC est une application qui a su évoluer au cours du temps en fonction de besoins multiples et variés. Une application désormais facile à déployer, adaptée à un chargement souple et performant de données externes, en amont d'applications statistiques. Une application qui devient centrale dans la mise en œuvre de la stratégie d'utilisation des données administratives visant à la fois la diversification des sources, la rapidité de leur traitement, la capacité à les mettre au service de processus statistiques différents, et cela en toute sécurité.

L'ouvrir vers des utilisations plus exploratoires, réalisés en « self service » par des statisticiens, peut permettre d'explorer plus facilement le potentiel de nouvelles sources de données, au service de l'innovation statistique.

25 Le DevOps est un mouvement en ingénierie informatique et une pratique technique visant à l'unification du développement logiciel et de l'administration des infrastructures informatiques.
26 Le DataOps est une méthode automatisée pour améliorer la qualité et réduire le temps de cycle de l'analyse des données. *https://dataopsmanifesto.org/fr/.*

▶ 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 16 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497083?sommaire=4497095*.

- BISEUL, Xavier, 2023. Data lineage : définition, principes et outils. In : *Journal du Net*. [en ligne]. 28 février 2023. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.journaldunet.fr/web-tech/guide-du-big-data/1516833-data-lineage-definition-principes-et-outils/*.

- 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 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168396?sommaire=4168411*.

- COTTON, Franck et DUBOIS, Thomas, 2019. Pogues, un outil de conception de questionnaires. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 17-28. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254216?sommaire=4254170*.

- COTTON, Franck et HAAG, Olivier, 2023. L'intégration des données administratives dans un processus statistique - Industrialiser une phase essentielle. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 104-125. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635829?sommaire=7635842*.

- DUBRULLE, Bertrand, ROSEC, Olivier et SUREAU, Christian, 2023. Une norme d'échange pour alimenter des référentiels et en assurer la qualité. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N°N9, pp 126-146. [Consulté le 18 juin 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635835?sommaire=7635842*.

- DONDON, Alexis et LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N°N9, pp 86-103. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.

- ERIKSON, Johan, 2020. Le modèle de processus statistique en Suède – Mise en œuvre, expériences et enseignements. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee. N° N4, pp. 122-141. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497085?sommaire=4497095*.

- 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 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3647025?sommaire=3647035*.

- 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 31 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- RENNE, Catherine, 2018. Bien comprendre la déclaration sociale nominative pour mieux mesurer. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 35-44. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3647029?sommaire=3647035*.

PRÉSENTATION DU NUMÉRO N11

Vous avez aimé découvrir l'histoire de la statistique publique dans le numéro N9 du Courrier sur le thème « Statistiques publiques et débat démocratique (1946-1987) », alors n'hésitez plus et lisez le 2e épisode. Peu à peu, la construction européenne façonne la production de statistiques publiques, une nouvelle ère d'ouverture et de gratuité s'ouvre et de nouveaux thèmes apparaissent.

Comment faciliter la navigation dans l'océan de données mis à disposition sur le site de l'Insee ? C'est le sujet de l'article suivant, qui pointe les indispensables métadonnées, l'importance d'un catalogue, et les possibilités d'accès à des « hypercubes ».

Les dessous de la quantification dans le secteur de l'énergie sont ensuite dévoilés, au moment où la transition écologique est toujours plus d'actualité.

Les quatre autres articles de ce numéro constituent un dossier, organisé autour du Répertoire statistique des individus et des logements (Résil).

Si le premier présente le projet Résil dans son ensemble, avec ses principes directeurs, le second nous révèle la démarche de concertation engagée par l'Insee, pour assurer la légitimité de ce répertoire, et répondre aux enjeux juridiques et éthiques. Deux étapes du processus Résil nécessitaient une attention particulière. Ainsi le troisième article du dossier porte-t-il sur les appariements : finalités, méthodologie, mise en pratique et évaluation de la qualité. Enfin, le dernier papier s'attelle de façon pédagogique à expliquer l'outil ARC (accueil-réception-contrôle) : appliqué dans un premier temps à la déclaration sociale nominative (DSN), il a été généralisé pour le projet Résil.

ISSN 2107-0903
ISBN 978-2-11-162412-2

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE