N°8

COURRIER DES STATISTIQUES

Novembre 2022

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE

Rédaction en chef
Odile Rascol
Contribution
Insee : Lionel Espinasse, Ali Hachid, Xavier Helfenstein, Marie Leclair, Pascal Rivière, Valérie Roux, Willy Thao Khamsing
Cnav : Joseph Préveraud de Vaumas
DGCL : Luc Brière
Directeur de la publication
Jean-Luc Tavernier
Directeur de la collection
Pascal Rivière

Rédaction
Catherine Fresson-Martinez, Pierre Glénat, Marine Le Roux, David Martineau, Odile Rascol, Pascal Rivière
Composition
Agence LATITUDE Nantes
5, rue Jacques Brel
« Les Reflets » Bâtiment A
44800 SAINT-HERBLAIN
0407/22
02 51 25 06 06
www.agence-latitude.fr
Photo de couverture
Adobe Stock®

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

www.insee.fr

Courrier des statistiques N8

SOMMAIRE

PRÉSENTATION DU NUMÉRO

Avec cette nouvelle édition, le Courrier des statistiques livre son huitième numéro. La revue se donne une fois de plus pour ambition d'aborder, avec une tonalité qui se veut pédagogique, l'univers des grandes problématiques auxquelles se confronte la *statistique publique*.

Le statisticien, débutant ou expert, comme le citoyen intéressé par la « fabrique de la donnée », y trouvera de nouveau un large panel d'articles permettant de mieux comprendre cette galaxie et de témoigner de la volonté des services de la statistique publique d'innover pour remplir leur mission. Les articles y sont écrits avec soin par des auteurs et autrices qui décrivent notre capacité collective à s'adapter tant sur les méthodes et outils, que sur des questions institutionnelles ou encore juridiques. Témoin des travaux menés dans la sphère de la statistique publique, la revue reste attentive aux pratiques extérieures, en France comme à l'étranger, à les placer dans leur contexte pour nourrir nos réflexions, et, finalement, asseoir nos recherches sur des bases solides.

Ce huitième numéro consacre cinq de ses sept articles à la galaxie des répertoires. Étonnement, le Courrier des statistiques n'avait jusque-là jamais réglé aussi précisément sa focale sur leurs propriétés, sur la manière de les construire et de les faire vivre. Pourtant, ils occupent le centre de gravité de bon nombre de systèmes d'information développés par le statisticien. Ils endossent par ailleurs de multiples rôles en tant que référence centrale, voire opposable, dans des processus de gestion, notamment administratifs. Ils s'adressent tant aux utilisateurs individuels qu'institutionnels aux intérêts parfois divergents, mais qui convergent cependant toujours sur leurs exigences d'y trouver un niveau de service élevé mêlant qualité du contenu et interopérabilité.

Si le choix a été fait de consacrer une large place aux répertoires et référentiels, le Courrier s'arrête, en ouverture de ce numéro 8, sur l'univers des statistiques dédiées aux collectivités locales mais aussi sur l'enquête TeO qui explore de manière singulière comment les origines des immigrés ou des enfants d'immigrés influent sur leurs trajectoires et conditions de vie.

C'est donc la deuxième édition de l'enquête TeO (Trajectoires et origines) qui ouvre la revue. ***Willy Thao Khamsing*** dissèque cette opération co-réalisée par l'Insee et l'Ined dès 2008-2009 et rééditée en 2019-2020. La réédition de TeO va beaucoup plus loin qu'une simple actualisation puisque de nombreuses innovations dans la conception, le protocole ou encore la méthodologie de l'enquête y ont été introduites.
L'Insee et les services statistiques des ministères sont chargés de produire des statistiques visant à mesurer la diversité, fondées sur des données objectives et fiables relatives à la situation des immigrés, des enfants d'immigrés mais également à leur parcours. Dans un contexte où les thèmes de l'immigration et de la diversité de la population française sont au cœur du débat public, objectiver ces éléments reste essentiel.
L'auteur nous plonge au cœur de l'enquête et décrit avec précision les méthodes sur mesure déployées pour constituer les échantillons et redresser la non-réponse.

Le second article propose une analyse à 360 degrés du Service statistique ministériel (SSM) collectivités locales. L'auteur, ***Luc Brière***, responsable du SSM, y analyse cet appareil statistique complet et centralisé dédié aux collectivités. Répondant au départ principalement

à des besoins de suivi en matière de démographie, de finances des collectivités ou des structures intercommunales, cet appareil a progressivement évolué afin de répondre à des demandes de connaissances plus fines. Les données produites s'inscrivent dans des processus complexes de production, de validation et d'enrichissement de sources principalement administratives.

Les répertoires sont à l'honneur dans les cinq articles qui suivent. ***Pascal Rivière*** ouvre le bal dans un article général et très documenté. Il y définit les répertoires, ces « référentiels indispensables et pourtant méconnus » comme une source d'information reconnue, contenant des données « maître » dans laquelle les utilisateurs viennent puiser. Il y détaille les cinq propriétés fondamentales (centralité, qualité, stabilité, unité de sens et interopérabilité) qui les caractérisent, souligne également l'aspect dynamique d'un répertoire et met en évidence le système d'information qui gravite autour de celui-ci. On retrouve, à divers degrés, ces différentes notions dans les articles qui suivent.

Les deux articles suivants nous font pénétrer dans les constellations mêlées du Répertoire national d'identification des personnes physiques (RNIPP) et du système national de gestion des identifiants (SNGI).

Pour commencer, ***Lionel Espinasse et Valérie Roux*** nous présentent le RNIPP, répertoire fondamental géré à l'Insee, qui contient l'état civil de toutes les personnes nées en France ou ayant vécu en France. Au moment de son immatriculation dans le répertoire, chacun se voit attribuer un numéro d'identification (le NIR), plus connu sous le mal nommé « numéro de sécurité sociale ». Géré par l'Insee, le RNIPP est alimenté à partir des actes d'état civil transmis par les communes pour les personnes nées en France et par la Caisse nationale d'assurance vieillesse (Cnav) pour les personnes nées à l'étranger. Jumelé au SNGI géré par la Cnav, il est une pièce maîtresse du système social français. L'accès aux informations personnelles contenues dans le RNIPP est strictement encadré par des textes réglementaires. Il est utilisé principalement pour la certification des états civils ou encore pour vérifier le statut vivant ou décédé des personnes. Le RNIPP est aussi un répertoire pivot au service des autres répertoires comme Sirene ou le Répertoire électoral unique. Initialement cantonné à la sphère sociale, le RNIPP joue un rôle structurant dans le système administratif français. Ces dernières années, il enregistre près de 30 millions de demandes d'identifications par an. Plus récemment, l'arrivée de *FranceConnect* en a encore accru les usages et les standards de qualité et de disponibilité.

Le SNGI est le référentiel des identités, pour les besoins des organismes de la protection sociale. Créé en 1988 par la Cnav, il traite les états civils et le NIR des ayants droit de la sécurité sociale. Au fil du temps, il s'est imposé comme référentiel socle notamment parce qu'il permet l'attribution d'un NIR aux individus nés hors de France. Construit à partir des fichiers des assurés du régime général de retraite de la Cnav, il a progressivement été synchronisé avec le RNIPP de l'Insee.
Joseph Préveraud de Vaumas décrit dans son article son enrichissement permanent ainsi que ses fonctionnalités de consultation et de recherche d'identité.

Avec l'article suivant, on quitte le domaine des individus pour s'intéresser aux entreprises. Avec ***Ali Hachid et Marie Leclair*** nous parcourons la constellation Sirus, outil indispensable au statisticien d'entreprises qui doit pouvoir s'appuyer sur un répertoire pour établir sa base de sondage, comparer ses données d'enquêtes ou administratives à des valeurs de référence. En France, le système statistique sur les entreprises s'est appuyé sur le répertoire Sirene. Ce dernier, conçu pour des besoins administratifs et géré par l'Insee, a dû concilier au fil du temps des réponses à des besoins contradictoires, venant alourdir la gestion. C'est pourquoi, depuis dix ans, l'Insee a choisi de développer un répertoire des unités statistiques, Sirus. Adossé à Sirene pour en exploiter la fraîcheur mais développant de nouveaux concepts utiles aux statisticiens (l'entreprise au sens économique, la cessation économique, etc.), Sirus fournit à la statistique d'entreprise une référence commune.

Last but not least, l'article de ***Xavier Helfenstein*** nous plonge dans une singularité de l'appareil statistique français à travers l'analyse de la base permanente des équipements qui recense et géolocalise les équipements, services et infrastructures accessibles à la population sur l'ensemble du territoire chaque année. Héritière de l'inventaire communal qu'elle a éclipsé, la constitution de cette base s'appuie aujourd'hui sur la collecte de sources, principalement administratives, riches et bien moins onéreuses à produire et à mettre à jour. Si son processus de production est simple, la multitude des sources intégrées et leur hétérogénéité nécessitent d'adapter les traitements en continu. Sa refonte, actuellement en cours, s'appuie sur une démarche qualité ambitieuse dont l'objectif principal est de rationaliser les travaux et de mieux prendre en compte les demandes des utilisateurs. Enfin, gérer un système complet de métadonnées pour faciliter les échanges avec les producteurs de données et les utilisateurs constitue une rénovation attendue.

Odile Rascol
Rédactrice en chef, Insee

Une nouvelle enquête Trajectoires et Origines, dix ans après la première

Évolutions et innovations dans le protocole de collecte

*Willy Thao Khamsing**

Trajectoires et Origines est une enquête de référence sur les thèmes de l'immigration et de la diversité de la population française. En explorant l'histoire migratoire des personnes ou de leurs parents, en décrivant leurs parcours (scolaire, professionnel, résidentiel, familial), la transmission des langues et de la religion dans le cadre familial, cette enquête cherche à étudier comment les origines géographiques, nationales, culturelles ou sociales sont susceptibles d'influer sur les conditions de vie et les trajectoires des individus.

La première édition de Trajectoires et Origines (TeO) a été réalisée par l'Insee et l'Ined en 2008-2009 et largement exploitée ; dix ans plus tard, dans un contexte social qui a évolué, la mise à jour des connaissances sur ces sujets devenait nécessaire. La réédition de 2019-2020 va plus loin qu'une simple actualisation des données : elle laisse la place à des évolutions et des innovations dans la conception, le protocole et la méthodologie de l'enquête.

Trajectoires et Origines *is a reference survey on the themes of immigration and the diversity of the French population. By exploring the migratory history of individuals or their parents, by describing their educational, professional, residential and family backgrounds, and the transmission of languages and religion within the family, this survey seeks to study how geographical, national, cultural or social origins are likely to influence the living conditions and trajectories of individuals.*

The first edition of Trajectories and Origins (TeO) was carried out by INSEE and INED in 2008-2009 and has been widely used; ten years later, in a social context that has changed, it became necessary to update knowledge on these subjects. The 2019-2020 reissue goes further than a simple update of the data: it leaves room for evolutions and innovations in the design, the protocol and the methodology of the survey.

* À la date de rédaction de l'article, chef de projet statistique de l'enquête TeO2, DSDS, Insee, *willy.thao-khamsing@insee.fr*

> ***La France est de longue date un pays d'immigration.***

La France est de longue date un pays d'immigration. En 2020, 10 % de la population résidant sur le territoire national était immigrée, c'est-à-dire née étrangère à l'étranger (voir définitions en ***encadré 1*** et *(Insee, 2022a)* : 48 % des immigrés vivant en France sont nés en Afrique, 32 % en Europe, 14 % en Asie, 6 % en Amérique et Océanie. Un peu plus d'un tiers de ces 6,8 millions d'immigrés ont acquis la nationalité française. Les descendants d'immigrés nés en France et y résidant en 2020 sont un peu plus nombreux : 7,4 millions étaient comptabilisés sur le territoire[1], dont près de 5 millions avaient plus de 18 ans.

Quel est le parcours familial, résidentiel et professionnel des immigrés depuis leur entrée en France ? Y a-t-il une persistance de l'influence des origines sur les trajectoires sociales des deuxième et troisième générations ? L'intégration des immigrés en France diffère-t-elle selon les origines ? Existe-t-il des discriminations à l'encontre des immigrés et des descendants d'immigrés et, si oui, quelles formes prennent-elles ? Autant de questions qui font l'objet de nombreux débats, voire d'idées reçues.

Comme souvent dans ce genre de contexte, la statistique publique et la recherche sont mises à contribution pour apporter des informations fiables, détaillées et actualisées, permettant de combler un déficit de connaissance et d'objectiver le débat. On peut dans un premier temps vouloir mesurer la diversité d'une société. Ainsi, certains

▶ Encadré 1. De qui parle-t-on dans l'enquête TeO ?

Un **immigré** est une personne née étrangère à l'étranger et résidant en France. Les personnes nées françaises à l'étranger et vivant en France ne sont donc pas comptabilisées. À l'inverse, certains immigrés ont pu devenir français, les autres restant étrangers. Un individu continue à être immigré même s'il acquiert la nationalité française. Les personnes nées françaises à l'étranger et vivant en France ne sont pas immigrées.

Un **descendant d'immigrés de 2e génération** est une personne née en France ayant au moins un parent immigré. Cette appellation ne concerne donc pas les personnes ayant migré étant enfant avec leurs parents.

Un **descendant d'immigrés de 3e génération** est une personne née en France, ayant au moins un grand-parent immigré et dont aucun parent n'est immigré. Elle peut avoir de un à quatre grands-parents immigrés.

Le terme de 1re génération désigne donc les immigrés (ils sont la première génération à vivre en France). Par suite, les descendants d'immigrés sont dits de 2e ou 3e génération selon qu'ils ont des parents ou des grands-parents immigrés.

Un **natif d'outre-mer** est une personne née dans un territoire d'outre-mer. Un **descendant de natif d'outre-mer** est une personne née en France ayant au moins un parent natif d'outre-mer.

Les **personnes sans ascendance migratoire directe** sont celles qui ne sont ni immigrées, ni descendantes d'immigrés (de 2e génération), et dans le cas spécifique de TeO, ni natives d'outre-mer, ni descendantes de natifs d'outre-mer.

1 Hors département de Mayotte.

pays réalisent ce qui est parfois appelé des « statistiques ethniques » : c'est le cas par exemple des recensements américain, canadien, brésilien, irlandais et britannique, qui comportent des questions d'appartenance à un groupe « ethno-racial ». Le terme de « statistiques ethniques » ne se voit toutefois pas attribuer le même sens par tous *(Le Minez, 2020)*. En France et conformément au cadre juridique en vigueur, l'Insee et les services statistiques ministériels produisent de nombreuses statistiques fondées sur des données objectives : pays de naissance, nationalité à la naissance, nationalité actuelle, autant d'informations disponibles dans de nombreuses enquêtes et dans le recensement de la population. Muni de ces critères, on peut commencer à identifier et préciser d'éventuelles discriminations ou inégalités de situations auxquelles font face certains groupes de population.

Mais, tout en respectant le cadre juridique, il est aussi possible de se baser sur des informations plus subjectives, en recueillant le « ressenti d'appartenance » des individus *((Le Minez, 2020) et antérieurement (Insee, 2016))*. En 2008-2009, une enquête de la statistique publique et de la recherche, baptisée *Trajectoires et Origines* (TeO), a comporté pour la première fois à grande échelle des questions sur le ressenti d'appartenance. En effet, outre le pays de naissance et la nationalité à la naissance des parents, étaient abordés le lien avec le pays d'origine, la religion, les langues, l'image de soi et le regard des autres. Cette enquête comportait également des questions subjectives sur les injustices et discriminations ressenties dans différents domaines de la vie sociale (accès au logement, à l'emploi, études poursuivies, etc.) et leurs motifs (âge, sexe, état de santé ou handicap, couleur de peau, origines ou nationalité, orientation sexuelle, etc.). Elle a été rééditée une décennie plus tard en 2019-2020, dans un contexte où les thèmes de l'immigration et de la diversité de la population française demeurent au cœur du débat public en France[2]. Dix ans plus tard, TeO a élargi son périmètre d'observation, intégrant de nouvelles populations dont on attend de comprendre si elles font l'objet des mêmes discriminations que celles identifiées avec l'enquête de 2008.

▶ *Trajectoires et Origines* : une enquête originale

Depuis sa première édition en 2008, *Trajectoires et Origines* est une enquête de l'Ined, l'institut national des études démographiques, et de l'Insee, l'institut national de la statistique et des études économiques, qui mêle de ce fait des questions intéressant tant la recherche que la statistique publique. TeO étudie la diversité des populations en France et la situation des populations d'origine immigrée ; l'enquête a pour objectif de fournir des informations fiables sur des questions importantes du débat public, comme l'intégration, les inégalités selon l'origine et les discriminations dans la société française.

TeO est originale par la population qu'elle étudie : les personnes immigrées et leurs descendants nés en France, mais également les natifs d'outre-mer et leurs descendants nés en France. Pour disposer d'éléments de comparaison, l'échantillon comprend également des personnes qui sont françaises depuis deux générations ou plus, appelées personnes de la « population majoritaire » ou « sans ascendance migratoire directe » (voir *infra*). Avec la deuxième édition, TeO2 s'intéresse de plus aux descendants d'immigrés de 3e génération.

2 Cet article est pour l'essentiel une synthèse d'un document de travail paru en juillet 2022 *(Thao Khamsing, Guin, Merly-Alpa, Paliod, 2022)*, qui a également bénéficié des contributions de *Cris Beauchemin, Mathieu Ichou* et *Patrick Simon* (Ined), *Odile Rouhban* et *Pierre Tanneau* (Insee).

La finalité de TeO est également singulière : comprendre comment les origines géographiques, nationales ou culturelles modifient l'accès aux ressources de la vie sociale qui définissent la place de chacun dans la société. TeO couvre ainsi de nombreux aspects des trajectoires individuelles (migratoires, scolaires, professionnelles). Elle collecte des informations d'une nature particulière : pratiques, ressentis et opinions. Certaines questions sont de ce fait « sensibles », notamment celles concernant la religion, la santé et la vie citoyenne : il s'agit autant d'une sensibilité qui peut être ressentie par les enquêtés eux-mêmes que d'une catégorie de données définie dans la loi Informatique et Libertés[3] : « *données à caractère personnel qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses* [...] ». Les traitements utilisant ce type de données sont en général interdits ; TeO s'inscrit dans les exceptions prévues par la loi.

▶ Une première édition largement exploitée depuis 2009

TeO1, la première édition de l'enquête (2008-2009), a été réalisée auprès d'un large échantillon (21 800 personnes) (*Algava et Lhommeau 2009*) dans le but de combler un déficit de données sur les immigrés et les enfants nés en France de parents immigrés. Différentes thématiques y étaient abordées :

- l'entourage familial, l'histoire matrimoniale, les relations sociales ;
- l'accès à l'éducation, au logement, à l'emploi, la santé et la vie citoyenne ;
- les différentes dimensions des origines et appartenances culturelles (lien avec le pays d'origine, religion, langues, image de soi et regard des autres).

Le questionnaire comprenait des éléments rétrospectifs sur les parcours individuels (trajectoires scolaires, professionnelles, familiales et résidentielles), afin d'analyser les processus d'insertion.

Pour mener une enquête statistique, il faut pouvoir constituer une base de sondage appropriée. Or s'il existe des sources statistiques permettant d'identifier les immigrés (recensement, enquête emploi en continu, etc.), celles-ci ne sont pas toujours adaptées pour étudier les descendants d'immigrés. Par ailleurs, les descendants d'immigrés, lorsqu'ils sont identifiés dans les enquêtes de la statistique publique, constituent souvent des échantillons trop petits pour pouvoir établir des analyses selon l'origine détaillée. Les sources administratives ne permettent pas non plus de répondre au besoin : même si elles peuvent informer sur le pays de naissance, elles ne peuvent pas en général informer sur la nationalité à la naissance, ce qui ne permet pas d'identifier les immigrés. Elles ne permettent pas plus d'identifier les descendants d'immigrés.

Afin de produire des informations fiables sur ces populations, l'enquête TeO1 a nécessité un échantillon de 21 800 individus résidant en France métropolitaine, répartis en cinq groupes :

- des immigrés, personnes nées étrangères à l'étranger (parfois dites de « première génération ») ;
- des enfants nés en France de parents immigrés, personnes nées en France ayant un ou deux parents immigrés (descendants de « deuxième génération ») ;

3 Voir références juridiques en fin d'article.

- des personnes nées dans un département d'outre-mer ;
- des personnes nées en France métropolitaine dont au moins un parent est né dans un département d'outre-mer ;
- des personnes représentatives de la population générale, qui en majorité n'appartiennent à aucun des groupes précédents et à partir de laquelle seront identifiées les personnes sans ascendance migratoire directe.

L'enquête TeO1 a donné lieu à une production scientifique importante ***(figure 1)*** et sert de référence pour les chercheurs et administrations travaillant sur les questions de la place de l'origine dans les processus d'intégration, de discrimination et de construction identitaire au sein de la société française. TeO1 a permis de mesurer l'ampleur des inégalités sociales et économiques liées aux origines, mais également de rendre compte des formes de participation à la société des immigrés et de leurs descendants et des trajectoires de mobilité sociale qu'ils peuvent suivre.

> ***TeO1 a permis de mesurer l'ampleur des inégalités sociales et économiques liées aux origines.***

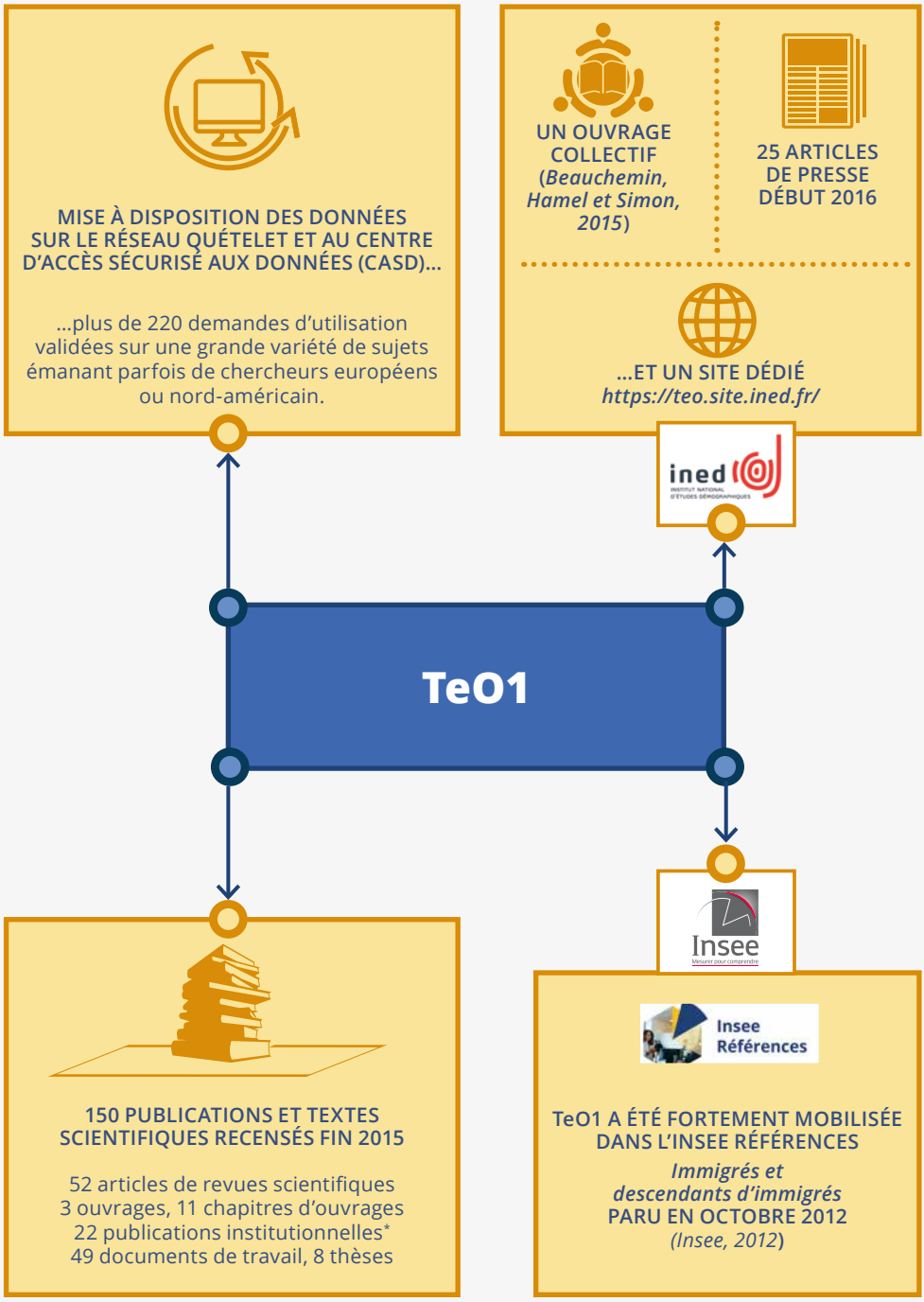
Mesurer dix ans d'évolution : continuité de la méthode, adaptation du questionnement

Dix ans après la première enquête, les thèmes de l'immigration et de la diversité de la population française sont toujours au cœur du débat public en France. Mais le contexte social a évolué, avec notamment une diversification des origines des immigrés et une attention accrue aux questions de discriminations. La réédition de l'enquête était donc attendue par les pouvoirs publics, la société civile et les chercheurs ***(encadré 2)***. Un des enjeux essentiels de cette deuxième édition était d'assurer des comparaisons avec TeO1 : une décennie plus tard, les constats de TeO1 sont-ils toujours d'actualité ? Le principe de continuité a donc prévalu dans les choix de conception de TeO2, qu'il s'agisse de l'échantillon ou du questionnaire.

Cependant, les acteurs du projet ***(encadré 3)*** ont introduit des innovations dans les thématiques abordées et le protocole de collecte, pour répondre à certains enjeux, notamment :

- prendre en compte l'évolution de la structure par origine de la population immigrée et apporter des connaissances nouvelles sur des groupes méconnus (chinois, réfugiés, etc.) ;
- étudier les trajectoires des enfants nés en France de parents « Français de l'étranger » (rapatriés d'Algérie ou d'autres territoires coloniaux, expatriés, etc.) ;
- identifier et interroger la « troisième génération » dans le but de mesurer l'impact des origines sur les trajectoires sociales et l'accès aux ressources au fil des générations.

Figure 1 - L'exploitation de l'enquête *Trajectoires et Origines 2008* a alimenté nombre d'études et de publications



**Insee Première, Population & Sociétés, Infos migrations, Dares, Analyses, etc.*

▶ Un questionnaire aux thématiques élargies

Depuis la première édition, le questionnaire couvre un grand nombre de thèmes (*Beauchemin et alii, 2015)* et comprend de nombreux éléments rétrospectifs relatifs aux trajectoires. Le recueil des éléments constitutifs de l'origine, aussi bien géographique que sociale, culturelle ou résidentielle, fait l'objet d'une attention particulière (voir *supra*). Le thème des discriminations est abordé de façon transversale dans différents modules, et fait également l'objet d'un module spécifique.

Avec TeO2, de nouveaux développements ont été introduits pour élargir les thématiques des études sur les sujets suivants :

- l'analyse de la ségrégation résidentielle et de ses effets sur les processus de discrimination scolaire et professionnelle ;
- les conséquences sur les trajectoires socio-économiques des migrants de leurs caractéristiques pré-migratoires ;
- l'impact des trajectoires légales (statut administratif) des migrants sur leurs conditions d'intégration ;
- les comportements de santé des populations immigrées et leur intégration au sein du système de soins français.

▶ Encadré 2. Une deuxième édition très attendue

Le souhait d'un renouvellement de l'enquête a été exprimé à plusieurs reprises dans diverses instances :

- en 2010, le rapport du Comité pour la mesure de la diversité et l'évaluation des discriminations (COMEDD) recommandait la réalisation d'une enquête, renouvelée périodiquement, dédiée à l'étude des discriminations et des inégalités ;
- le souhait de renouveler l'enquête TeO a été exprimé lors de la réunion de la commission « démographie et questions sociales » du CNIS d'octobre 2015, ainsi que par le Défenseur des droits lors de la conférence de presse de présentation des résultats de l'enquête de 2008-2009 le 8 janvier 2016 à l'Ined ;
- TeO2 a, par ailleurs, été inscrite dans le Plan national de lutte contre le racisme et l'antisémitisme (2018-2020), élaboré par la délégation interministérielle à la Lutte contre le racisme, l'antisémitisme et la haine anti-LGBT (DILCRAH) ;
- un indicateur du fort intérêt pour la première édition de l'enquête a été le nombre de demandes d'utilisation des données effectuées auprès du réseau Quételet qui a dépassé les 220.

▶ Encadré 3. TeO2, fruit d'une co-maîtrise d'ouvrage entre l'Insee et l'Ined et d'un large financement

La conception et le pilotage de la réalisation de TeO2 sont le fruit d'un travail collectif, qui a associé chercheurs de l'Ined et statisticiens de l'Insee *(Ined, 2022)*. Les deux parties sont intervenues conjointement à tous les stades de l'enquête TeO2. L'Insee a assuré la collecte et son organisation, ainsi que la méthodologie d'enquête et les développements informatiques. L'Ined a piloté les questions d'opportunités de recherche, de la conception du questionnaire à l'exploitation de l'enquête.

Le projet a mobilisé de nombreux acteurs, dans toute une série de groupes ou comités :

- le *comité de pilotage :* pour prendre les décisions stratégiques, préciser le contour du projet et s'assurer de l'atteinte des objectifs en termes de réalisation, coût et délai ;
- le *comité de suivi* : pour instruire les questions techniques, coordonner toutes les phases de l'enquête, assurer le suivi de l'exécution, veiller au respect du calendrier de travail et des coûts associés et préparer les éléments pour la prise de décision par le comité de pilotage ;
- le *conseil scientifique* : instance de consultation, sollicitée par l'un ou l'autre des deux maîtres d'ouvrage en cas de difficulté dans la préparation ou la conduite de l'enquête ;
- le *groupe technique dédié à l'échantillonnage :* pour instruire la constitution de l'échantillon, faire partager les constats et les contraintes et ainsi faciliter les arbitrages du comité de pilotage ;
- le *groupe de conception du questionnaire*, assisté de différents groupes thématiques consacrés à l'évolution de chacun des modules ;
- le *groupe d'exploitation* : pour aider à l'apurement des données et la constitution des fichiers d'étude, puis assurer l'exploitation de l'enquête. Composé des membres du groupe de conception et d'autres chercheurs, chargés d'études et représentants de bailleurs.

Le coût total de l'enquête est estimé à 5,7 millions d'euros dont 3,8 millions d'euros liés à la collecte. Le financement a mobilisé, outre l'Insee et l'Ined, onze organismes :

- Agence nationale de la cohésion des territoires (ANCT) ;
- Caisse nationale des allocations familiales (CNAF) ;
- Défenseur des Droits (DDD) ;
- Délégation interministérielle à la lutte contre le racisme et l'antisémitisme et la haine anti-LGBT (DILCRAH) ;
- Direction générale de la cohésion sociale (DGCS) ;
- France Stratégie ;
- Institut national de la Jeunesse et de l'éducation populaire (INJEP) ;
- Ministère de la Culture ;
- Secrétariat d'État à l'égalité Femmes / Hommes ;
- SSM Travail (Dares, direction de l'animation de la recherche, des études et des statistiques) ;
- SSM Immigration (DSED, département des statistiques, des études et de la documentation du ministère de l'Intérieur).

▶ Un échantillonnage toujours aussi original…

Les populations d'intérêt de l'enquête sont toutes des personnes résidant en France métropolitaine. Comme pour la première édition de l'enquête, l'échantillon de TeO2, un peu plus de 50 000 individus, est constitué de cinq sous-échantillons : immigrés (1re génération à vivre en France), descendants d'immigrés de 2e génération, natifs d'outre-mer, descendants de natifs d'outre-mer, et personnes représentatives de la population générale ***(figure 2)***.

Les quatre premiers sous-échantillons visent à assurer une représentation minimale des groupes dont on pense qu'ils peuvent être soumis à des discriminations du fait de leur origine, quelle que soit leur nationalité.

L'enquête TeO vise aussi à diffuser des résultats selon des **groupes d'origines** : ce sont des groupes de pays, dont les personnes qui en sont originaires ont potentiellement des histoires très différentes les unes des autres. Ceci implique de disposer de suffisamment de répondants pour chacun d'entre eux. Les sous-échantillons d'immigrés et de descendants d'immigrés ont été stratifiés selon ces groupes d'origines, avec des taux de sondage variables. Les cibles de collecte par groupe ont fait par ailleurs l'objet d'une attention toute particulière.

Pour la 1re génération, on a distingué 11 groupes pour TeO2 : les personnes nées en Algérie ; au Maroc ou en Tunisie ; en Afrique sahélienne ; en Afrique centrale ou dans le golfe de Guinée ; en Espagne, Italie ou Portugal ; dans les autres pays d'Union européenne ; en Turquie ; en Asie du Sud-Est ; en Chine ; dans des pays avec de nombreux réfugiés ; et dans les autres pays. Pour la 2e génération, les groupes étaient légèrement différents (on sépare le Portugal de l'Espagne et de l'Italie, on ne distingue pas la Chine, ni les pays avec de nombreux réfugiés).

▶ De nouveaux groupes d'origines : Chinois, réfugiés

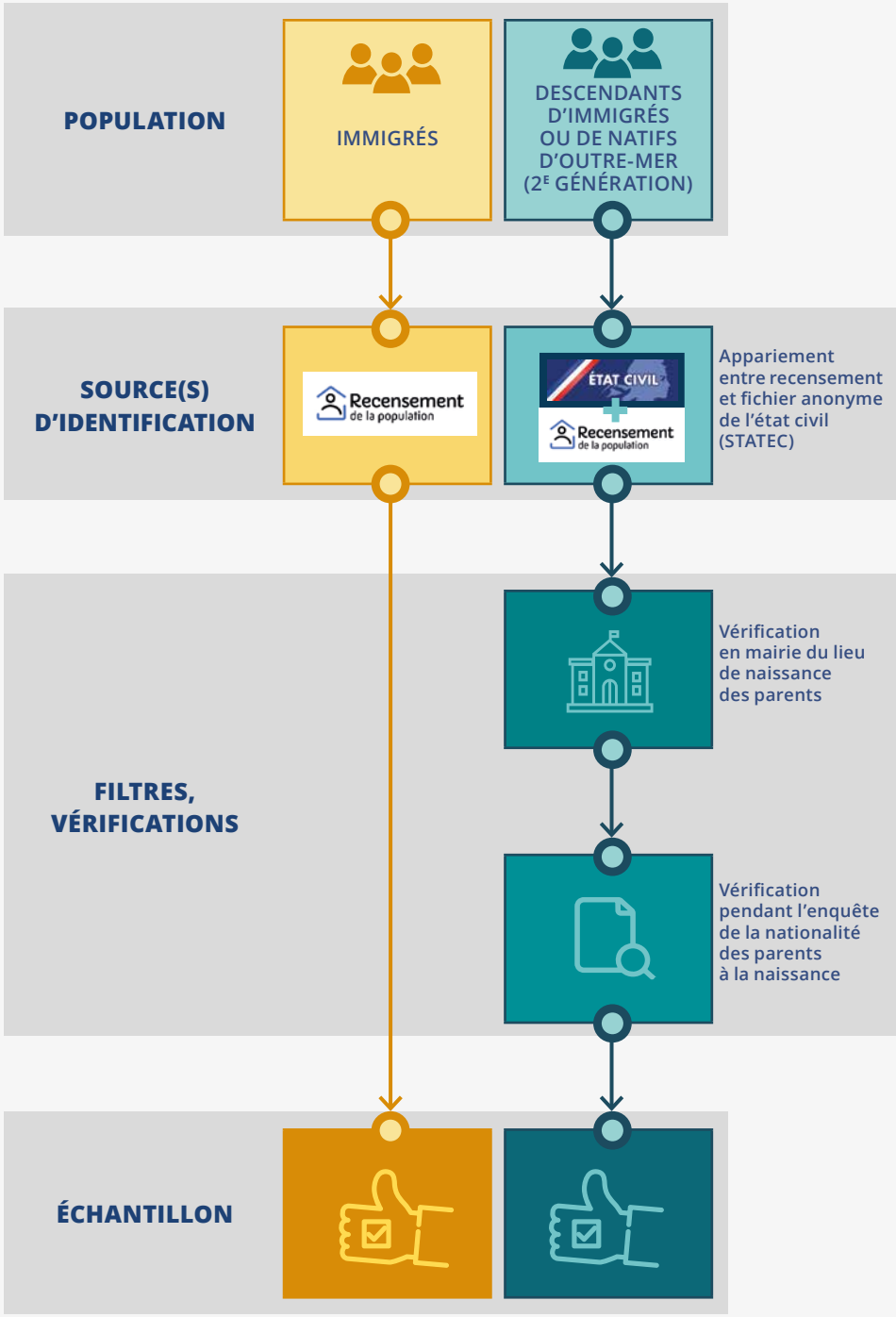
Afin de tenir compte de l'évolution de la structure par origine de la population immigrée et d'apporter des connaissances nouvelles sur des groupes méconnus, il a été décidé de surreprésenter dans l'échantillon de TeO2 :

- un groupe « Chine », dont la population immigrée a crû rapidement au cours des dix dernières années ;
- un groupe constitué de pays d'où sont originaires un grand nombre de réfugiés[4] : Angola, Sri Lanka, Thaïlande, Pakistan, Haïti, Russie, pays de l'ex-Yougoslavie hors Union européenne.

Par ailleurs, l'objectif visé était d'augmenter les effectifs de réfugiés déjà présents dans les groupes d'origine d'intérêt (notamment les deux Congo compris dans le groupe « Afrique centrale et golfe de Guinée » ou le Vietnam, le Laos et le Cambodge inclus dans l'« Asie du Sud-Est »).

4 Pays qui comptaient plus de 20 % de réfugiés dans TeO1 et dont l'effectif est supérieur à 1 000 individus dans l'enquête annuelle du recensement de 2013.

Figure 2 - L'échantillonnage des première et deuxième générations.



▶ Première interrogation de la « troisième génération »…

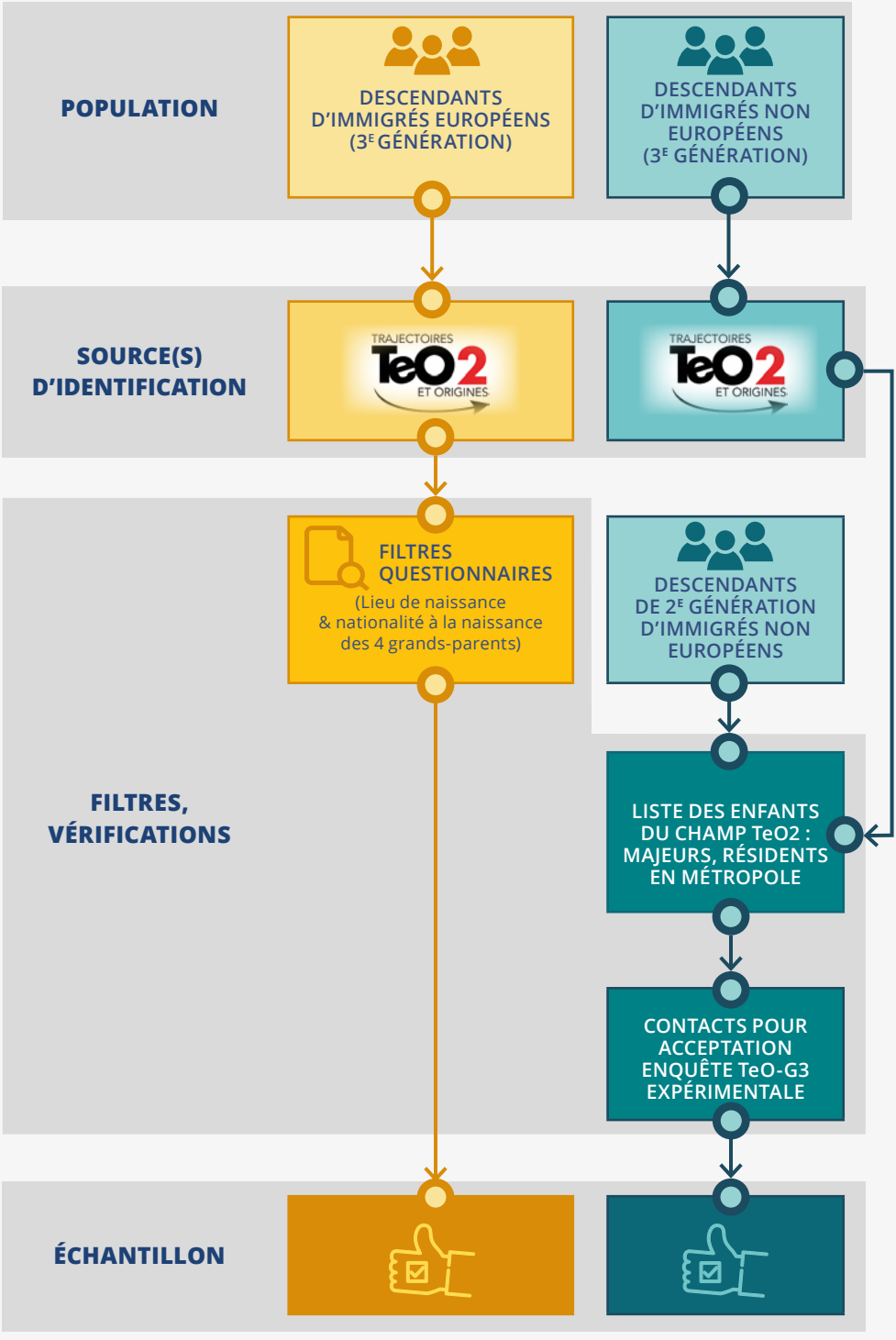
À la différence de la première édition cependant, l'enquête TeO2 s'intéresse également spécifiquement aux personnes qui ont, soit une ascendance migratoire au-delà de la deuxième génération, soit au moins un parent né français à l'étranger (y compris dans une ancienne colonie). Depuis 2008, davantage d'enfants des descendants d'immigrés de 2e génération sont devenus adultes (voir *supra*) et la question de la persistance de l'influence des origines sur leurs trajectoires sociales se pose. TeO1 a en effet montré que la situation socio-économique ou le ressenti des discriminations des enfants nés en France d'au moins un parent d'origine maghrébine, africaine subsaharienne, asiatique ou turque s'amélioraient peu ou se dégradaient par rapport aux perceptions des immigrés de même origine. Il s'agissait de savoir si la situation observée pour la deuxième génération d'origine non européenne, se maintient ou disparaît à la génération suivante.

> ***Identifier les personnes de la troisième génération.***

Un des objectifs de l'enquête TeO2 a donc été d'identifier les personnes de la troisième génération, c'est-à-dire les personnes ayant au moins un grand-parent immigré, sans être eux-mêmes ni immigrés, ni descendants d'immigrés. Comme pour la deuxième génération, il n'existe pas de base de sondage permettant d'identifier les descendants de troisième génération. Cela a conduit à mettre en place une méthodologie particulière, avec une stratégie à deux niveaux ***(figure 3)***. D'une part, les descendants de troisième génération d'origine européenne sont identifiés directement *via* le questionnaire de TeO2, grâce à de nouvelles questions sur la nationalité et le pays de naissance des quatre grands-parents. Suffisamment nombreux dans le sous-échantillon de la population générale, ils peuvent faire l'objet d'analyses sans nécessiter de surreprésentation particulière. D'autre part, les descendants d'immigrés de troisième génération d'origine non européenne, *a priori* moins nombreux, ont fait l'objet d'une enquête complémentaire expérimentale (TeO2-G3). Ses résultats devront éclairer les situations de discrimination ou de racisme perçues par une partie de la population majoritaire dans la première enquête, et qui pourraient être expliquées par la présence, au sein de cette population, de personnes d'origines migratoires plus anciennes. L'échantillon de cette enquête complémentaire a été constitué, par sondage indirect : l'enquête TeO2 sur les descendants d'immigrés comportait en effet des questions sur les enfants des enquêtés, qu'ils résident ou non dans le logement de leurs parents. Le protocole prévoyait un module de prise de contact et de sélection des personnes de troisième génération à la toute fin du questionnaire pour vérifier l'éligibilité des enfants et collecter leurs coordonnées (adresses postale et électronique, téléphone). Un délai a été laissé aux parents pour échanger avec leurs enfants et s'assurer de leur accord, cette partie du questionnement n'étant pas obligatoire.

Sur le millier d'individus adultes éligibles, on ne disposait de coordonnées que pour un tiers : 240 personnes ont été interrogées entre mars 2020 et janvier 2021, à travers quatre vagues de collecte. Le questionnaire était identique à celui de l'enquête TeO2.

Figure 3 - L'échantillonnage de la troisième génération.



▶ … et des enfants nés en France de parents « Français de l'étranger »

Dans TeO1, les personnes nées en France de parents nés de nationalité française à l'étranger[5], étaient considérées comme hors champ. Elles n'ont donc répondu qu'à un questionnaire tronqué. Ainsi, pour environ 3 000 individus, l'entretien débutait mais était interrompu dès que l'enquêté répondait que son ou ses parents nés à l'étranger étaient nés français. Dans l'enquête TeO2, il a été décidé de poursuivre le questionnement jusqu'au bout. Les trajectoires de ces personnes constituent un domaine de recherche encore peu étudié. Avec TeO2, on va pouvoir, par exemple, comparer leur parcours migratoire à celui des personnes dont les parents immigrés sont nés dans le même pays.

▶ Améliorer le suivi des enquêtés qui ont déménagé

Comme TeO1, l'enquête TeO2 est réalisée auprès des individus et menée en face-à-face par les enquêteurs de l'Insee. Ces enquêtes présentent une difficulté supplémentaire par rapport aux enquêtes classiques auprès des ménages, où ce sont des logements qui sont échantillonnés *(Sillard et alii, 2020)*. En effet, les individus ayant déménagé doivent faire l'objet d'un traitement particulier, afin de pouvoir les interroger.

Dans le cadre de l'enquête TeO2 ***(encadré 4)***, la base de sondage étant le recensement de 2018, lorsqu'une personne a été tirée, l'enquêteur dispose pour la joindre de ses nom, prénom et adresse en 2018. Or, les immigrés et leurs descendants déménagent plus souvent que la population sans ascendance migratoire. Les enquêteurs ont donc pour consigne de chercher la nouvelle adresse de ces personnes.

Pour retrouver les personnes ayant déménagé, les enquêteurs ont eu recours à :

- des informations issues du terrain (habitants actuels du logement, voisinage, mairie, etc.) ;
- la consultation de l'annuaire et la recherche internet ;
- les coordonnées (adresse mail et téléphone) récupérées grâce à l'appariement de la base de sondage TeO2 avec les fichiers de la taxe d'habitation. Cet appariement non standard est techniquement possible et l'utilisation des données des fichiers d'imposition des personnes est juridiquement autorisée. Cela a fait l'objet d'une déclaration générique par l'Insee à la Cnil, couvrant l'ensemble des enquêtes.

▶ Comment interroger des personnes non francophones ?

Répondre à une enquête multi-thématique et relativement longue comme TeO2 demande une bonne maîtrise de la langue française. De plus, sont surreprésentées dans l'échantillon des personnes potentiellement peu francophones, par exemple, les immigrés chinois. Il était donc essentiel de mettre en œuvre une méthodologie permettant d'enquêter dans des conditions satisfaisantes la population immigrée, quel que soit son niveau en français.

5 Par exemple les rapatriés d'Algérie ou d'autres territoires coloniaux, les expatriés, etc.

Le protocole dans TeO1 supposait que l'enquêteur puisse contacter un interprète s'il identifiait des difficultés rédhibitoires pour la conduite de l'entretien en français. La difficulté majeure était de réussir à mettre en place des rendez-vous à trois (enquêteur, enquêté et traducteur). Pour l'enquête TeO2, un nouveau protocole a été mis en place pour faciliter ces entretiens :

- en l'absence d'un tiers pouvant traduire (cas d'enquêtés non francophones isolés), les enquêteurs étaient chargés d'identifier la langue parlée par l'enquêté. Pour cela, lors de leur visite, ils disposaient d'une fiche de contact traduite en 22 langues[6] ;
- l'entretien était réalisé dans un deuxième temps par des enquêteurs traducteurs de l'Ined (enquête de rattrapage des non francophones).

▶ Encadré 4. De la première à la deuxième génération, l'échantillonnage se complexifie

Pour les immigrés, les natifs d'outre-mer, leurs descendants de 2e génération, le principe général de la méthode d'échantillonnage de TeO1 *(Algava et Lhommeau, 2013)* a été reconduit dans TeO2 *(Merly-Alpa, Paliod et Thao Khamsing, 2021 ; 2022)* : les immigrés et natifs d'outre-mer sont sélectionnés directement dans l'enquête annuelle de recensement* ; mais ce n'est pas possible pour les descendants d'immigrés ou de domiens, car l'information sur le lieu de naissance des parents et sur leur nationalité à la naissance est absente de cette source, on a donc conçu un dispositif *ad hoc.*

On recherche des candidats potentiels dans l'enquête annuelle du recensement, *i.e.* toutes les personnes de 18-59 ans, nées en France métropolitaine, hors immigrés et natifs d'outre-mer, soit plus de 2 millions d'individus.

Puis on apparie ces données avec le fichier anonyme** d'état civil dont dispose l'Insee pour les individus nés après 1968, en vue d'enlever les personnes qui n'ont aucun parent né à l'étranger ou dans les Dom. Ce sont de potentiels descendants d'immigrés ou de natifs d'outre-mer.

On vérifie ensuite, dans les bulletins de naissance en mairie, qu'il s'agit bien de descendants de personnes nées à l'étranger ou dans les Dom ; au préalable et afin de limiter les coûts des recherches, un tirage d'individus a été effectué. Cette consultation des bulletins de naissance en mairie, sur lesquels figurent les lieux de naissance des parents (mais pas leur nationalité), est une opération atypique et spécifique à TeO. Elle a concerné environ 100 000 personnes.

L'opération aboutit à la construction d'une base de sondage de descendants de personnes nées à l'étranger ou dans les Dom.

Pour les personnes dont au moins un parent est né à l'étranger, le tirage avec surreprésentation des groupes d'origines peut être effectué, en se basant sur le pays de naissance des parents. Il restera à vérifier la nationalité de naissance des parents, pour déterminer précisément les descendants d'immigrés***. C'est à partir du questionnaire de l'enquête TeO2 que l'on recueille cette information. Bien que coûteuse et risquée à l'origine, cette méthode d'échantillonnage est apparue comme un succès.

* *EAR de 2018 pour TeO2.*

** *La base STATEC contient une information sur le sexe, la date et le lieu de naissance. L'appariement sur ces traits d'identité permet de récupérer un ou plusieurs individus correspondant, dont on connaît le lieu de naissance des parents.*

****L'opération permet de ne pas confondre les descendants d'immigrés avec les descendants de Français nés à l'étranger.*

6 Anglais, turc, arabe, vietnamien, portugais, espagnol, mandarin, laotien, russe, tamoul, serbe, croate, allemand, khmer, bengali, thaï, pachto, roumain, albanais, géorgien, lituanien et tchèque.

Des vidéos pour former les enquêteurs… et informer le grand public

L'originalité de l'enquête et son protocole particulier justifiait d'utiliser le support vidéo pour la formation des gestionnaires d'enquête et des enquêteurs. Ainsi, six courtes séquences vidéos ont été produites entre mars et septembre 2019 pour expliquer les objectifs de l'enquête, ses enjeux et le protocole de collecte.

La première séquence explique les objectifs de l'enquête et ses enjeux. Les deux suivantes (nationalités, trajectoires légales des migrants) portent sur deux modules du questionnaire qui recouvrent des concepts peu standards dans les enquêtes de la statistique publique. Il a paru important que les enquêteurs s'approprient au mieux ces notions avant la réalisation de l'enquête. Deux autres séquences (interrogation des non francophones, interrogation des troisièmes générations) portent sur les protocoles de collecte particuliers. Enfin, le questionnaire comportant des questions sensibles dans trois de ses modules (religion, santé, vie citoyenne), il a semblé nécessaire de sensibiliser les enquêteurs et le grand public sur cet aspect dans une sixième vidéo.

L'équipe de conception a également proposé de mettre à disposition du grand public deux de ces vidéos : celle sur la présentation de l'enquête et celle sur le traitement des données sensibles. Celles-ci sont mises en ligne sur le site de l'Insee *(Insee, 2021)*, le site TeO2 *(Ined, 2022)* et sur la chaîne YouTube de l'Insee *(Insee, 2019a et 2019b)*.

Le bilan de cette opération a été très positif. Ces vidéos ont été très appréciées, que ce soit par les enquêteurs ou enquêtés et par les gestionnaires pour leur formation auprès des enquêteurs, comme en témoignent les bilans d'enquête réalisés auprès de ces acteurs.

L'enquête principale s'est déroulée entre juillet 2019 et novembre 2020. 27 000 personnes, âgées de 18 à 59 ans, de toutes origines, y ont répondu (54 % de l'échantillon initial). Environ 240 personnes ont répondu à l'enquête expérimentale sur 360 qui pouvaient être contactées ***(encadré 5)***.

Une fois la collecte réalisée, l'enjeu était ensuite de s'assurer de la représentativité des répondants. Classiquement, l'étape du redressement vise à corriger l'échantillon de ses éventuelles déformations par rapport à la population cible, dues à la non-réponse.

> ***Cette enquête a posé quelques défis méthodologiques en termes de redressement.***

Cependant, cette enquête a posé quelques défis méthodologiques en termes de redressement, que ce soit dans les opérations dédiées (correction de la non-réponse, partage des poids, calage sur marge) ou dans l'articulation de ces trois opérations *(Guin et Thao Khamsing, 2021a et 2021b)*.

Une correction de la non-réponse qui préserve les groupes d'origines…

La collecte de l'enquête TeO2 s'est traduite par une déperdition de l'échantillon mis en collecte initialement. Afin de maintenir la représentativité de l'échantillon final obtenu, des travaux méthodologiques de redressement[7] de l'échantillon ont donc été menés.

La non-réponse à l'enquête réduit la taille de l'échantillon exploitable[8]. Ce dernier n'est dès lors plus représentatif de l'ensemble de la population. Pour compenser le biais introduit par les non-répondants, il est nécessaire de corriger les poids attribués aux individus de l'échantillon de façon à réaffecter le poids des personnes qui n'ont pas répondu. L'étape de correction de la non-réponse totale consiste à estimer parmi les répondants et les non-répondants, la probabilité de répondre en fonction des caractéristiques disponibles sur l'ensemble de ce champ. Le poids initial des répondants (inverse de la probabilité d'inclusion dans l'échantillon tiré dans la base de sondage) est ensuite corrigé en le multipliant par l'inverse de la probabilité de réponse estimée. Cette opération permet de corriger la non-réponse totale et de rendre ainsi l'échantillon des répondants représentatif de l'ensemble du champ de l'enquête.

Encadré 5. Une collecte longue qui a dû s'adapter à la crise sanitaire

La collecte de l'enquête TeO2 était organisée en deux vagues. La première vague concernait la collecte du sous-échantillon des immigrés, des natifs d'outre-mer et de la population générale et a eu lieu du 1er juillet au 31 décembre 2019, sur 26 semaines ininterrompues.

La deuxième vague de l'enquête TeO2 a débuté le 1er janvier 2020. Elle concernait les sous-échantillons des personnes dont les parents sont nés à l'étranger et dans les Dom. À compter du mois de mars 2020, la collecte a été interrompue trois mois en raison de la crise sanitaire. À la suite du déconfinement, une expérimentation a été mise en place afin de tester la reprise de la collecte en face-à-face. À l'issue de celle-ci, le protocole a été adapté (nouvelles lettres avis reprenant le protocole sanitaire, autorisation partielle du téléphone, etc.) et la fin de la collecte a été repoussée de deux mois jusque fin novembre 2020. À la suite de l'annonce du deuxième confinement (29 octobre 2020), il a été décidé que la fin de collecte de l'enquête TeO2 se ferait exclusivement par téléphone.

L'enquête TeO2 a permis d'interroger environ 27 000 individus.

La collecte de l'enquête expérimentale TeO2-G3 a débuté le 10 mars 2020 et a été presque immédiatement interrompue par le premier confinement. Après la reprise de la collecte en juillet 2020, le taux de collecte est resté faible. Avec le deuxième confinement, il a été décidé de prolonger la collecte des individus de troisième génération déjà repérés jusqu'au 15 décembre 2020, puis d'interroger un quatrième lot d'individus au mois de janvier 2021.

L'enquête expérimentale TeO2-G3 a permis d'interroger environ 240 individus, pour un objectif initialement fixé à 500.

7 Ne sont présentées ici que les idées générales. La chaîne de redressement est décrite dans le détail dans *(Thao Khamsing, Guin, Merly-Alpa, Paliod, 2022)*. Cette partie n'aborde pas le redressement de l'enquête expérimentale sur les descendants de 3e génération (TeO2-G3).

8 Et augmente de ce fait la variance des estimateurs.

La prise en compte des enjeux de diffusion de l'enquête par groupes d'origines a orienté la correction de la non-réponse vers une approche multi-modèle par groupes d'origines[9].

▶ Un partage des poids pour tenir compte de l'échantillon en population générale

Parmi les cinq sous-échantillons cités plus haut, celui issu de la population générale est représentatif de l'ensemble de la population. Il comprend donc des individus appartenant à toutes les populations d'intérêt de l'enquête. Par construction, immigrés, natifs d'outre-mer, descendants de 2e génération ou descendants de natif d'outre-mer, chacun pouvait potentiellement être sélectionné *via* le sous-échantillon de la population générale ou *via* le sous-échantillon spécifique à ces groupes d'intérêt.

L'opération de partage des poids *(Lavallée, 2009)* a consisté à repérer *a posteriori* (avec les informations de la collecte sur les répondants), les individus qui potentiellement pouvaient être sélectionnés dans deux sous-échantillons et à corriger leur poids. Encore une fois pour répondre aux enjeux de diffusion de l'enquête par groupes d'origines, l'opération du partage des poids a été adaptée pour prendre en compte les origines des individus (méthode des liens pondérés)[10].

▶ Un calage sur marges qui s'appuie sur le recensement

À l'issue de l'opération du partage des poids, on obtenait donc un jeu de pondération unique sur l'échantillon des répondants. L'objectif du calage sur marges *(Deville et Särndal, 1992)* est double :

- assurer une cohérence entre les totaux estimés de certaines variables (dites variables de calage) et les vrais totaux de ces variables connus, par une information externe, sur la population ;
- réduire la variance des estimateurs des variables d'intérêt de l'enquête.

La procédure du calage a porté sur trois populations : les immigrés, les natifs d'outre-mer, les « autres » (ni immigrés, ni natifs d'outre-mer).

Le calage nécessite de disposer d'une source externe qui fournisse les totaux connus de variables auxiliaires (les marges). La source de référence retenue pour le calcul des marges a été l'enquête annuelle de recensement de l'année 2019, couplée avec celle de 2020 (qui donne une photographie de la population en moyenne sur l'année 2019). La taille de cette source – près de 5 millions d'observations sur le champ de TeO2 – est son principal avantage. En particulier, elle permet de mesurer des caractéristiques sur les populations rares comme les natifs d'outre-mer (qui représentent moins de 1 % de la population).

9 L'enquête TeO2 est déjà stratifiée en groupes d'origines ; il s'agit ici d'envisager des agrégations de ces groupes d'origines.
10 Méthode des liens pondérés proposée par *(Davezies, Landré, Murat et Rousseau, 2005).*

Une vigilance a été portée dans le choix des marges[11] pour éviter les différences de concept entre les marges de la population estimée à partir de la source externe et les variables correspondantes collectées pour l'échantillon de l'enquête. Par exemple, le niveau de diplôme est une information qui existe dans le recensement mais qui n'a pas été retenue dans le calage : l'interrogation est en effet très différente entre le recensement (variable auto-déclarée) et TeO (codification sur le poste de collecte).

▶ Une collecte riche qui va nourrir de nombreuses études

Dans le système statistique français, l'enquête TeO est une enquête particulière à plusieurs titres. Réalisée deux fois à ce stade, elle est le fruit d'une collaboration entre la recherche et la statistique publique. Si cette dernière produit de nombreuses études sur l'immigration et l'intégration à partir d'autres sources comme l'enquête Emploi ou le recensement de la population, l'enquête TeO aborde des thèmes habituellement peu traités comme les discriminations, la religion, la vie citoyenne et interroge des populations difficiles à identifier et à caractériser dans les sources administratives et les enquêtes existantes.

Tout en cherchant à répondre à des objectifs de comparabilité et d'actualisation, TeO2 a laissé place à des évolutions (questionnaire, choix d'interroger de nouveaux groupes de personnes) mais également à des innovations en particulier pour identifier les personnes de la troisième génération. L'absence d'une base de sondage permettant d'échantillonner ces personnes a conduit à développer des méthodes sophistiquées dans le plan de sondage, le protocole d'enquête et les travaux post collecte.

> ***De multiples compétences ont été mobilisées et il a également fallu être réactif et s'adapter tout au long de la conception et la réalisation de l'enquête.***

Ainsi, de multiples compétences ont été mobilisées et il a également fallu être réactif et s'adapter tout au long de la conception et la réalisation de l'enquête. Bien que longue et perturbée par le contexte sanitaire, la collecte a tout de même permis d'atteindre une taille d'échantillon suffisamment importante (plus de 27 000 individus) pour mener des analyses détaillées selon la plupart des groupes d'origines ciblés par l'enquête. Les premiers résultats issus de l'enquête ont été diffusés par l'Insee et l'Ined en juillet 2022 (*(Lê, et alii,, 2022)*, *(Lê, Simon et Coulmont, 2022)* et *(Beauchemin, Ichou, Simon et alii, 2022)*). Les prochaines années devraient voir la publication de nombreuses études et analyses valorisant les nombreuses dimensions de TeO2, que ce soit pour actualiser les constats de TeO1 ou pour exploiter les innovations de cette deuxième édition. Chacun est invité à contribuer à cette exploitation.

11 Sexe, âge, couple, type de logement, nationalité, pays de naissance d'origine, région de résidence, tranche d'unité urbaine, etc.

▶ Bibliographie

- ALGAVA, Élisabeth et LHOMMEAU, Bertrand, 2009. T'es où TeO ? À la recherche de la 2e génération pour l'enquête Trajectoires et origines. In : *Actes des 10e Journées de Méthodologie Statistique (JMS)*. [en ligne]. 23-25 mars 2009. Insee, Session 24. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *http://www.jms-insee.fr/2009/S24_1_ACTE_ALGAVA_JMS2009.PDF*.
- ALGAVA, Élisabeth et LHOMMEAU, Bertrand, 2013. *À l'origine de l'enquête TeO : enjeux de l'échantillonnage, collecte et pondérations de l'enquête TeO.* [en ligne]. 19 avril 2013. Insee, Documents de travail, n° F1304. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1381084*.
- BEAUCHEMIN, Cris, HAMEL, Christelle et SIMON, Patrick, 2016. *Trajectoires et origines. Enquête sur la diversité des populations en France.* Ined, Collection Grandes Enquêtes, ouvrage collectif : ISBN 978-2-7332-8004-1. *https://www.ined.fr/fr/publications/editions/grandes-enquetes/trajectoires-et-origines/*.
- BEAUCHEMIN, Cris, ICHOU, Mathieu, SIMON, Patrick *et alii*, 2022. Familles immigrées : le niveau d'éducation progresse sur trois générations mais les inégalités sociales persistent. In : *Population et Sociétés 2022/7.* [en ligne]. Juillet-août 2022. Ined, n° 602, pp. 1-4. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://doi.org/10.3917/popsoc.602.0001*.
- DAVEZIES, Laurent, LANDRÉ, Cédric, MURAT, Fabrice et ROUSSEAU, Sylvie, 2005. Problèmes théoriques et pratiques de la mise en œuvre d'une sur-représentation des ZUS dans les échantillons d'enquête : le cas de l'enquête IVQ. In : *Actes des 9e Journées de Méthodologie Statistique (JMS)*. [en ligne]. 14-16 mars 2005. Insee, Session 3. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *http://jms-insee.fr/2005/S03_5_ACTE_DAVEZIES-LANDRE-MURAT-ROUSSEAU_JMS2005.pdf*.
- DEVILLE, Jean-Claude et SÄRNDAL, Carl-Erik, 1992. *Calibration Estimators in Survey Sampling*. In : *Journal of the American Statistical Association.* Juin 1992. Taylor & Francis, *Ltd. on behalf of the AAS.* Vol. 87, n° 418, pp. 376-382. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://doi.org/10.1080/01621459.1992.10475217*.
- GUIN, Olivier et THAO KHAMSING, Willy, 2021a. Partage des poids pour l'enquête Trajectoires et Origines 2 (TeO2). In : *Actes du 11e Colloque International Francophone sur les Sondages.* [en ligne]. 6-8 octobre 2021. Société française de statistique (SFdS) et Université libre de Bruxelles (ULB). [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://sondages2020.sciencesconf.org/resource/page/id/15*.
- GUIN, Olivier et THAO KHAMSING, Willy, 2021b. Redressements de l'enquête Trajectoires et Origines 2 (TeO2). In : *Actes du 11e Colloque International Francophone sur les Sondages.* [en ligne]. 6-8 octobre 2021. Société française de statistique (SFdS) et Université libre de Bruxelles (ULB). [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://sondages2020.sciencesconf.org/resource/page/id/15*.

- THAO KHAMSING, Willy, GUIN, Olivier, MERLY-ALPA, Thomas et PALIOD, Nicolas, 2022. *Enquête Trajectoires et Origines 2. De la conception à la réalisation*. [en ligne]. 21 juillet 2022. Insee, Documents de travail, n°2022/02. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6478465*.
- INED, 2022. Enquête sur la diversité des populations en France. In : *site de l'Ined.* [en ligne]. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://teo.site.ined.fr/fr/*.
- INSEE, 2012. *Immigrés et descendants d'immigrés en France*. Édition 2012. [en ligne]. 10 octobre 2012. Insee Références. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1374025*.
- INSEE, 2016. *Pour comprendre… La mesure des populations étrangère et immigrée.* [en ligne]. Collection Insee en bref. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/2416930/insee-en-bref-immigration.pdf*.
- INSEE, 2019a. *Enquête TeO2 – Trajectoires et origines 2.* [en ligne]. 10 juillet 2019. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.youtube.com/watch?v=4AdOzdxeTZc*.
- *INSEE, 2019b. Les données sensibles dans TeO2.* [en ligne]. 23 août 2019. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.youtube.com/watch?v=P8vvrli4lI4*.
- INSEE, 2021. Trajectoires et Origines 2 : enquête sur la diversité des populations en France. In : *site de l'Insee.* [en ligne]. 09 mars 2021. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4172158*.
- INSEE, 2022a. *L'essentiel sur… les immigrés et les étrangers.* [en ligne]. 10 août 2022. Insee, Chiffres clés. [Consulté le 11 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/3633212*.
- LAVALLÉE, Pierre, 2009. *Indirect sampling.* Éditions Springer, Science & Business ISBN 978-0-387-70778-5.
- LE MINEZ, Sylvie, 2020. Oui, la statistique publique produit des statistiques ethniques. Panorama d'une pratique ancienne, encadrée et évolutive. In : *Le blog de l'Insee.* [en ligne]. 31 juillet 2020. Insee. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://blog.insee.fr/statistique-publique-produit-des-statistiques-ethniques/*.
- LÊ, Jérôme, ROUHBAN, Odile, TANNEAU, Pierre, BEAUCHEMIN, Cris, ICHOU, Mathieu et SIMON, Patrick, 2022. *En dix ans, le sentiment de discrimination augmente, porté par les femmes et le motif sexiste.* [en ligne]. 5 juillet 2022. Insee Première n°1911. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6473349*.
- LÊ, Jérôme, SIMON, Patrick et COULMONT, Baptiste, 2022. *La diversité des origines et la mixité des unions progressent au fil des générations.* [en ligne]. 5 juillet 2022. Insee Première n°1910.[Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6468640*.

- MERLY-ALPA, Thomas, PALIOD, Nicolas et THAO KHAMSING, Willy, 2021. Échantillonnage de l'enquête Trajectoires et Origines 2. In : *Actes du 11e Colloque International Francophone sur les Sondages*. [en ligne]. 6-8 octobre 2021. Société française de statistique (SFdS) et Université libre de Bruxelles (ULB). [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://sondages2020.sciencesconf.org/resource/page/id/15*.
- MERLY-ALPA, Thomas, PALIOD, Nicolas et THAO KHAMSING, Willy, 2022. Échantillonnage de l'enquête Trajectoires et Origines 2. In : *Actes des 14e Journées de Méthodologie Statistique (JMS).* [en ligne]. 29-31 mars 2022. Insee, Session 24. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *http://www.jms-insee.fr/2022/S24_3_ACTE_PALIOD_JMS2022.pdf*.
- SILLARD, Patrick, FAIVRE, Sébastien, PALIOD, Nicolas et VINCENT, Ludovic, 2020. Pour les enquêtes auprès des ménages, l'Insee rénove ses échantillons. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee. N° N4, pp. 81-100. [Consulté le 25 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497081/courstat-4-6.pdf*.

▶ Fondements Juridiques

- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. In : *site de Légifrance.* [en ligne]. Mise à jour le 26 janvier 2022. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460*.

Le SSM Collectivités locales : décryptage d'un appareil statistique à la fois robuste et en évolution

Luc Brière*

Les collectivités locales, du fait de l'extension de leurs compétences au fil des années et des missions essentielles qu'elles assurent, jouent un rôle majeur aux niveaux économique et social. Pour mieux les comprendre, les observer ou encore les accompagner, le service statistique ministériel relatif aux collectivités locales au sein de la direction générale des collectivités locales (DGCL) a développé au fil du temps un appareil statistique complet et robuste. Pour répondre aux besoins anciens en matière de suivi démographique, la DGCL et le Service Statistique Ministériel (SSM) Collectivités locales s'appuient sur le recensement de la population. Les besoins portent également, et depuis longtemps, sur le domaine des finances locales, et plus récemment sur le suivi des ressources humaines dans les collectivités, et les structures intercommunales. Le SSM propose également davantage d'analyses pour comprendre l'hétérogénéité des collectivités locales.

Les données produites s'inscrivent dans des processus de production, de validation et d'enrichissement des sources administratives. La consolidation des opérations comptables entre des entités qui entretiennent des flux croisés entre elles est l'un des processus majeurs de la construction de statistiques robustes. Enfin, le SSM Collectivités locales contribue à élargir l'accès aux informations statistiques, avec la mise en place récente d'un portail dédié.

Because of the extension of their powers over the years and the essential missions they perform, local authorities play a major role in both our society and our economy. In order to better understand, observe and support them, the Ministerial Statistical Office (MSO) for Local Authorities within the General Directorate of Local Authorities (DGCL) has developed over time a comprehensive and robust statistical system. To meet long-standing needs for demographic monitoring, the DGCL and the MSO rely on the population census. Needs are also, and for a long time, in the area of local finances, and more recently on the monitoring of human resources in local authorities, and the inter-municipal structures. The MSO offers more analyses to understand the heterogeneity of local authorities.

The data produced are part of a process of production, validation and enrichment of administrative sources. The consolidation of accounting operations between entities that have cross-flow between them is one of the major processes in the construction of robust statistics. Finally, the MSO contributes to widening access to statistical information, with the recent implementation of a dedicated portal.

* Chef du service statistique ministériel relatif aux collectivités locales, DGCL, ministère de l'Intérieur et des outre-mer, *luc.briere@dgcl.gouv.fr*

En cinquante ans, le paysage des collectivités territoriales a profondément évolué, notamment à la suite d'un processus de décentralisation initié au début des années quatre-vingt, et qui s'est poursuivi jusqu'à récemment au travers de plusieurs réformes ***(encadré 1)***. Favorisant et accompagnant le développement économique local *(Bouvier, 2020)*, de nombreuses compétences ont ainsi été transférées aux communes, aux départements et aux régions, aux collectivités d'outre-mer et à d'autres collectivités à statut particulier. Parallèlement, les communes se sont regroupées au sein d'établissements publics de coopération intercommunale, afin de mettre en commun leurs moyens. L'extension des compétences des collectivités s'est logiquement accompagnée d'un accroissement de leurs ressources financières et de leurs moyens humains. Ces évolutions contribuent à donner aujourd'hui aux collectivités territoriales un rôle d'acteur public majeur de la vie démocratique et économique de notre pays.

> ***Ces évolutions contribuent à donner aujourd'hui aux collectivités territoriales un rôle d'acteur public majeur de la vie démocratique et économique de notre pays.***

Les collectivités territoriales et leurs groupements[1] forment quatre niveaux d'administration locale exerçant chacun des compétences particulières[2] : les régions, les départements, les intercommunalités et les communes ***(encadré 2 et figure 1)***. La connaissance de ces structures diversifiées, de leur fonctionnement, de leur rôle, notamment économique et social, nécessite de disposer d'un appareil statistique fiable, homogène et complet. Il est centralisé en grande partie au sein du service statistique ministériel relatif aux collectivités locales[3].

L'information statistique sur les collectivités locales repose sur des sources principalement de nature administrative : leur élaboration échappe donc en partie à la statistique publique. Ces sources administratives portent sur trois domaines principaux : finances et fiscalité locales, emploi et salaires des agents employés dans les collectivités locales et composition et compétences des structures territoriales. Elles se fondent sur des référentiels d'unités statistiques et des nomenclatures qui leur sont propres[4].

Cette offre statistique permet de répondre aux nombreuses demandes des utilisateurs :

- les ministères, interlocuteurs des collectivités locales ;
- les collectivités locales elles-mêmes ;
- le monde universitaire ;
- la société civile, etc.

1 La Constitution établit trois niveaux de collectivités **territoriales** : communes, départements et régions. Par extension, les établissements publics de coopération intercommunale (EPCI), qui font partie du « bloc communal », sont intégrés à l'ensemble plus vaste des collectivités **locales**.

2 Certaines collectivités peuvent exercer simultanément des compétences relevant de plusieurs niveaux : ainsi la métropole de Lyon est à la fois une intercommunalité et un département, et la Ville de Paris exerce à la fois les compétences d'une commune et d'un département.

3 Le département des Études et des statistiques locales (DESL) au sein de la direction générale des collectivités locales, au ministère de l'Intérieur. Voir (*Bouinot, 1978)* pour la description des premiers travaux statistiques à la DGCL et Insee, 2008 pour un premier état de lieux du SSM collectivités locales.

4 On ne prétend pas établir ici un panorama exhaustif des statistiques et données disponibles. En particulier, les collectivités elles-mêmes disposent de plus en plus d'informations quantitatives.

Les demandes ont longtemps porté en grande majorité sur des résultats agrégés par catégorie de collectivité, ou par strate de nombre d'habitants des collectivités. Au fil du temps, ces demandes deviennent à la fois plus nombreuses et visent à mieux rendre compte des disparités et de la diversité au sein de cet ensemble hétérogène que constituent les collectivités locales. Le domaine du SSM Collectivités locales rejoint ainsi une tendance de fond des travaux de l'ensemble de la statistique publique. Comme sur d'autres thématiques, des demandes d'accès aux données individuelles les plus détaillées, en *open data*, sont récemment apparues.

▶ Encadré 1. La montée en puissance des collectivités territoriales et de leurs compétences

« *La France est une République indivisible [...] Son organisation est décentralisée.* »
(article premier de la Constitution)

« *Les collectivités territoriales de la République sont les communes, les départements, les régions, les collectivités à statut particulier et les collectivités d'outre-mer [...]. Les collectivités territoriales ont vocation à prendre les décisions pour l'ensemble des compétences qui peuvent le mieux être mises en œuvre à leur échelon. [...] ces collectivités s'administrent librement par des conseils élus et disposent d'un pouvoir réglementaire pour l'exercice de leurs compétences* » (article 72).

La décentralisation en France résulte d'un processus institutionnel, qui s'est construit en plusieurs actes (*Rimbaut, Verpeaux, Waserman, 2021*) :

- **Prologue :** pour la première fois, la Constitution de la IVe République (1946) consacre un titre* aux collectivités territoriales. Mais c'est la Constitution de la Ve République (1958) qui affirme le principe de libre administration des collectivités territoriales dans son article 72. La première étape de la création des intercommunalités remonte à 1966 avec la création de quatre communautés urbaines (Bordeaux, Lille, Lyon, Strasbourg). À partir des années quatre-vingt, la décentralisation s'accélère ;
- **Acte I de la décentralisation** (1982 – 1986) : 25 lois, complétées par environ 200 décrets, modifient la répartition des compétences entre les communes, les départements, les régions et l'État, dans de nombreux domaines (urbanisme, action sociale, formation professionnelle, collèges et lycées). La fonction publique territoriale est créée en 1984 ;
- **Acte II de la décentralisation** (2003 – 2004) : plusieurs lois constituant l'organisation décentralisée de la République ; aucune collectivité ne peut exercer une tutelle sur une autre, le principe de l'autonomie financière est posé ; de nouvelles compétences sont transférées par l'État aux collectivités territoriales ;
- **Plus récemment** (2014 – 2015) : la loi de modernisation de l'action publique territoriale et d'affirmation des métropoles (MAPTAM) opère notamment une refonte du statut des métropoles ; puis la loi portant nouvelle organisation territoriale de la République (NOTRe) supprime la clause de compétence générale aux régions et aux départements, renforce le rôle économique des régions, affirme le partage des compétences entre collectivités en matière de culture, sport et tourisme, et renforce l'intercommunalité en prônant la réduction de leur nombre.

La décentralisation des années quatre-vingt et deux mille contribue par ailleurs à accroître le poids économique des collectivités locales. Ainsi, au sens de la comptabilité nationale, l'investissement des administrations publiques locales représente en moyenne au cours des années 2018 à 2020 près de **60 % de l'investissement total des administrations publiques** ; en 1980, cette part était d'environ 52 %.

* Un titre est un chapitre de la Constitution. La Constitution française comporte un préambule et 17 titres.

Démographie et finances des collectivités : des objets statistiques déterminants

> ***Le suivi des populations dites « légales », issues du recensement de la population et authentifiées chaque année par décret, constitue ainsi un enjeu essentiel pour les communes elles-mêmes.***

Historiquement, les travaux statistiques sur les collectivités locales répondaient à des besoins de connaissance et d'analyse avant tout démographiques et financiers. Ces deux aspects ne sont d'ailleurs pas totalement indépendants compte tenu des modalités de détermination d'une partie des ressources des collectivités, les communes notamment, reçues sous forme de dotations de l'État.

Le suivi des populations dites « légales », issues du recensement de la population et authentifiées chaque année par décret, constitue ainsi un enjeu essentiel pour les communes elles-mêmes. Actuellement, ce facteur démographique conditionne toujours lui-même l'évolution de la dotation forfaitaire versée par l'État aux communes. Pour déterminer le montant de cette dotation, une notion spécifique de population a même été élaborée : la population « DGF »[5].

Encadré 2. Collectivités territoriales, une construction administrative progressive

Les **communes** constituent l'échelon le plus ancien et le plus proche des citoyens. Elles ont succédé en 1789 aux anciennes paroisses. En 2016, leur regroupement au sein de « communes nouvelles », a ramené leur nombre sous la barre symbolique des 36 000. Fruit d'une organisation administrative très ancienne, la France est ainsi l'un des pays qui compte le plus de communes.

Les **départements** ont été également créés en 1789. À l'origine circonscriptions d'action de l'État (représenté par le préfet), ils ne sont devenus aussi des collectivités territoriales qu'en 1871.

Les **régions** sont de création plus récente. Établissements publics dans les années soixante, elles deviennent des collectivités territoriales de plein exercice en 1982.

L'échelon intercommunal permet la mutualisation des moyens pour la mise en œuvre des politiques publiques locales, en partageant des services publics ou l'élaboration de certaines politiques. Les **établissements publics de coopération intercommunale** (EPCI), avec ou sans fiscalité propre, sont intégrés aux collectivités locales.

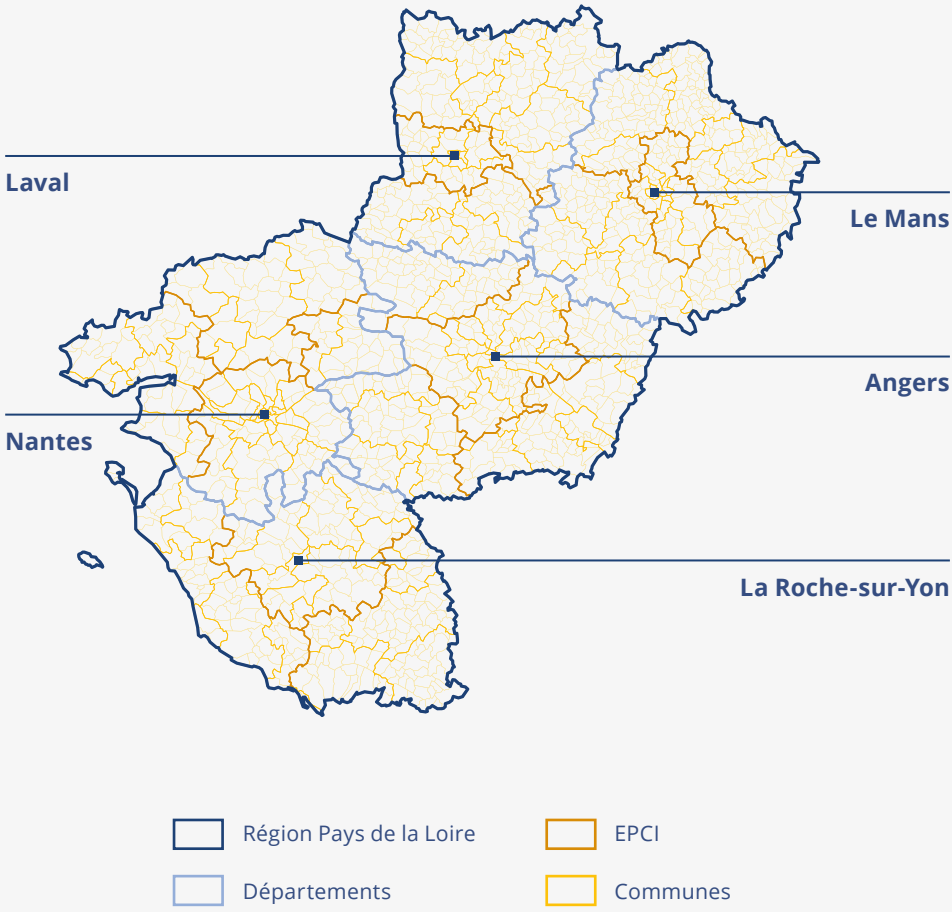
Par ailleurs, certaines collectivités disposent d'un statut particulier. Il s'agit de territoires pour lesquels une seule assemblée exerce des compétences dévolues en droit commun à plusieurs niveaux de collectivités : la région et le département dans le cas de la Guyane, de la Martinique (depuis 2015), de la Corse (depuis 2018), ou le département et le groupement fiscalité propre dans le cas de la Métropole de Lyon (créée en 2015), ou bien la commune et le département dans le cas de la Ville de Paris (nouveau statut créé en 2019) ; le département de Mayotte est également une collectivité à statut particulier mais exerce principalement les compétences d'un département.

5 La dotation globale de fonctionnement (DGF) des communes a été créée par la loi du 3 janvier 1979.

▶ Figure 1 - Les Pays de la Loire, illustration de l'emboîtement des quatre niveaux de collectivité

Les collectivités locales forment 4 échelons territoriaux.

La région
… se décompose en départements
… qui se découpent en communes,
… et les EPCI* à fiscalités propres regroupent des communes sans respecter obligatoirement les limites départementales ou régionales.



* EPCI à fiscalité propre (établissement public de coopération intercommunale).

Ainsi, pour la répartition de la DGF, la population totale authentifiée par l'Insee est majorée du nombre de résidences secondaires situées sur le territoire de la commune (ainsi que des places de caravane situées sur les aires d'accueil des gens du voyage). Ce concept « élargi » de population communale vise ainsi à se rapprocher de la population que la commune est conduite à administrer et qui est susceptible de lui faire supporter des charges supplémentaires. Les communes touristiques, sur le territoire desquelles les résidences secondaires sont davantage répandues, sont donc ici particulièrement concernées.

Une partie de la dotation versée annuellement à chaque commune[6] est ainsi fonction de sa population. Cette relation est affectée d'un coefficient logarithmique qui vise à tenir compte de l'observation selon laquelle la croissance tendancielle du niveau des charges par habitant n'est pas linéaire et proportionnelle à la population de la commune, mais que le surcoût des charges par habitant a tendance à diminuer à mesure que la population communale augmente *(Gilbert et Guengant, 2005).*

Plus largement, la population des collectivités est utilisée dans le calcul des dotations et fonds de péréquation, à la fois en étant combinée à d'autres indicateurs de ressources et de charges (potentiel financier par habitant, revenu moyen…), mais aussi en constituant le déterminant principal du niveau de l'attribution au titre des différentes enveloppes allouées[7].

Dans le même temps, le suivi des recettes (y compris fiscales) et des dépenses (de fonctionnement et d'investissement) a longtemps constitué le cœur des travaux statistiques relatifs aux collectivités locales. Du point de vue économique, le rôle d'investisseur des collectivités locales, en particulier des communes, nécessite en effet un suivi précis et régulier, en raison notamment de son effet sur le tissu économique local. Pour la commune elle-même, ces éléments de connaissance sont également indispensables afin de préparer au mieux son budget *(DGCL et DGFiP, 2020).*

▶ Appréhender les collectivités locales dans leur rôle d'employeurs

Le paysage des collectivités locales est complexe : aujourd'hui, ce sont plus de 60 000 structures[8] qui entretiennent des liens entre elles ou avec les autres niveaux d'administration publique dont l'État. Le besoin de compréhension s'est progressivement élargi à de nouvelles questions, en lien avec leurs moyens en personnel : fin 2020, l'ensemble des collectivités locales employait plus de 1,9 million d'agents, soit plus du tiers des effectifs de la fonction publique.

6 La répartition de la DGF, dans ses différentes composantes, est réalisée selon des critères très précis, par le bureau des concours financiers de l'État de la direction générale des collectivités locales.

7 À titre d'exemple, l'attribution au titre de la fraction « bourg-centre » de la dotation de solidarité rurale (DSR) est déterminée en multipliant le nombre d'habitants « DGF » par un indice de potentiel financier par habitant et d'effort fiscal.

8 En comptant les établissements publics locaux (comme les caisses des écoles ou les centres communaux d'action sociale).

Au-delà de la connaissance du nombre d'agents au sein des collectivités locales et son évolution, les besoins portent sur une représentation des spécificités des emplois occupés (catégories hiérarchiques[9], filières , cadres d'emploi, grades), au regard des autres versants de la fonction publique[10]. De même, les analyses font ressortir les différences entre collectivités selon leur type et leur nombre d'habitants.

Parallèlement, la demande porte aussi sur les rémunérations. Il s'agit alors de mesurer le salaire moyen, ou les quantiles de la distribution des salaires au niveau global, par niveau de collectivité, par catégorie hiérarchique, par cadre d'emploi[11], par filière. Les utilisateurs cherchent aussi à connaître la part entre le traitement indiciaire (le salaire) et le régime indemnitaire (les primes) pour calibrer des mesures salariales nationales.

Suivre l'évolution des structures et des compétences exercées

Le suivi statistique des structures locales a d'abord consisté en un décompte annuel du nombre de communes ***(encadré 3)***. Cette mission est assurée par l'Insee à travers le Code officiel géographique.

Le niveau communal[12] pour la France correspond aux unités administratives locales pour Eurostat. Plus globalement, en Europe, *« les niveaux d'administration locale se caractérisent dans une majorité de pays par un modèle à deux niveaux, le niveau communal et le niveau régional (Autriche, Hongrie, Irlande, Pays-Bas, Portugal, Suède, Suisse) et pour une minorité, les plus grands d'entre eux, majoritairement fédéraux, par un modèle à trois niveaux, présentant un niveau intermédiaire entre les communes et les régions (Allemagne, Espagne, Italie, Pologne). Parmi les pays européens, l'intercommunalité est largement répandue même si, le plus souvent, il s'agit de mécanismes souples et non institutionnalisés » (Sénat, 2009).*

> ***Le développement de nouvelles structures intercommunales et les modifications dans la répartition des compétences ont généré une demande d'outils de suivi statistique précis sur ces sujets.***

En France, principalement à partir des années quatre-vingt-dix, le développement de nouvelles structures intercommunales et les modifications dans la répartition des compétences ont généré une demande d'outils de suivi statistique précis sur ces sujets. Des informations ont ainsi été progressivement recueillies sur les groupements de communes à fiscalité propre[13] ou sans fiscalité propre, notamment leur périmètre, leur mode d'organisation et de financement ainsi que leurs compétences.

9 Les emplois des fonctionnaires sont répartis en trois catégories hiérarchiques (A, B et C), suivant le niveau de recrutement.

10 En France, la fonction publique est composée de trois versants (État, territoriale et hospitalière), qui emploient 5,61 millions d'agents, soit un salarié sur cinq (DGAFP, 2021).

11 Les cadres d'emplois regroupent les fonctionnaires territoriaux qui sont soumis au même statut particulier. Les cadres d'emplois comprennent plusieurs grades et sont eux-mêmes regroupés en « filière ». La fonction publique territoriale compte 53 cadres d'emplois répartis en 10 filières.

12 Au 1er janvier 2010 on dénombrait 36 682 communes et au 1er janvier 2022, on en compte 34 955, soit une diminution de 1 727 communes, du fait du développement récent des fusions de communes, sous la forme de « communes nouvelles ».

13 Un groupement de communes à fiscalité propre est une structure intercommunale qui à la différence d'un EPCI sans fiscalité propre (syndicats) est pourvu d'un pouvoir en matière fiscale.

Le recueil des informations relatives aux compétences exercées localement par les intercommunalités est essentiel pour établir une vision d'ensemble sur ce domaine, en répondant aux besoins des décideurs publics, avec des résultats mobilisables à différentes échelles géographiques. Ce suivi est réalisé grâce aux mises à jour effectuées par les préfectures dans un système d'information dédié[14], à partir des textes codifiant les statuts de chaque intercommunalité.

▶ Mieux apprécier les disparités entre collectivités locales

> ***Depuis quelques années, les travaux statistiques relatifs aux collectivités locales visent à mesurer les disparités des situations.***

Depuis quelques années, les travaux statistiques relatifs aux collectivités locales visent à mesurer les disparités des situations, en tenant compte des contextes socio-économiques qui leur sont propres. L'examen révèle une grande hétérogénéité des structures, que ce soit sur les plans financiers (du point de vue des ressources, des dépenses, du recours à l'endettement), fiscaux ou des ressources humaines.

Les statistiques par taille de population, tout particulièrement pour la situation financière des communes et des intercommunalités, sont une demande assez ancienne qui a pu être satisfaite depuis déjà plusieurs années. Elle répond en partie au besoin d'indicateurs de disparités. De ce point de vue, les résultats en termes de ratios financiers des communes selon leur taille (dépenses ou recettes par habitant, dépenses d'équipement rapportées à la population, part des dépenses de personnel dans le total des dépenses de fonctionnement, délai de désendettement, etc.) permettent de dresser des analyses donnant du sens pour caractériser la grande diversité des situations, à la fois en termes statiques et dynamiques[15].

Les communes peuvent être différenciées selon leur nombre d'habitants, leur caractère touristique, leur appartenance à des zones de montagne, leur caractère urbain ou rural. Ces paramètres peuvent être également croisés entre eux.

▶ Le recours aux sources administratives comptables…

Pour répondre à l'ensemble de ces besoins, plusieurs sources de données sont disponibles. Elles couvrent à la fois les domaines d'intérêt plus anciens, comme la démographie et les finances, et les sujets plus récents, ressources humaines des collectivités et suivi des structures territoriales ou des politiques publiques mises en œuvre.

14 Il s'agit de l'application *Aspic-Banatic* (Accès des services publics aux informations sur les collectivités – Base nationale sur l'intercommunalité) de la DGCL. Celle-ci est en cours de refonte pour améliorer en particulier les traitements relatifs au suivi des compétences. Voir *https://www.banatic.interieur.gouv.fr/V5/accueil/index.php*.

15 Voir par exemple (*Büsch, 2018*).

▶ Encadré 3. Une grande diversité parmi les 60 000 structures locales

Au 1er janvier 2022, la France métropolitaine et les DOM comptent :

14 RÉGIONS

...au sens de conseils régionaux

94 DÉPARTEMENTS

...au sens de conseils départementaux*

6 COLLECTIVITÉS À STATUT PARTICULIER

- La collectivité de Corse
- Les collectivités territoriales uniques de Martinique et de Guyane
- La Métropole de Lyon
- La Ville de Paris
- Le département de Mayotte

1 254 EPCI À FISCALITÉ PROPRE**

- 21 métropoles
- 14 communautés urbaines
- 227 communautés d'agglomération
- 992 communautés de communes

34 955 COMMUNES

Mais aussi :

25 PÔLES MÉTROPOLITAINS***

114 PÔLES D'ÉQUILIBRE TERRITORIAL ET RURAL (PETR)****

11 ÉTABLISSEMENTS PUBLICS TERRITORIAUX (EPT)

...au sein de la métropole du Grand Paris

8 882 SYNDICATS DE COMMUNES OU MIXTES

15 600 ÉTABLISSEMENTS PUBLICS LOCAUX (EPL)*****

- Centres communaux ou intercommunaux d'action sociale (CCAS – CIAS)
- Caisses des écoles
- Régies autonomes
- Centres de gestion de la fonction publique territoriale
- Services départementaux d'incendie et de secours (SDIS)

* La création au 1er janvier 2021 de la Collectivité européenne d'Alsace a fait baisser ce nombre d'une unité.
** Établissements publics de coopération intercommunale.
*** Structures créées en 2010 pour renforcer des territoires urbains qui ne peuvent prétendre à devenir des métropoles.
**** Établissements publics à la disposition des territoires situés hors métropoles, ruraux ou non.
***** Données de 2020.

À l'exception du recensement de la population et de la base Siasp (voir *infra*), la connaissance des collectivités locales s'appuie d'abord sur une offre de données conçues en dehors de la statistique publique : il s'agit des sources administratives sur les comptes des collectivités ou relatives à leurs moyens humains[16]. En revanche, l'outil de suivi des intercommunalités[17], de leur périmètre communal et de leurs compétences est sous la seule responsabilité du SSM Collectivités locales, même si les préfectures contribuent à sa mise à jour.

L'analyse de la situation financière des collectivités repose sur l'examen de l'ensemble de leurs dépenses et recettes et de leur endettement. Elle se fonde sur l'exploitation des sources de la Direction générale des Finances publique du ministère des Finances (DGFiP). Les données utilisées proviennent des comptes de gestion disponibles par année civile pour chaque niveau de collectivité. Une version provisoire, déjà très complète, des comptes de l'année *n* est mise à disposition du SSM Collectivités locales fin avril *n+1*.

Les écritures comptables sont retracées dans des comptes ; la liste de ces comptes figure dans des *nomenclatures*, mises à jour chaque année en fonction de l'évolution des besoins et de la réglementation.

La comptabilité est toujours tenue par nature. Ce classement regroupe les opérations selon les caractéristiques du mouvement comptable enregistré (constructions, achats de terrain, frais financiers, frais de personnel, subventions, prestations sociales versées, etc.). Il facilite l'analyse financière et le contrôle des comptes. Il s'applique à toutes les communes.

Le budget, en revanche, peut être présenté et voté par nature, mais aussi par fonction dans les communes de plus de 3 500 habitants. Le classement par fonction regroupe les opérations selon leur destination ou usage (enseignement, culture, santé, aménagement, sport, etc.) ; il facilite l'analyse économique des opérations engagées, notamment les dépenses. Par la présentation croisée, il permet de connaître la nature des dépenses et des recettes pour chaque usage.

▶ … repose à la fois sur des nomenclatures par nature…

Les informations des comptes de gestion sont enregistrées selon **des nomenclatures par nature**, dans lesquelles sont codées les dépenses et les recettes, à la fois en section de fonctionnement et en section d'investissement[18], par poste comptable fin, défini pour chaque millésime selon des instructions comptables exhaustives[19]. Des opérations de bilan sont également disponibles dans les comptes de gestion. Le recours à l'endettement est uniquement possible pour financer des dépenses de la section d'investissement.

La codification des opérations comptables s'effectue selon un emboîtement qui permet d'aller du niveau le plus agrégé de la dépense ou de la recette (code sur un caractère) au niveau le plus détaillé (code jusqu'à 7 caractères).

16 Les balances comptables de l'ensemble des collectivités locales, le fichier des budgets primitifs des collectivités, le recensement des éléments d'imposition produits par la DGFiP, et dans le domaine de l'emploi, les données issues du rapport social unique (anciennement bilans sociaux).

17 *Aspic-Banatic* (voir *supra*).

18 Au sens d'un compte de résultat pour reprendre une notion de la comptabilité privée. Le décret n° 2012-1246 du 7 novembre 2012 relatif à la gestion budgétaire et comptable publique (art. 56) impose aux collectivités publiques le respect du plan comptable du secteur privé.

19 Pour plus de détail sur les balances comptables de la DGFiP et les traitements réalisés par le SSM collectivités locales sur ces données, voir (*Niel, 2021*).

Historiquement, il existait une nomenclature comptable propre à chaque niveau de collectivité (M14 pour les communes et intercommunalités, M52 pour les départements, M71 pour les régions – d'autres nomenclatures existent également pour certaines catégories d'établissement ou certains types de budgets). Une nouvelle nomenclature a été mise en place ces dernières années (M57) pour unifier les enregistrements comptables[20]. Ce référentiel est utilisé par un nombre croissant de collectivités. Il deviendra la norme à compter du 1er janvier 2024.

▶ … et sur des nomenclatures par fonction

Parallèlement, les dépenses sont retracées selon des **nomenclatures fonctionnelles** qui permettent de préciser la finalité économique de la dépense et reposant sur un niveau agrégé (en M57 : services généraux, sécurité, enseignement, formation professionnelle et apprentissage, etc.), ou détaillé : par exemple, la fonction « *enseignement, formation professionnelle et apprentissage* » de la nomenclature M57 se divise au niveau des sous-fonctions en « *services communs* », « *enseignement du premier degré* », « *enseignement du second degré* », « *enseignement supérieur* », « *cités scolaires* », « *formation professionnelle* », « *apprentissage* », etc. Des « sous-sous-fonctions » sont également prévues.

> ***L'utilisation de ces nomenclatures fonctionnelles n'est pas immédiate.***

L'utilisation de ces nomenclatures fonctionnelles n'est pas immédiate : outre le fait qu'elles doivent faire l'objet de travaux d'articulation entre elles pour proposer une nomenclature commune aux différents niveaux[21], le travail de codification par les collectivités elles-mêmes peut déboucher sur des résultats inexploitables du fait de codes aberrants. On retrouve ici une difficulté propre à l'utilisation des sources administratives issues de systèmes de gestion, et utilisées ensuite par le statisticien *(Rivière, 2018)*.

Sur la base des informations disponibles dans ces mêmes comptes de gestion, l'Insee prend en charge l'élaboration du compte des administrations publiques locales au sens de la comptabilité nationale, ce qui permet de connaître leur contribution au déficit (ou à l'excédent) public et à la dette publique, calculés selon les critères de Maastricht.

Les nomenclatures par nature et fonctionnelle des collectivités sont alors traduites dans les nomenclatures propres à la comptabilité nationale, grâce à des tables de passage. Au niveau européen, la nomenclature par fonction des comptes nationaux est connue sous le nom de COFOG *(classification of functions of government)*.

20 Sur les instructions comptables et budgétaires, voir (*DGCL et DGFiP, 2022*).
21 Par exemple, la M14 (bloc communal) détaille davantage les dépenses culturelles et sportives que la M52 (départements), qui se penche plus sur les dépenses sociales, ou que la M71 (régions) qui propose davantage de détails pour les dépenses de transports.

Le recours à d'autres sources budgétaires, comptables et fiscales

La Direction générale des Finances publiques donne également accès aux budgets primitifs des collectivités, c'est-à-dire votés en début d'année pour fixer les prévisions de dépenses et de recettes de l'année. Ces données ne sont, en revanche, pas disponibles en termes de dépenses par fonction.

Par ailleurs, chaque collectivité est tenue de produire, voter et publier annuellement le compte administratif de son budget principal ainsi que les comptes administratifs de ses différents budgets annexes. Le SSM Collectivités locales est destinataire des comptes administratifs des plus grandes collectivités[22].

Enfin, le SSM mobilise également le fichier de recensement des éléments d'imposition[23]. Cette base de données très complète fournit les résultats détaillés par collectivité sur la fiscalité directe locale. Il comprend donc les données relatives aux bases fiscales, aux taux d'imposition relatifs à chaque taxe locale pour chacune des collectivités concernées et au produit résultant de ces deux paramètres, à la fois pour les taxes dites « ménages » ou les impôts dits « économiques », versés par les entreprises ***(figure 2)***.

Quantifier les moyens humains des collectivités : la statistique publique à la manœuvre

Pour analyser les ressources humaines employées par les collectivités, le SSM utilise le système d'information des agents des services publics (Siasp), élaboré au sein de la statistique publique à partir de sources administratives exhaustives. Siasp rassemble des données individuelles sur les effectifs, les caractéristiques des emplois, les volumes de travail et les rémunérations des agents de l'ensemble de la fonction publique, dans ses trois versants (État, collectivités locales et hôpitaux). Cette base de données individuelles est produite par l'Insee chaque année. Elle était historiquement alimentée par les DADS (déclarations annuelles de données sociales) pour la partie relative à la fonction publique territoriale. Avec le passage progressif des collectivités locales à la DSN (déclaration sociale nominative, voir *(Humbert-Bottin, 2018)*), les informations de Siasp proviendront uniquement de cette source à partir du millésime 2023.

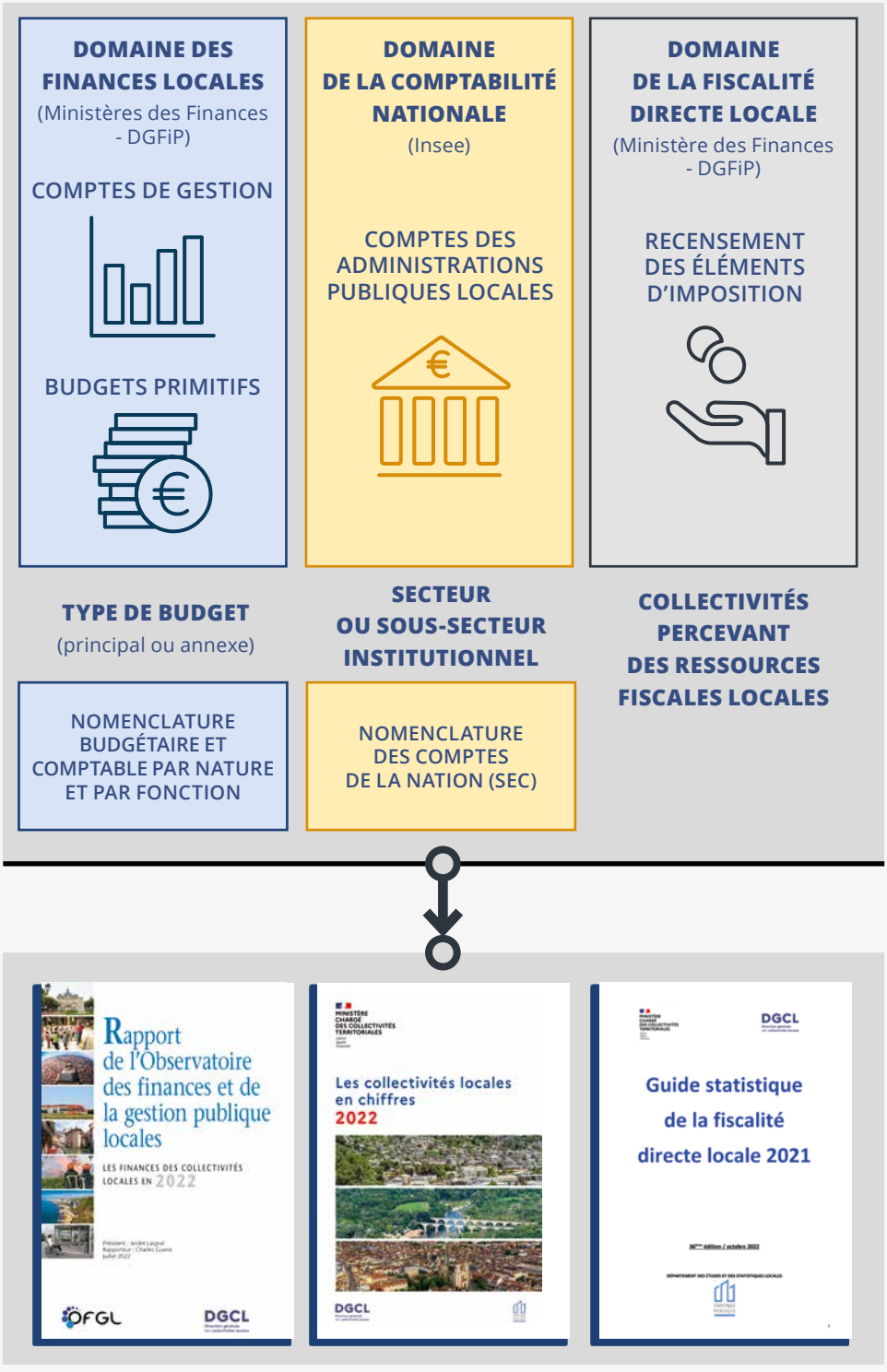
Avant la mise en place de Siasp pour la fonction publique territoriale, l'Insee réalisait chaque année, de 1980 à 2008, une enquête baptisée Colter[24] : celle-ci permettait de mesurer le niveau et les évolutions de l'emploi dans les collectivités territoriales et les services publics locaux. En parallèle, depuis 2003, l'information sur les rémunérations des agents des collectivités était disponible grâce aux DADS.

22 L'accès à ces informations, complémentaires à celles des comptes de gestion, permet au SSM de répondre à plusieurs besoins, comme l'identification de certains postes, notamment en matière d'emprunts structurés, dits aussi emprunts « toxiques ».

23 Le fichier de recensement des éléments d'imposition à la fiscalité directe locale (REI) est disponible sur la plateforme *https://www.data.gouv.fr/fr/* dans une version retraitée pour tenir compte du secret statistique et fiscal (données en *open data* mises à disposition par l'État).

24 Enquête sur les personnels de collectivités locales et des établissements publics locaux.

Figure 2 - D'amont en aval, les sources comptables et fiscales



Qualifier l'organisation RH des collectivités : une enquête administrative modernisée

Le SSM Collectivités locales dispose en matière de RH d'une source d'information qui lui est propre. Il s'agit d'une enquête *administrative*, historiquement connue sous l'appellation de « bilans sociaux », mais renommée « rapport social unique » (RSU) depuis la loi de transformation de la fonction publique d'août 2019. Cette enquête est réalisée en lien avec les centres de gestion de la fonction publique territoriale[25]. L'enquête est exploitée en collaboration avec le Centre national de la fonction publique territoriale (CNFPT). Les informations collectées portent sur des sujets complémentaires à ceux figurant dans le Siasp : l'organisation et le temps de travail, la santé et la sécurité au travail, les risques professionnels, la formation, les différents types de contractuels, l'action et la protection sociale ***(figure 3)***.

> ***Le rapport social unique est, par exemple, un moyen essentiel pour obtenir des informations sur les absences pour raison de santé ou sur la répartition des agents exerçant un emploi à temps non complet.***

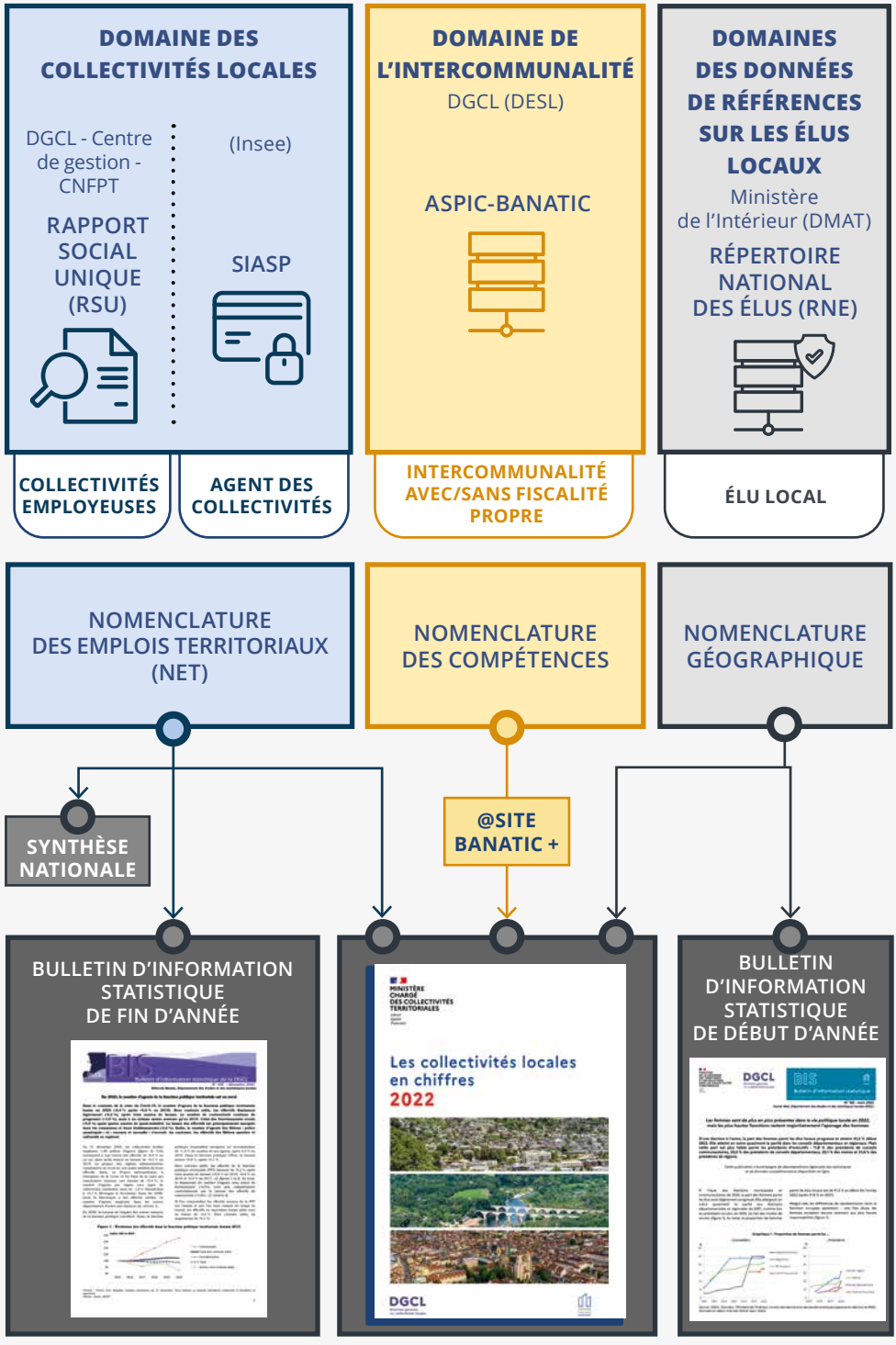
Son rôle est crucial pour la compréhension de différentes composantes de la fonction publique territoriale. Le rapport social unique est, par exemple, un moyen essentiel pour obtenir des informations sur les absences pour raison de santé ou sur la répartition des agents exerçant un emploi à temps non complet au sein des collectivités locales. Les postes à temps non complet sont une caractéristique de certains emplois des collectivités locales qui se définissent par le fait que c'est l'employeur lui-même qui propose un poste à temps non complet[26]. À la différence de Siasp qui rassemble des données individuelles sur près de 2 millions d'agents publics territoriaux, le rapport social unique ne fournit que des résultats agrégés pour chacune des 60 000 entités.

Le recueil de données auprès des collectivités, notamment sur ces sujets de ressources humaines, répond à des besoins de connaissance largement partagés, mais il n'est pas soumis au label d'intérêt général de la statistique publique ; il peut se heurter à des difficultés liées à l'équilibre fragile entre le principe de libre administration caractérisant la gouvernance des collectivités et le souhait de collecter des informations pour mieux suivre leurs activités.

Enfin, sur un sujet relatif à la gouvernance des collectivités et donc, par extension, lié au thème des ressources humaines, il existe une source particulière permettant de dénombrer et caractériser les élus locaux. Il s'agit du répertoire national des élus (RNE) produit par le ministère de l'Intérieur et disponible en *open data*. Le SSM Collectivités locales contribue à l'apurement des données et analyse les résultats, en particulier pour les sujets touchant à la parité parmi les élus.

25 Structures venant en appui des collectivités locales, en particulier pour les plus petites d'entre elles, dans le domaine du recrutement et de la gestion des ressources humaines.
26 Environ 15 % des agents de la fonction publique territoriale exercent un emploi sur un poste à temps non complet.

Figure 3 - D'amont en aval, les sources sur les ressources humaines, sur les intercommunalités et sur les élus



▶ Des outils spécifiques pour suivre les structures et leurs politiques publiques

Le suivi des structures territoriales et de leurs compétences repose principalement sur une application *ad hoc* (*Aspic-Banatic*[27]), mise à jour par les préfectures mais dont le SSM Collectivités locales a la responsabilité. Les informations disponibles portent sur la composition communale des EPCI (établissements intercommunaux de coopération intercommunale), leurs caractéristiques en termes d'immatriculation au répertoire Sirene, de gouvernance et de compétences exercées. Les compétences exercées par les intercommunalités[28] sont déterminées en fonction d'un référentiel fondé sur les textes juridiques en vigueur, en particulier dans le Code général des collectivités territoriales (CGCT) *(DGCL, 2019).*

Dans le domaine des politiques sociales, le service statistique ministériel relatif aux Solidarités et à la Santé (Drees) recueille chaque année des informations agrégées sur les moyens engagés et les actions menées par les départements en matière d'insertion des bénéficiaires du RSA, d'accompagnement des personnes handicapées ou en perte d'autonomie, de protection de l'enfance, de protection maternelle et infantile ainsi que sur d'autres activités d'action sociale. Un recueil de données individuelles est également mis en place actuellement, afin d'enrichir les analyses par un suivi des parcours de ces bénéficiaires. S'agissant des communes et intercommunalités, la Drees élabore ponctuellement des enquêtes et travaux d'analyse sur l'action sociale de ces collectivités. *(Drees, 2021 ; 2022).*

▶ Derrière la collectivité locale, plusieurs unités statistiques

En matière comptable et financière, les bases de données comptables de la Direction générale des Finances publiques identifient deux types d'unités statistiques :

- les collectivités sont d'abord prises en compte sous forme d'unités légales, auxquelles on a donc attribué un numéro Siren ;
- leur sont ensuite associés des « types de budget », identifiés chacun par des numéros Siret : un budget dit « principal » et plusieurs budgets dits « annexes ».

Le budget principal d'une collectivité correspond à l'équilibre de ses recettes et de ses dépenses dans ce qui constitue le cœur de son activité économique et sociale. Les budgets annexes représentent des activités que la collectivité isole en raison de leurs spécificités. Par exemple, la commune de Lyon, outre son budget principal, dispose de trois budgets annexes : l'auditorium-orchestre national de Lyon, le théâtre des Célestins, les Halles Paul Bocuse. Le poids des budgets principaux est largement majoritaire : en 2020, les dépenses comptabilisées dans les budgets annexes représentaient 10,7 % du total des dépenses des budgets principaux[29].

27 Voir *supra*.
28 Certaines des compétences sont exercées de manière obligatoire (par exemple, actions de développement économique dans les conditions prévues à l'article L. 4251-17), d'autres sont facultatives (comme les opérations programmées d'amélioration de l'habitat).
29 Jusqu'en 2017, les statistiques produites et diffusées par le SSM collectivité locales n'ont concerné que les budgets principaux.

En matière fiscale, le champ se restreint dans la mesure où l'unité statistique coïncide avec les seules collectivités pourvues du pouvoir fiscal en matière de vote des taux de fiscalité directe locale : communes, intercommunalités à fiscalité propre, département et régions, y compris collectivités territoriales uniques.

En matière de données sur le personnel, les informations disponibles reposent sur une source constituée de données individuelles (SIASP). L'unité statistique est donc ici l'emploi occupé par les agents des collectivités, distingué selon qu'il s'agit d'un emploi principal, d'un emploi secondaire (un agent pouvant occuper à la fois un emploi principal et un emploi secondaire au sein de plusieurs collectivités), ou d'un emploi annexe pour les très faibles volumes de travail. Les effectifs peuvent ensuite être agrégés au niveau d'une collectivité, d'un type de collectivité, etc.

Pour la source du « rapport social unique », les unités statistiques sont constituées de toutes les collectivités ayant le statut d'employeur et étant repérées comme unités légales dans le répertoire Sirene. Le champ est défini en fonction de la catégorie juridique au sens de Sirene.

En matière de suivi des intercommunalités, l'unité statistique correspond à l'unité légale enregistrée sous son numéro Siren.

▶ Mettre à disposition des statistiques agrégées et détaillées

> ***L'offre d'informations statistiques porte sur plusieurs types de produits adaptés à des besoins différents.***

L'offre d'informations statistiques porte sur plusieurs types de produits adaptés à des besoins différents.

En premier lieu, un document de synthèse sous forme d'annuaire statistique (***Les collectivités locales en chiffres***) offre, à la fois pour le grand public et les décideurs publics, un panorama d'ensemble des statistiques disponibles sur les collectivités locales dans les domaines des structures territoriales, financiers, fiscaux, des ressources humaines et des élus locaux. En complément de la version papier, la version électronique propose des séries longues téléchargeables *(DGCL, 2022)*.

Par ailleurs, des études thématiques, de nature récurrente ou inédite, sont publiées dans ***les bulletins d'information statistique (BIS)***. Avec une douzaine de numéros par an, ces publications visent à couvrir l'ensemble des sujets attendus par les utilisateurs (grand public ou spécialistes du domaine, chercheurs, élus locaux, etc.) : finances et fiscalité locales, emploi, rémunérations, temps de travail dans les collectivités, suivi des structures intercommunales, parité chez les élus locaux.

Sur le thème des intercommunalités, le portail dédié ***Banatic*** propose des fiches de synthèse par structure intercommunale ou des données en téléchargement (composition communale et compétences des intercommunalités, par exemple) destinées aux acteurs locaux eux-mêmes, aux autres décideurs publics mais aussi à la société civile et aux universitaires, notamment.

Des tableaux détaillés sont accessibles sur le portail Études et statistiques locales, en premier lieu ceux sur les comptes des collectivités et leurs budgets prévisionnels. Ceux-ci présentent, pour les départements et les régions, collectivité par collectivité, les principales recettes et dépenses, à la fois par nature et par fonction. D'autres tableaux détaillés sont également disponibles : le guide de la fiscalité directe locale, les résultats des bilans sociaux[30] et du rapport social unique (voir *supra*) et des bilans statistiques sur l'intercommunalité.

Ces informations statistiques peuvent être utilisées dans le cadre de travaux développés en dehors du SSM et portant sur des évaluations de politiques publiques. À titre d'exemple, certaines des publications du SSM Collectivités locales viennent appuyer ou éclairer des rapports de la Cour des comptes[31], voire des réponses contradictoires des collectivités elles-mêmes à la suite d'observations contenues dans les rapports de la Cour des comptes[32].

En outre, lors de leurs travaux d'évaluation et d'enquête, les missions d'inspection ou la Cour des comptes peuvent faire appel à des exploitations statistiques spécifiques de la part du SSM, illustrant ainsi une tendance à l'accroissement des exigences en termes de données. Ces demandes visent le plus souvent à suivre au mieux l'activité des collectivités territoriales en tenant compte de leur hétérogénéité. Les collectivités elles-mêmes peuvent également adresser des demandes de ce type dans un souci de comparabilité ou d'approfondissement statistique.

▶ L'apport du SSM dans les travaux de validation et d'enrichissement

Les données utilisées par le SSM Collectivités locales et sur lesquelles il élabore ses analyses sont en grande partie produites en dehors de son strict domaine d'intervention. Néanmoins, ses travaux contribuent activement à la validation ou à l'enrichissement des informations de base qui lui sont transmises.

Tout d'abord, dans le domaine des finances et de la fiscalité locale, le SSM enrichit les résultats « bruts » disponibles directement dans les bases de la DGFiP. Ainsi, un chantier très approfondi de consolidation des comptes entre budgets principaux et budgets annexes, d'une part, et entre niveaux de collectivités, d'autre part, est conduit chaque année : il permet de neutraliser les flux croisés entre ces différentes entités et d'obtenir ainsi une vision plus juste des opérations en jeu lorsqu'elles sont agrégées. Ce travail de consolidation est le fruit d'un investissement méthodologique conséquent mené pour la première fois en 2018 *(Niel, 2018).*

Par ailleurs, à partir des balances comptables, le calcul des ratios comptables dits « obligatoires », par taille de collectivités, complète l'information disponible en permettant de situer plus aisément chaque collectivité par rapport à une référence. Ces travaux sont également utiles aux collectivités locales elles-mêmes qui, dans le cadre de documents d'information publics à produire, ont besoin de situer leurs propres résultats.

30 Voir (*CNFPT et DGCL, 2022*).
31 Par exemple le rapport de la Cour des comptes sur les finances publiques locales 2020 (*CDC et CRDC, 2021a*), s'appuie page 42 sur le bulletin d'information statistique n°148, portant sur les effectifs de la fonction publique territoriale.
32 Voir la réponse du président de l'association des Régions de France au rapport de la Cour des comptes sur les finances publiques locales 2021 (*CDC et CRDC, 2021b*), page 174, faisant directement référence au bulletin d'information statistique n° 150.

De même, en matière de fiscalité locale, des travaux spécifiques sont menés par le SSM pour calculer les "effets base" et les "effets taux". Ils permettent de décomposer les variations des produits de chaque taxe locale entre ces deux facteurs. L'effet base correspond à l'évolution que les produits fiscaux auraient connue à taux d'imposition constants, c'est-à-dire si les bases avaient été les seules à évoluer ; l'effet taux explique la part restante de l'évolution globale. Par exemple, quand un groupement perçoit une année une taxe alors qu'il ne la percevait pas l'année précédente, l'augmentation du produit qui en résulte est intégralement retranscrite dans l'effet taux : sa base imposable n'a pas changé (à contour du groupement identique) et comme l'effet base est calculé en multipliant cette base inchangée à des taux d'imposition constants, cet effet base est nécessairement nul. Par déduction, toute l'augmentation du produit provient, dans ce cas, de l'effet taux.

Dans le domaine des statistiques sur les agents des collectivités locales, l'Insee produit la base Siasp (voir *supra)*, mais le SSM contribue à la phase de validation. Sa connaissance du domaine lui permet notamment d'identifier d'éventuels « trous » de collecte, de repérer et statuer sur des évolutions atypiques des effectifs ou des rémunérations, de recodifier certains enregistrements mal codés en fonction des référentiels utilisés, comme dans le cas de la nomenclature des emplois territoriaux utilisée pour coder les grades des agents, et d'analyser la cohérence par rapport à des mesures réglementaires ou législatives connues.

Concernant les bilans sociaux et le rapport social unique, le SSM intervient dans tout le processus de production, depuis la collecte (en lien avec les centres de gestion), l'apurement, le redressement, la pondération, le calage (sur les données administratives) jusqu'à l'élaboration des indicateurs diffusés.

> ***Des retraitements statistiques sont mis en œuvre, de manière à tenir compte des spécificités institutionnelles caractérisant l'organisation et le fonctionnement de certaines collectivités.***

Enfin, quel que soit le domaine, des retraitements statistiques sont mis en œuvre, de manière à tenir compte des spécificités institutionnelles caractérisant l'organisation et le fonctionnement de certaines collectivités, ou certaines opérations, dans le cas comptable. Cela est notamment le cas lors des changements de statut de collectivités : la création en 2015 de la métropole de Lyon et des collectivités territoriales uniques de Guyane et de Martinique, ou la recentralisation du revenu de solidarité active à La Réunion en 2020 et en Guyane et à Mayotte en 2019 sont autant de situations ayant nécessité des retraitements pour suivre des évolutions à champ constant. Le SSM, en étant parfaitement inséré au sein de la direction générale des collectivités locales, dispose d'un atout majeur pour être informé en temps réel du contenu précis de ces changements institutionnels.

Retracer l'hétérogénéité des collectivités

Comme indiqué précédemment, l'évolution des besoins oriente de plus en plus les travaux du SSM vers la mise en évidence de l'hétérogénéité des agents économiques que constituent les collectivités locales, ainsi que la diversité de leur comportement en termes économiques et sociaux.

Cette tendance a été accentuée avec la crise sanitaire de la Covid-19. Placées en première ligne, les collectivités ont pu réagir plus ou moins fortement au choc économique et financier, que ce soit du point de vue de leurs ressources ou de leurs dépenses. La crise a donc conforté le besoin de situer les collectivités au regard des disparités qui les caractérisent : ainsi, les résultats au niveau national doivent s'accompagner de décompositions par catégorie de collectivité et par taille, en comparant les situations par quantiles. Il est également possible de faire apparaître des disparités en termes de profils singuliers rassemblant des collectivités autour de critères homogènes et quantifiables[33].

Faciliter l'accès à des données fines : des partenariats efficaces et des projets

Le SSM Collectivités locales est étroitement associé aux travaux de l'Observatoire des finances et de la gestion publique locales (OFGL). Le rapport sur les finances locales publié chaque année en juillet est ainsi réalisé en quasi-totalité par l'équipe du SSM Collectivités locales *(OFGL et DGCL, 2022).*

L'observatoire a récemment créé un portail permettant d'accéder directement et facilement aux résultats comptables détaillés des collectivités à partir des balances comptables produites par la direction générale des finances publiques. Cette direction est également un partenaire essentiel à l'activité du SSM Collectivités locales, à la fois en transmettant les données comptables et proposant d'accéder, sur un site *web* partagé, à l'ensemble des comptes individuels des collectivités sous forme d'un tableau retraçant leur équilibre comptable.

> ***En matière d'offre de données complémentaires, les collectivités locales sont elles-mêmes de plus en plus directement productrices d'informations quantitatives sur leurs propres activités.***

En matière d'offre de données complémentaires, les collectivités locales sont elles-mêmes de plus en plus directement productrices d'informations quantitatives sur leurs propres activités : la mobilité, la gestion de l'énergie, de l'eau, des déchets, etc. Elles tendent d'ailleurs à ouvrir largement l'accès à ces données sur leur propre site *web*[34], les rendant ainsi disponibles au service des citoyens et pour accompagner ou orienter la mise en œuvre des politiques publiques territoriales.

33 Voir la fiche sur les finances des départements en 2020 dans (*OFGL et DGCL, 2021*).
34 Exemple de la région Île-de-France : *https://data.iledefrance.fr/pages/home-open-data/.*

▶ *Open collectivités*, une plateforme fédératrice

Plus largement, les demandes d'accès aux données individuelles en *open data* font désormais partie du paysage et le SSM Collectivités locales contribue à répondre à ces besoins émergents. La demande visant à gagner en accessibilité et visibilité pour l'ensemble des informations statistiques relatives aux collectivités locales a été adressée en 2019 par le Conseil national de l'information statistique au SSM *(Cnis, 2019).* De là est né le projet « *Open collectivités* » qui a consisté en la création d'une plateforme *web* fédératrice des informations statistiques déjà existantes sur les collectivités locales, mais ne provenant pas uniquement du Service statistique ministériel Collectivités locales.

Le portail *www.open-collectivites.fr* comporte un volet relatif aux publications de la statistique publique, visualisées grâce à un flux avec la bibliothèque numérique de la statistique publique (BNSP) et un autre volet valorisant des chiffres clés sur chaque collectivité (données socio-économiques et financières). Ce portail propose en outre un recensement des données en *open data* proposées par les collectivités locales elles-mêmes et facilite leur accès (*https://www.open-collectivites.fr/plateformes-open-data-locales/*).

Ces élargissements en matière d'accès à des données publiques tierces vont très certainement avoir un impact sur la production des statistiques relatives aux collectivités locales ainsi que leur valorisation. Cette évolution récente nécessite d'en prendre la mesure pour s'interroger sur leur utilisation éventuelle au sein du SSM, de la même manière que les nouvelles bases de données ouvertes questionnent l'ensemble des travaux de la statistique publique. Ces interrogations sont trop récentes pour avoir débouché sur une mise en œuvre, mais on peut anticiper quelques difficultés, compte tenu notamment du manque d'harmonisation entre ces jeux de données, et singulièrement l'absence de référentiels communs.

▶ Bibliographie

- BOUINOT, Jean, 1978. La statistique à la direction générale des collectivités locales du ministère de l'Intérieur. In : *Courrier des statistiques.* [en ligne]. Janvier 1978. Insee. N°5, pp. 3-5. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06z97k0/f1.pdf*.
- BOUVIER, Michel, 2020. *Les finances locales, 18e édition.* 7 juillet 2020. Librairie générale de droit et de jurisprudence (LGDJ). Collection Systèmes pratiques, Entreprise, économie & droit. EAN 978-2-275077826.
- BÜSCH, Faustine, 2018. *Diversité des communes, cinq profils budgétaires et financiers.* [en ligne]. Décembre 2018. Bulletin d'information statistique n°129. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.collectivites-locales.gouv.fr/sites/default/files/Accueil/Etudes%20et%20statistiques/Documents%20de%20synth%C3%A8se/BIS/2019/bis_129_0.pdf*.
- CDC et CRDC, 2021a. *Les finances publiques locales 2021. Fascicule 1. Rapport sur la situation financière et la gestion des collectivités territoriales et de leurs établissements publics en 2020.* [en ligne]. 30 juin 2021. Cour des comptes et Chambres régionales et territoriales des comptes. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.ccomptes.fr/fr/publications/les-finances-publiques-locales-2021-fascicule-1*.
- CDC et CRDC, 2021b. *Les finances publiques locales 2021. Fascicule 2. Rapport sur la situation financière et la gestion des collectivités territoriales et de leurs établissements.* [en ligne]. 23 novembre 2021. Cour des comptes et Chambres régionales et territoriales des comptes. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.ccomptes.fr/fr/publications/les-finances-publiques-locales-2021-fascicule-2*.
- CNIS, 2019. *Les données statistiques sur les collectivités locales.* [en ligne]. Octobre 2019. Note du Conseil national de l'information statistique. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/02/Note-Stat-sur-les-Coll-territoriales-def.pdf*.
- CNFPT et DGCL, 2022. *Bilans sociaux 2019. Synthèse nationale des rapports sur l'état des collectivités territoriales au 31 décembre 2019.* Disponible en décembre 2022.
- DGAFP, 2021. *Rapport annuel sur l'état de la fonction publique. Édition 2021. Politiques et pratiques de ressources humaines. Faits et chiffres.* [en ligne]. 22 octobre 2021. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.fonction-publique.gouv.fr/files/files/publications/rapport_annuel/RA_2021_web.pdf*.
- DGCL et DGFiP, 2020. *Le guide du Maire. Édition 2020.* [en ligne]. E 978-2-11-155544-0. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.collectivites-locales.gouv.fr/files/Accueil/guide_du_maire_2020.pdf*.
- DGCL, 2022. *Les collectivités locales en chiffres 2022.* In : *site Collectivites-Locales.gouv.fr.* [en ligne]. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.collectivites-locales.gouv.fr/collectivites-locales-chiffres-2022*.
- DGCL et DGFIP, 2022. Les instructions budgétaires et comptables relatives aux collectivités locales. In : *site Collectivites-Locales.gouv.fr.* [en ligne]. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.collectivites-locales.gouv.fr/finances-locales/instructions-budgetaires-et-comptables*.
- DGCL, 2019. *Signification des variables des fichiers en téléchargement sur BANATIC*. [en ligne]. Juillet 2019. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.banatic.interieur.gouv.fr/V5/ressources/documents/document_reference/Banatic_Metadonnees2019.xlsx*.

- DREES, 2021. *Minima sociaux et prestations sociales. Ménages aux revenus modestes et redistribution. Édition 2021*. [en ligne]. Collection Panoramas – Social. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://drees.solidarites-sante.gouv.fr/sites/default/files/2021-09/Minima%20sociaux%202021.pdf*.
- DREES, 2022. [en ligne]. Données mensuelles sur les prestations de solidarité. In : *site de la Drees*. [en ligne]. [Consulté le 2 septembre 2022]. Août 2022. Disponible à l'adresse : *https://data.drees.solidarites-sante.gouv.fr/explore/dataset/donnees-mensuelles-sur-les-prestations-de-solidarite/information/*.
- GILBERT, Guy et GUENGANT, Alain, 2005. Évaluation de la performance péréquatrice des concours financiers de l'État aux communes. In : *Économie et statistique.* [en ligne]. 1er avril 2005. N° 373-2004, pp. 81-108. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/1376450/es373d.pdf*.
- 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 22 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.
- INSEE, 2008. Le SSM collectivités locales : interview de Jean-Luc Heller. In : *Courrier des statistiques.* [en ligne]. Mai-octobre 2008. N°124, pp. 44-48. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt44g/f1.pdf*.
- NIEL, Xavier, 2018. *Consolidation des comptes des collectivités locales : quel impact sur la mesure de la croissance des dépenses ?* [en ligne]. Novembre 2018. DGCL. Bulletin d'information statistique, n°126. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.collectivites-locales.gouv.fr/sites/default/files/Accueil/Etudes%20et%20statistiques/Documents%20de%20synth%C3%A8se/BIS/2018/bis-126_comptes-consolides-des-cl.pdf*.
- NIEL, Xavier, 2021. *Les comptes des collectivités locales publiés par la DGCL.* [en ligne]. 25 février 2021. Ministère de l'Intérieur, département des études et des statistiques locales (DESL) de la direction générale des collectivités locales (DGCL). Note méthodologique. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.collectivites-locales.gouv.fr/files/Accueil/DESL/2021/0_CL_en_chiffres_2021.pdf*.
- OFGL et DGCL, 2021. *Rapport de l'Observatoire des finances et de la gestion publique locales. Les finances des collectivités locales en 2021.* [en ligne]. Juillet 2021. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.collectivites-locales.gouv.fr/files/Accueil/DESL/2021/OFGL/OFGL_Rapport2021.pdf*.
- RIMBAUT, Christine, VERPEAUX, Michel et WASERMAN, Franck, 2021. *Les collectivités territoriales et la décentralisation. 12e édition.* 15 juin 2021. Dalloz. La documentation française, Sciences humaines & sociales. EAN 978-2-111574366.
- RIVIÈRE, Pascal, 2018. Utiliser les déclarations administratives à des fins statistiques. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 14-24. [Consulté le 22 juillet 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647013/courstat-1-5.pdf*.
- SÉNAT, 2009. Une absence de « modèle européen » unique mais des pratiques et des principes communs. In : *Rapport d'étape sur la réorganisation territoriale.* [en ligne]. 11 mars 2009. Rapport d'information n° 264 (2008-2009) de M. Yves Krattinger et Mme Jacqueline Gourault, fait au nom de la mission Collectivités territoriales, pp. 58-66. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.senat.fr/rap/r08-264-1/r08-264-11.pdf*.

▶ Fondements juridiques

- Loi n° 79-15 du 3 janvier 1979 instituant une dotation globale de fonctionnement versée par l'État aux collectivités locales et à certains de leurs groupements et aménageant le régime des impôts directs locaux pour 1979. In : *site de Légifrance*. [en ligne]. Mise à jour le 24 février 1996. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886671*.
- Loi n° 2014-58 du 27 janvier 2014 de modernisation de l'action publique territoriale et d'affirmation des métropoles. In : *site de Légifrance*. [en ligne]. Mis à jour le 28 janvier 2022. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000028526298*.
- Loi n° 2015-991 du 7 août 2015 portant nouvelle organisation territoriale de la République. In : *site de Légifrance*. [en ligne]. Mise à jour le 23 février 2022. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000030985460*.
- Loi n° 2019-828 du 6 août 2019 de transformation de la fonction publique. In : *site de Légifrance*. [en ligne]. Mise à jour le 1er mars 2022. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000038889182*.
- Décret n° 2012-1246 du 7 novembre 2012 relatif à la gestion budgétaire et comptable publique. In : *site de Légifrance*. [en ligne]. Mis à jour le 1er janvier 2022. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000026597003*.

Qu'est-ce qu'un répertoire ?

De multiples exigences pour un système complexe

Pascal Rivière *

« Simple liste » d'objets (individus, entreprises) à laquelle on associe des caractéristiques stables, un répertoire n'a rien pour impressionner a priori, *comparé à d'autres composantes d'un système d'information.*

Pourtant, en tant que référence centrale voire opposable dans des processus de gestion, notamment administratifs, le répertoire charrie avec lui de multiples enjeux. Il invite à sa table des utilisateurs individuels ou institutionnels aux intérêts divergents, en attente d'un niveau de service élevé, mêlant qualité du contenu et interopérabilité. Pour atteindre ce degré d'exigence malgré l'hétérogénéité des usages, le répertoire doit posséder des propriétés particulières, et se fonder sur toute une organisation pour le construire et le faire fonctionner.

Il en résulte un système riche et dynamique, avec pour colonne vertébrale une sémantique rigoureusement définie, ainsi qu'une infrastructure juridique ou conventionnelle. La vie de cette structure complexe s'organise autour de processus d'alimentation bien identifiés et normalisés, d'une vaste palette de services, et d'une démarche de maîtrise de la qualité s'appuyant sur une équipe de gestionnaires dédiée. Le pilotage d'un tel système requiert des prises de décision à plusieurs niveaux, et par conséquent des instances de gouvernance adaptées. Nous sommes donc à des années-lumière de la « simple liste »…

As a "simple list" of objects (individuals, businesses) with stable attributes, a register does not seem impressive at first glance, compared to other components of an information system.

However, as a central reference for operational processes, particularly administrative processes, a register raises multiple issues. Firstly, it often has numerous users who do not necessarily share the same objectives and agenda. Its expected level of service is high, as it combines data quality and interoperability needs. Reaching this level despite heterogeneity of uses requires specific properties, and a strong organization to ensure developments and day-to-day activities. Finally, a register can be seen as a powerful and dynamic system that relies on rigorous semantics, a legal architecture and a solid technical infrastructure. Its lifecycle is based on standardized data flows, a wide range of services, and quality control supported by a dedicated team. Managing such a system implies multi-level decision making, in other words a governance of the register. All this is very far from the "simple list"…

* Chef de l'Inspection générale, Insee,
pascal.riviere@insee.fr

Dans nos sociétés, le besoin de références communes à visée opérationnelle se perd dans la nuit des temps. Dès l'Antiquité, les calendriers, nécessaires à l'organisation efficace des échanges et à toute forme de planification, apparurent et se perfectionnèrent, sans toutefois se coordonner. Dans la métrologie antérieure à la Révolution française, « *les unités* [étaient] *incohérentes, leurs mesures diverses et les procédés de mesure incertains : ni caractère national, ni doctrine d'ensemble* » *(Lacombe, 1979)*. En 1789, on dénombrait en France presque autant d'unités de mesure que de lieux et de corps de métiers. À Paris, merciers, drapiers et marchands de toile avaient chacun leur aune, par exemple[1]. Muid, minot ou boisseau pouvaient représenter des mesures très diverses selon les situations. Elles pouvaient être liées à la matière : setiers de sel, mires d'huile, corbes de foin, mines de grains, penses de fromage, etc. Si les paysans, commerçant localement, s'en accommodaient peu ou prou, certains négociants n'y trouvaient pas leur compte[2]. Passer à une métrologie universelle pour la longueur, le temps, la masse, n'allait pas de soi, loin de là. On n'imagine pas aujourd'hui l'extraordinaire épopée que fut la création du système métrique *(Débarbat et Quinn, 2019)*, ni le rôle décisif qu'eurent les circonstances politiques favorables des débuts pour lancer le processus.

Dans tous les domaines d'activité, on s'est ainsi attelé à construire des cadres de référence communs, partagés, reconnus, publiés : des **référentiels**. Ce ne fut pas sans douleur, sans conflits de pouvoir *(Alder, 2005)*. À l'ère de l'informatisation et dans un monde de plus en plus complexe, les référentiels n'ont cessé de se développer, incluant de nouvelles dimensions applicatives, techniques, systémiques, sans que ne soit démentie la difficulté que représente leur construction et leur mise en œuvre.

Au sein des systèmes d'information, on peut distinguer plusieurs types de référentiels, mais deux en particulier méritent d'être isolés. En premier lieu, les **nomenclatures** *(Guibert, Laganier et Volle, 1971)*, ou classifications, répondent à un besoin permanent de catégoriser *(Bowker et Star, 2000)*[3] : nomenclature de professions *(Amossé, 2020)*, nomenclature d'infractions *(Camus, 2022)*, classification internationale des maladies (CIM), etc. Mais il existe un second type de référentiel particulièrement crucial, d'apparence simple en première approche : les **répertoires,** répondant à un besoin de clarté sur une « population[4] » de référence. Mais en quoi consistent-ils ? Selon quels principes, avec quels outils, méthodes, contraintes la tenue d'un répertoire s'organise-t-elle ? Répondre à ces questions et donner ses lettres de noblesse à une brique majeure et pourtant mal connue de l'univers des données est l'objet de cet article.

▶ Un répertoire, qu'est-ce que c'est ?

Commençons par une définition générale, abstraite : *un répertoire, c'est une liste d'instances d'une même entité.* Par exemple, pour une entreprise, la liste de ses fournisseurs, celle de ses clients, la liste de ses produits, etc. Les « instances » ne sont pas des concepts, des catégories qu'on invente mais des objets précis, qui naissent et qui peuvent disparaître.

1 L'instauration du système métrique décimal, 1790-1837 : *https://unesco.delegfrance.org/L-instauration-du-systeme-metrique-decimal-1790-1837-Focus-Memoire-du-Monde-2639*.
2 Ces préoccupations commerciales conduisirent à la revendication d'une uniformisation des mesures, dans les cahiers de doléances présentés aux États généraux en 1789.
3 Le chapitre introductif de l'ouvrage s'intitule « *To classify is human* ».
4 Terme à prendre au sens large : ce peut être une population de logements, par exemple.

Répertorier, c'est en quelque sorte inventorier, cataloguer, et le répertoire résulte de cette opération. Il se caractérise donc essentiellement par sa **structure** : c'est une liste « à plat », une liste d'objets avec une certaine pérennité, une certaine épaisseur temporelle (répertoire de véhicules, mais pas répertoire d'accidents, par exemple). La structure de répertoire n'implique pas systématiquement la dimension référentielle : un répertoire d'amis sur un *smartphone*, ou bien le répertoire d'une cantatrice (liste d'œuvres) n'ont pas nécessairement vocation à servir de référence.

Car « référentiel » signifie « qui a trait à la référence ». Le terme existe depuis longtemps dans de nombreux domaines comme la physique (référentiel galiléen), la linguistique, la psychologie ou l'éducation *(Cros et Raisky, 2010)*. Dans un système d'information, un référentiel est une *source d'information reconnue*, contenant des données « maître » (*master data*), et dans laquelle on peut puiser. Les répertoires sont une de ces sources. Dans ce qui suit, les répertoires désigneront implicitement des **répertoires avec un statut référentiel** (on utilisera indifféremment l'un ou l'autre terme). On peut citer par exemple des répertoires d'individus[5], d'étudiants[6], de professionnels de santé, d'électeurs[7] *(Demotes-Mainard, 2019)*, d'entreprises *(Bernard, 1995)*, de véhicules routiers, d'établissements d'enseignement[8], etc.

Mais quelles sont les caractéristiques permettant d'affirmer qu'un répertoire possède un tel statut ? Pour cela, *(Bizingre et alii, 2013)* distinguent cinq propriétés fondamentales : centralité, qualité, stabilité, unité de sens et interopérabilité ***(figure 1)***.

▶ Un répertoire : un positionnement central…

La centralité d'un répertoire signifie qu'il est « la référence » pour un ensemble d'acteurs, qui le reconnaissent en tant que tel, lui conférant ainsi sa légitimité. Ce serait en quelque sorte, et toutes proportions gardées, l'équivalent de *l'universalité* pour le système métrique. Cette légitimation implique un certain degré d'officialisation, qui se matérialise souvent par un arsenal juridique : par exemple, l'arrêté du 13 novembre 2013 relatif à la mise en place d'un répertoire national des établissements sanitaires donne un statut officiel à Finess[9], en décrivant notamment ses finalités, son contenu et le service qui le gère.

Dans certains cas, cette forte légitimité donne un poids important au répertoire, allant parfois jusqu'à le rendre opposable pour des actes administratifs. Cette position centrale est facilitée par la neutralité de l'organisme en charge de sa gestion. Mais cette centralité ne s'exprime pas dans l'absolu, pour tous usages : elle s'inscrit dans le cadre des finalités attribuées au répertoire[10], des enjeux auxquels il est censé répondre. Il existe ainsi des répertoires dont la finalité est essentiellement statistique : ainsi, pour le répertoire statistique des véhicules routiers (RSVERO), la finalité principale est « *la production de statistiques, d'études ou de rapports d'évaluation des dispositifs de politiques publiques* ». Pour d'autres,

5 Cf. l'article de Lionel Espinasse et Valérie Roux sur le RNIPP dans ce même numéro.
6 Avec l'INE (identifiant national étudiant).
7 On constate au passage que l'objet est lié à un rôle : un même individu peut être électeur, étudiant, client, etc. et ce sont là des objets bien distincts, associés à des systèmes d'information totalement différents.
8 Répertoire académique et ministériel sur les établissements du système éducatif (RAMSESE).
9 Fichier national des établissements sanitaires et sociaux.
10 Au minimum, on peut dire que les finalités d'un répertoire sont transverses, et non spécifiques à un métier.

> *L'encadrement juridique ou conventionnel imposé à un répertoire peut s'accompagner de contraintes, comme l'obligation d'alimenter le répertoire… qui offrent en contrepartie des garanties de qualité, fort utiles.*

les finalités, parfois nombreuses, sont administratives : ainsi, l'une des finalités de Finess est d'être l'autorité de référence pour les établissements qui demandent des accès numériques aux différents systèmes de santé.

Faire jouer au répertoire un rôle central requiert d'associer de nombreux acteurs à travers des instances de gouvernance : les organismes concernés (producteurs ou utilisateurs) s'expriment avec chacun leurs propres enjeux, qui ne sont pas toujours cohérents avec ceux du répertoire. Des arbitrages sont donc indispensables.

Figure 1 - Cinq propriétés d'un répertoire avec statut de référentiel

Enfin, l'encadrement juridique ou conventionnel imposé à un répertoire peut s'accompagner de contraintes, comme l'obligation d'alimenter le répertoire… qui offrent en contrepartie des garanties de qualité[11], fort utiles. Réciproquement, le fait que les données soient reconnues comme fiables lui confère une plus grande légitimité, facilitant sa position de pivot : en effet, décréter qu'un répertoire est central, sans se préoccuper de la réalité de son contenu est insuffisant ; il faut qu'il soit de qualité acceptable… Comment appréhender cette qualité ?

▶ … une exigence de qualité…

La qualité des données revêt en pratique des aspects divers *(Di Ruocco, Scheiwiler et Sotnykova, 2012)* : pertinence, exactitude, complétude, consistance, accessibilité, etc. Vis-à-vis de cette grille d'analyse générale, le cas d'un répertoire présente des spécificités. Ainsi, de par sa position de référence ultime, il est par construction plus difficile d'analyser sa qualité en le comparant avec d'autres sources de données… car un référentiel, c'est justement ce à quoi les autres sources se comparent *(Bizingre et alii, 2013)*[12]. Il soulève ainsi des difficultés particulières pour élaborer des indicateurs *(Rivière, 2005)*, par exemple pour mesurer le taux de « faux actifs »[13] .

En dehors des délicates questions d'exactitude, on peut insister sur :

- l'exhaustivité, sur un champ préalablement défini *(Wallgren et Wallgren, 2016)* ; deux défauts symétriques sont possibles, la sur-couverture (unités présentes à tort, par exemple des doublons), ou la sous-couverture (des unités qui devraient être présentes, ne le sont pas). La sous-couverture est toujours plus délicate à traiter, car on ne sait pas où sont les manques ;
- la fiabilité des flux d'entrée, ce qui requiert le développement d'une batterie de contrôles avant mise à jour du répertoire, tests de diverses natures pour vérifier que les données sont conformes et s'assurer que le risque d'introduire des erreurs est limité *(Sureau et Merlen, 2021) ;*
- la fraîcheur, c'est-à-dire la rapidité de prise en compte des événements de mise à jour : cela peut varier d'un répertoire à l'autre, voire d'une variable à l'autre, en fonction des besoins, des usages. Pour un répertoire administratif comme Sirene, l'exigence est plus importante : le délai de prise en compte d'une création d'entreprise y est très court, en raison des conséquences directes pour les entreprises. À l'inverse, les répertoires dits « statistiques », sans enjeu administratif individuel, peuvent avoir une exigence de fraîcheur moins forte ;
- … bien d'autres aspects tels que la traçabilité des événements, voire l'historicisation des données, la documentation, ainsi qu'une autre dimension de la qualité, la qualité de service (voir *infra* la partie sur l'*interopérabilité*).

Les deux propriétés suivantes, stabilité et unité de sens, renvoient à la sémantique des données.

11 Le répertoire national commun de la protection sociale (RNCPS) constitue un bon contre-exemple : il est moins contraignant que d'autres, ce qui peut avoir des impacts sur la qualité, du fait de l'absence de déclaration de certains événements.
12 Cf. chapitre 6, consacré à la qualité.
13 Unités que le répertoire suppose « vivantes » mais qui ne le sont pas en réalité.

▶ … une nécessaire stabilité…

La **stabilité** est une propriété très particulière d'un répertoire *(Régnier-Pécastaing, Gabassi et Finet, 2008)*[14], qui le distingue fondamentalement d'autres bases de données et en particulier des bases de gestion. Il doit en effet contenir des données structurantes, qui ne changent pas en permanence. À cette aune, l'adresse du siège[15] d'une entreprise est une information de référence, mais pas le montant de ses investissements. *Pour chaque entité d'un répertoire,* le principe doit être que les données changent peu dans le temps[16]. « Stable » ne signifie pas que les données sont « immuables » : par exemple, une facture est une information figée, dont le contenu ne changera plus, mais qui n'a rien à voir avec les référentiels.

Il est difficile de fonder la notion de stabilité sur la fréquence de mise à jour, trop dépendante de phénomènes métier et difficilement accessible. Plus simplement, la stabilité des données d'un répertoire peut être caractérisée par le fait que leurs évolutions doivent être indépendantes des processus métier qui les utilisent. Par exemple, dans un référentiel d'hôpitaux, le nombre de lits pourrait avoir une dimension référentielle[17], alors que le nombre de lits *occupés*, lié aux processus métier (accueil et départ des patients) et qui peut changer tous les jours, n'est en rien de nature référentielle.

La conception d'un répertoire suppose donc de porter une grande attention au choix des variables et de veiller à se limiter à celles qui sont essentielles, structurantes, stables et caractérisant l'objet : ainsi, lorsqu'une entreprise gère un répertoire de clients, on y trouvera par exemple des données de contact (numéro de téléphone, adresse mail), le canal de contact préférentiel (téléphone, courriel, site *web*, courrier, etc.), voire le type d'achat, mais certainement pas les achats eux-mêmes.

▶ … le respect d'une unité de sens…

Avec la notion d'**unité de sens**, on souligne que le répertoire doit posséder une certaine homogénéité sémantique : on ne peut donc pas y mettre n'importe quoi pour céder aux demandes. L'idée est d'éviter que le répertoire devienne un fourre-tout dans lequel on ajoute des informations sous prétexte qu'elles sont utiles à certains utilisateurs. Cela vaut d'abord pour les entités : les responsables Sirene ont ainsi été sollicités pour immatriculer des éoliennes, ou des ruches par exemple… ce qui n'a rien à voir avec la notion d'entreprise ou d'établissement. De telles demandes se comprennent : les utilisateurs

> ***Le répertoire doit posséder une certaine homogénéité sémantique : on ne peut donc pas y mettre n'importe quoi pour céder aux demandes.***

14 Pp. 45-47.
15 La variable adresse est un sujet délicat ; c'est volontairement que l'on a pris l'exemple de l'adresse du siège. En effet, une adresse est liée à un rôle : adresse administrative, de l'accueil, du siège, de livraison, etc. On peut aussi parler d'adresse géographique, d'adresse postale, ce qui complique encore les choses…
16 Ne pas se méprendre ici : quand le répertoire contient des millions d'entités, de nombreuses mises à jour seront effectuées sur l'ensemble du répertoire… mais très peu entité par entité. On peut penser par exemple à la rareté des événements affectant un individu (naissance, décès, mariage, changement de nom, etc.).
17 On simplifie beaucoup en écrivant les choses ainsi. Le nombre de lits renvoie plus généralement à la notion de capacité, mais il peut y avoir plusieurs capacités, comme la capacité administrative, par exemple.

du répertoire peuvent voir ce dernier comme un outil utile pour faciliter leur travail, sans se préoccuper du « bien commun », perçu comme théorique. La gouvernance d'un répertoire donne donc naturellement lieu à des tensions entre besoins des utilisateurs et nécessité de cohérence au niveau global.

L'unité de sens s'applique également aux données : dans un répertoire d'électeurs *(Demotes-Mainard, 2019)*, on va se limiter aux variables utiles pour le *rôle* d'électeur (commune de rattachement, bureau de vote), et ne pas ajouter des données sans rapport avec ce rôle (par exemple, nombre d'enfants). Plus généralement, un répertoire doit contenir « relativement peu » de variables en principe, car il est coûteux d'en assurer la qualité, et en particulier la fraîcheur ou encore la minimisation des données à caractère personnel, au titre du Règlement général sur la protection des données.

▶ … et la garantie d'une interopérabilité

Cette dernière propriété aborde un aspect plus technique, qui n'est pas spécifique aux référentiels : le service rendu aux utilisateurs par le « système » répertoire, ainsi que son insertion technique dans l'ensemble du système d'information. En effet, si le répertoire possède toutes les propriétés évoquées précédemment, mais qu'il n'est pas pratique à utiliser et à intégrer, il perd tout son intérêt.

L'interopérabilité[18] revient à considérer le référentiel non pas en termes de contenu, mais de services offerts, au sens informatique du terme[19] *(Régnier-Pécastaing, Gabassi et Finet, 2008)*, et d'interconnexion efficace avec les systèmes d'information qui l'environnent, partageant avec ceux-ci un même langage, en quelque sorte. Dire qu'il est interopérable, c'est dire qu'il est ouvert et accessible (par exemple *via* des API ou *Application Programming Interface*)[20].

On distingue traditionnellement interopérabilité technique (exigences de performance, de disponibilité, de sécurité, d'utilisation de standards techniques, etc.), interopérabilité syntaxique (format utilisé, par exemple XML[21], JSON[22]), et interopérabilité sémantique (identifiants, concepts, nomenclatures, etc.)[23].

Les répertoires se caractérisent souvent par un grand nombre d'utilisateurs, sans pouvoir connaître et encore moins maîtriser leurs usages ; d'où ce fort besoin qu'ils soient interopérables. À titre d'exemple, la présence d'identifiants reconnus (NIR, numéro Siren, etc.) est un facteur d'interopérabilité, car elle permet de faire le lien avec d'autres processus (par exemple l'utilisation du numéro Siren dans la déclaration sociale nominative ou DSN[24]).

18 Cf. la présentation du RGI (Référentiel général d'interopérabilité) dans *(DINSIC, 2015)*.
19 Cf. chapitres 5 et 6, sur les spécificités des référentiels en la matière.
20 Voir Définition de la CNIL : *https://www.cnil.fr/fr/definition/interface-de-programmation-dapplication-api*.
21 *Extensible Markup Language.*
22 Le *JavaScript Object Notation* (JSON) est un format standard utilisé pour représenter des données structurées comme les objets *JavaScript*.
23 On a cité plus haut le RGI, on peut aussi évoquer le cadre d'interopérabilité des systèmes d'information de santé *(Agence du numérique en santé, 2022)*.
24 *(Humbert-Bottin, 2018)*.

Derrière cette notion, c'est aussi tout le sujet de l'accostage du référentiel aux systèmes d'information qui est en jeu : qui doit s'accoster, comment… Cela comporte une dimension technique (par exemple une API à utiliser), mais aussi des règles de fonctionnement auxquelles on s'astreint : obligation pour telle application de s'accoster à tel référentiel, obligation de ne s'alimenter qu'à tel répertoire, etc. *Last but not least*, il ne faut pas oublier le sujet de la cohérence des répertoires entre eux, de leurs interactions, et les enjeux associés (par exemple, les liens entre Finess et Sirene, ou bien la mise en cohérence mutuelle SNGI – RNIPP[25]), qui relèvent clairement de l'interopérabilité.

▶ Le contenu d'un répertoire : des données d'identification…

Un répertoire est un ensemble d'entités de même type, et pour chacune d'elles, on trouve toujours peu ou prou le même type de données, liées au statut référentiel.

En premier lieu, l'identifiant : par exemple, le numéro Siret pour un établissement, le numéro RPPS[26] pour un professionnel de santé, le numéro INE[27] pour un élève, étudiant ou apprenti. Il doit être présent pour tout répertoire, puisque c'est ce qui désigne l'objet et l'officialise. Attribuer un identifiant à un objet se nomme *immatriculation* et revient à créer l'objet dans le répertoire. Le processus d'immatriculation est une opération majeure qui demande la plus grande rigueur ; en effet, ce processus ouvre accès à des services, ce qui poserait problème au cas où l'immatriculation est faite avec une fausse identité. L'identifiant est ainsi la clé d'accès à l'entité, clé qui permet aussi de faciliter les appariements entre fichiers portant sur le même objet. Une bonne pratique consiste à faire en sorte que cet identifiant soit non signifiant, c'est-à-dire qu'il ne contienne aucune information porteuse de sens ; cette pratique n'est d'ailleurs pas toujours respectée (voir le NIR).

> ***Les traits d'identité sont distincts de l'identifiant et ne le contiennent pas.***

Mais l'identifiant ne peut être créé à partir de rien ; il faut savoir « de qui on parle ». Ainsi, on doit disposer de *traits d'identité*, grâce auxquels on peut repérer sans ambiguïté l'objet : par exemple, nom-prénom-date de naissance pour un individu, raison sociale-adresse pour un établissement. Les traits d'identité sont distincts de l'identifiant et ne le contiennent pas.

Une précision importante : les traits d'identité ne doivent être présents (avec statut référentiel) que pour certains types de répertoires, dits **répertoires-socles** : ce sont eux qui formalisent le lien entre identifiant et traits d'identité. Les autres répertoires sont dits **répertoires adossés**. Par exemple, le REU, répertoire d'électeurs, est adossé au RNIPP, répertoire-socle d'individus ; autre cas, le RGCU, répertoire de carrières, est adossé au SNGI. Ainsi, dans un répertoire adossé, les traits d'identité, s'ils sont présents, ne constituent pas une référence : le lien entre traits et identifiant dépend complètement du répertoire-socle auquel il est adossé.

25 Cf. les articles de Joseph Préveraud de Vaumas sur le système national de gestion des identités (SNGI) et de Lionel Espinasse et Valérie Roux sur le répertoire national d'identification des personnes physiques (RNIPP) dans ce même numéro.
26 Répertoire partagé des professionnels de santé.
27 Identifiant national étudiant.

▶ … et d'autres données, référentielles ou non

Au-delà de ce noyau dur, un répertoire contient systématiquement des données de catégorisation, car les entités qui le composent ne sont pas toutes équivalentes. Par exemple, pour des établissements d'enseignement, on utilise un découpage du type : établissements du premier degré, du second degré, de l'enseignement supérieur ou de formation continue, etc. Un répertoire d'entreprises comme Sirene s'appuie quant à lui sur des nomenclatures officielles, ayant une existence et une gestion en dehors du répertoire : ainsi, pour caractériser l'activité principale, on se fonde sur la nomenclature d'activités française (NAF).

On trouve également, mais de façon moins systématique[28], des données pour positionner l'objet dans l'espace : adresse, géolocalisation, code lié à un zonage. Il faut également le situer dans le temps, à l'aide de dates-clés : dates d'ouverture, ou de création, et réciproquement de fermeture, ou de cessation. Plus généralement, la connaissance de l'historique des objets (étapes de création, cessation-fermeture, mais aussi fusions ou scissions, dans le cas des entreprises par exemple) peut se révéler indispensable.

Il est souvent nécessaire de compter parmi les données les identifiants d'autres objets liés[29], ou des références à des conventions[30].

À toutes ces données de fond, il faut ajouter des données dites *de gestion* ; ces données ne sont pas de nature référentielle, mais facilitent la gestion courante du répertoire. Par exemple, les données de contact[31] telles que l'adresse mail, l'adresse physique ou le numéro de téléphone associées à l'entité sont importantes lorsqu'il faut justement la contacter, pour effectuer des vérifications notamment.

Enfin, il est fort utile de disposer, pour certaines données, des métadonnées : provenance de la donnée, ou date de mise à jour, par exemple. Les conventions peuvent aussi se présenter comme des métadonnées.

Ainsi, construire une base contenant toutes les données nécessaires est une chose. Faire en sorte que cet ensemble de données soit de nature référentielle, et donc respecter les cinq propriétés citées précédemment, en est une autre. Toute une organisation doit être mise en place pour obtenir un tel résultat. Comment fait-on, concrètement ?

28 Ainsi, l'adresse va être référentielle pour une entreprise (sachant que, pour pimenter le tout, il peut y avoir plusieurs adresses), et ne le sera en général pas pour un individu.

29 Ainsi, pour un répertoire de professionnels de tel ou tel domaine (professionnels de santé, par exemple), il est important d'avoir, parmi les variables, l'identifiant de l'établissement auquel la personne est rattachée.

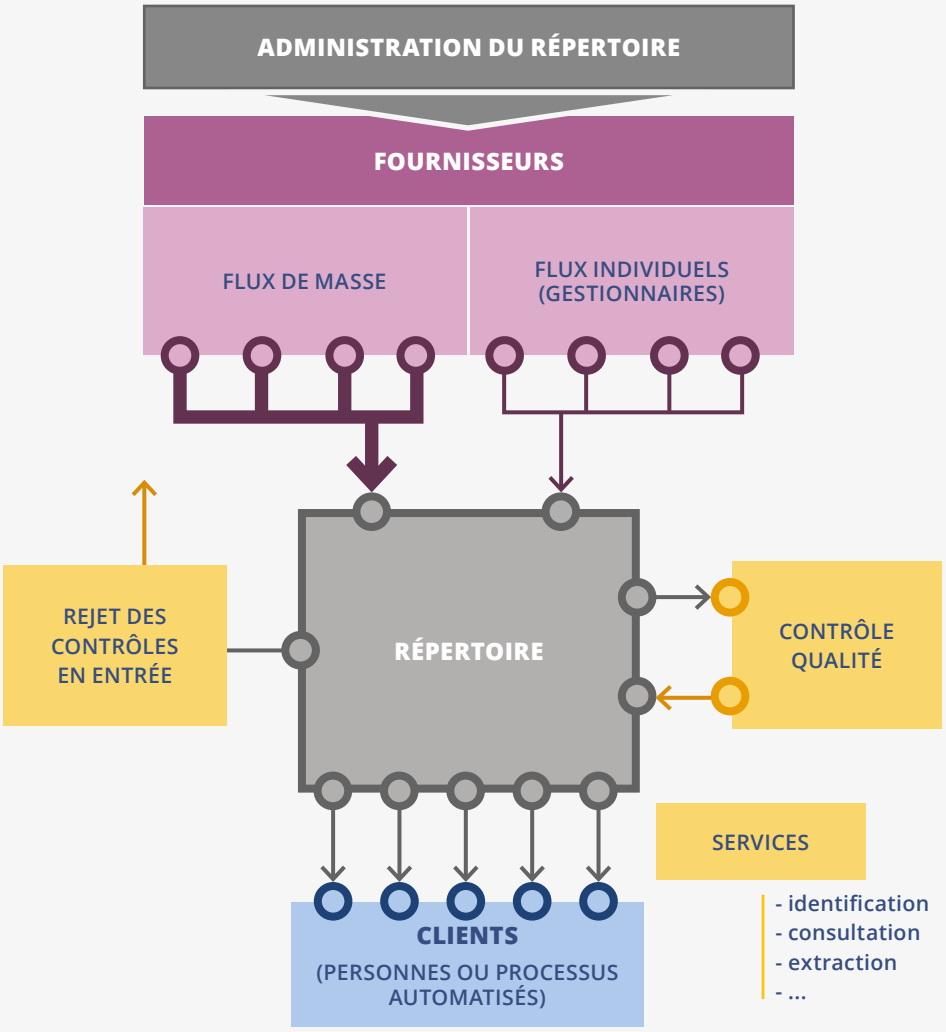
30 Lorsqu'un répertoire inclut une catégorisation d'activités (par exemple pour les répertoires d'entreprises, d'établissements de santé, d'établissements d'enseignement), certaines peuvent être encadrées par des conventions, dont il faut conserver la référence.

31 Cette différence entre données référentielles et de gestion est liée au rôle du répertoire, à ses enjeux. En particulier, les « données de contact » seront de nature référentielle dans le cas d'un répertoire de clients.

Un répertoire est un système vivant

L'image courante, lorsqu'on évoque une information de référence, consiste à dire qu'elle est « *gravée dans le marbre* ». Mais cette image est tout à fait inadaptée dans le cas d'un répertoire. En effet, ce dernier ne peut pas être figé, car le monde réel qu'il est censé décrire ne l'est pas : de nouvelles instances apparaissent (naissances, créations, ouvertures), d'autres disparaissent (cessations, fermetures, décès), et les données descriptives peuvent évoluer (par exemple la capacité d'accueil d'un établissement de santé). Ainsi, les grands répertoires sont affectés quotidiennement par de nombreux événements de mise à jour.

Figure 2 - Le répertoire, au centre d'un échange de flux



Un répertoire se caractérise également par les forts liens qu'il entretient avec de nombreux systèmes d'information : en amont pour le nourrir, en aval pour les services rendus aux utilisateurs. Ainsi, une image plus appropriée d'un répertoire serait celle d'une pompe aspirante refoulante d'informations, d'un cœur battant, régulièrement alimenté par des événements de mise à jour, et irriguant en permanence d'autres systèmes ***(figure 2)***. Car il n'a pas d'intérêt pris isolément, et ne prend sa signification que vis-à-vis de ses utilisateurs (fournisseurs ou consommateurs), variés et nombreux.

> ***Une image appropriée d'un répertoire serait celle d'une pompe aspirante refoulante d'informations, d'un cœur battant, régulièrement alimenté par des événements de mise à jour, et irriguant en permanence d'autres systèmes.***

Plus qu'une simple base de données, un répertoire est un véritable *système* dont les données de référence constituent le centre. À ce titre, on peut le caractériser selon plusieurs dimensions (entrée, sortie, contenu, boucle de rétroaction[32]) :

- les **flux d'alimentation** (unitaires, ou de masse), avec les différents niveaux de contrôle automatique de ces flux entrants ;
- les **services** rendus aux utilisateurs, aux autres systèmes d'information, qui engendrent des flux sortants ;
- la **structure** des informations qu'il contient (schéma de données, pour l'essentiel), donc le contenu, déjà évoqué ;
- les procédures de **contrôle qualité** internes, mêlant contrôles automatiques et vérifications manuelles effectuées par les gestionnaires.

▶ Maîtriser les flux d'entrée : le besoin de normaliser les échanges

Le système vivant ainsi décrit doit être alimenté régulièrement, avec des informations fraîches (autant que possible), à travers des « événements » qui requièrent une mise à jour : création, fermeture, changement d'adresse ou d'activité, toute forme de changement de statut, etc.

Ces modifications ne relèvent pas d'une mise à jour classique dans une base de données : le statut référentiel du répertoire entraîne des exigences particulières de maîtrise de la qualité, de traçabilité des modifications. Il faut donc en particulier mettre en place des contrôles automatisés permettant de filtrer les entrées dans le répertoire.

Plus précisément, on vérifie que le flux d'entrée transmis, le « message », est conforme à un standard : on peut imager cela en disant qu'on « branche » le flux de données sur le répertoire et on vérifie que la prise est bien conforme. Dans notre cas, le standard se présente sous la forme de règles à respecter pour toutes les données transmises, que l'on appelle **norme d'échange**.

32 Au sens systémique *(Wiener, 1948)* : entrée, sortie, boucle de rétroaction.

Quelles sont ces règles ? Tout d'abord, l'ordre dans lequel se présentent les données du message : par exemple nom, prénoms, date de naissance, département, commune de naissance. Plus généralement, on vérifie que la structure du message est conforme à une structure attendue. On effectue également des contrôles de type (numérique, alphanumérique, date, chaîne de caractères) et de domaine d'appartenance. Par exemple, une date doit respecter un certain format (comme *JJMMAAAA*), et des règles spécifiques (mois ≤ 12, jour ≤ 31, etc.). Un code de département, ou un code d'activité doivent appartenir à une liste précise et prédéfinie.

En amont, la normalisation est également sémantique : il faut avoir défini le sens de chaque donnée, de chaque valeur dans une liste d'items ou une nomenclature, pour maîtriser l'information transmise. Enfin, standardiser les échanges, c'est formaliser la cinématique des flux, un processus complet dans lequel sont décrites les modalités d'envoi entre émetteur et récepteur ou les notifications transmises, et ce selon le type d'événement.

Normer les échanges est indispensable dans le cas de grands répertoires largement utilisés : norme EDI-CFE pour Sirene, norme A pour le SNGI[33], norme R pour le RGCU[34], ce qui donne lieu à de vastes documentations techniques[35].

Au-delà des référentiels, la standardisation des flux de données est une pratique essentielle dans les systèmes d'information alimentés de façon industrielle : par exemple, la norme GTFS (*General Transit Feed Specification*) est un format standardisé pour communiquer, entre autres, des horaires de transports en commun et les informations géographiques associées ; la norme Odette fait référence dans l'automobile ; le monde de la protection sociale fait l'objet d'une forte rationalisation des échanges *(Gratieux et Le Gall, 2016)*, et utilise en particulier la norme NEODeS pour la déclaration sociale nominative *(Humbert-Bottin, 2018)*.

▶ Définir les services rendus par le répertoire

Le répertoire, ainsi nourri et même rassasié de flux d'information réguliers et standardisés, ne demande désormais qu'à être utilisé. Point d'articulation entre systèmes d'information, il se comporte comme un fournisseur de services envers des utilisateurs variés. Dans le cas du référentiel de métrologie (seconde, kilogramme, mètre), des « services » avaient en quelque sorte été développés en dehors du référentiel : respectivement, le chronomètre, la balance, le mètre ruban. C'est très différent pour un répertoire, évoluant dans un système d'information : les services lui sont intégrés, en sont une composante à part entière.

> ***Il faut isoler en premier lieu les services spécifiques d'un répertoire-socle : l'immatriculation et l'identification.***

Quels sont ces services ? Il faut isoler en premier lieu les services spécifiques d'un répertoire-socle : l'immatriculation et l'identification. Immatriculer un objet, cela consiste à le créer dans le répertoire[36],

33 Cf. l'article sur le SNGI dans ce même numéro.
34 *(Sureau et Merlen, 2021)*, déjà cité.
35 Sur la difficulté de l'écriture des normes en général, et ce qu'elles impliquent en termes de gouvernance, on consultera avec profit *(Mallard, 2000)*.
36 Dans le cas du service d'immatriculation, l'utilisateur est en même temps fournisseur ; la frontière amont-aval n'est pas si simple…

à lui donner naissance dans le système d'information. Pour cela, à partir des traits d'identité de l'objet (nom-prénoms-date et lieu de naissance pour un individu, raison sociale – adresse pour un établissement), on vérifie automatiquement qu'il n'existe pas déjà, puis on lui attribue un identifiant (NIR, numéro Siret, etc.)[37].

La fonction d'identification est tout à fait différente, car elle ne crée aucun objet. À partir de traits d'identité connus (par exemple, « *Leyla Garcia, née le 2 novembre 1999 à Lille* »), on cherche s'il existe une personne ayant des caractéristiques identiques ou proches (par exemple « *Leila* » et non « *Leyla* »), pour obtenir son identifiant (ici le NIR)[38]. Lorsque l'information est imparfaite, incomplète (par exemple en absence de l'année de naissance) ou qu'il reste des ambiguïtés, l'algorithme d'identification doit proposer plusieurs possibilités, plusieurs échos, et les classer par pertinence.

D'autres fonctions viennent naturellement à l'esprit : consulter les données d'une entité (par exemple, toute l'information disponible dans Finess sur tel établissement de santé), ou bien effectuer une sélection (par exemple, la liste, au 31/12/2021, des établissements de santé de l'Oise, créés après 2010). Le fait de pouvoir sélectionner tout ou partie des objets du répertoire à un instant t, de « prendre une photographie », sont des services particulièrement précieux pour les statisticiens, car cela permet de constituer une base de sondage, et plus généralement d'avoir une population de référence pour toute statistique.

Des services peuvent aussi être déclenchés[39] par un mécanisme d'abonnement : par exemple, une caisse de retraite demande à être informée automatiquement de mises à jour concernant les personnes affiliées à son régime.

Les répertoires ont aussi vocation à être utilisés au sein de processus automatisés, sans intervention humaine, en proposant des API intégrables à des processus de production (par exemple, le RGCU fait appel au moteur d'identification de Sirene, pour déterminer les établissements dans lesquels une personne a travaillé pendant sa carrière).

Au total, un répertoire doit proposer une palette de services standard, et différentes modalités d'appel de ces services. Attention cependant, un répertoire n'est certainement pas là pour offrir des services « sur mesure » à chaque utilisateur : par exemple, fournir toutes les « photographies » possibles (contenu, fréquence, filtre, etc.). Ce n'est pas son rôle. La gestion courante en pâtirait, paralysée par la multiplicité des demandes, avec le risque de perdre de vue l'objectif principal, la vocation de référentiel[40].

Pour finir, soulignons qu'il ne faut pas aborder la notion de service en se limitant aux seuls services numériques : dans les répertoires importants, des équipes de gestionnaires assurent en général un travail de réponse aux sollicitations (téléphone, courriel). Cette activité est essentielle pour la qualité du répertoire, et elle fait bien sûr, partie intégrante du service rendu aux utilisateurs.

37 … en s'assurant qu'on ne réutilise pas un identifiant existant.
38 Cf. l'article sur le SNGI, déjà cité.
39 Une petite subtilité : de façon générale, les services sont déclenchés par une demande. Et cette demande est un flux entrant, un message, auquel on va appliquer des contrôles figurant dans la norme d'échange.
40 Une telle limite, auto-imposée, est l'équivalent de l'unité de sens, appliquée aux services.

▶ Le contrôle de la qualité : des traitements automatisés…

Il en va d'un répertoire comme de tout processus d'approvisionnement : il est toujours plus pratique et plus efficace de contrôler la marchandise à son arrivée, immédiatement. Ainsi, avant l'entrée dans le répertoire, on vérifie la provenance et la conformité à des standards, et cela se fait automatiquement (contrôle d'habilitation, application de la norme d'échange). Ces contrôles sont bloquants : si le moindre d'entre eux n'est pas respecté, il y a rejet. Des contrôles non-bloquants peuvent compléter le dispositif : contrôle de cohérence entre données du flux, ou bien comparaison entre la donnée du flux et celle présente dans le référentiel, pour déterminer si l'évolution est plausible. L'élaboration de ces contrôles non-bloquants et leur mise en œuvre sont un sujet en soi[41] non détaillé ici.

On ne peut se limiter à vérifier le flux d'entrée : il faut aussi analyser le répertoire dans son ensemble, de façon macroscopique. Par exemple, repérer des doublons, pour éviter le phénomène de sur-couverture du répertoire, ou bien repérer des données manquantes, ou plus généralement vérifier que les données du référentiel, « en stock », respectent la structure attendue (en termes de typage et de domaine d'appartenance notamment). C'est ce que proposent les systèmes avancés de contrôle, généralement réunis sous le vocable de *data quality tools (Boydens, Hamiti et Van Eeckhout, 2021)*, avec des techniques dites de *profiling (Olson, 2003)*.

▶ … et un vaste travail de vérification humaine

Il serait illusoire de penser que la démarche qualité d'un répertoire se résume à une approche purement mécaniste et algorithmique, en grande partie automatisée. Qui dit répertoire dit équipe de gestionnaires, chargés de vérifications plus complètes et plus fines qui échappent à la machine, et aussi d'échanges directs avec les entités concernées (téléphone, courriel), avec le terrain en quelque sorte, pour une meilleure compréhension en remontant à la source *(Denis, 2018)*[42]. L'organisation de ce travail, souvent intitulée « administration de référentiel », prévaut dans tous les grands répertoires administratifs.

Entre autres exemples :

- le travail des gestionnaires Sirene fait l'objet d'une riche documentation composée de fiches pour chaque cas de figure, indiquant les vérifications à effectuer ;
- pour le SNGI, l'équipe du Sandia vérifie la conformité des pièces justificatives (passeports étrangers) ;
- les gestionnaires RNIPP échangent régulièrement avec les mairies ;
- dans l'organisation du RGCU, figure une cellule d'administration de référentiel, en complément du travail de vérification de carrière faits par les techniciens retraite.

41 Cf. le cas du RGCU, déjà cité.
42 Jérôme Denis explicite dans le détail toutes les composantes du travail de vérification manuelle des données, activité souvent méconnue mais qui est absolument essentielle pour obtenir des données de qualité.

Administrer le répertoire, choisir ses cibles d'intervention, requiert une vision d'ensemble, quantifiée, sous forme de tableaux de bord, notamment pour fixer des priorités et ainsi orienter le travail des gestionnaires de répertoire. Cela comporte des tâches récurrentes (comme le suivi du traitement des rejets) et des tâches projet (par exemple un projet d'amélioration de la qualité du stock). Cette administration inclut également, de façon plus macroscopique, un suivi de production, qui peut utilement mettre en évidence des anomalies : par exemple, lors d'un suivi mensuel du nombre de mises à jour du répertoire, une chute brutale un mois donné alerte sur un possible dysfonctionnement dans les flux d'entrée.

▶ De l'importance de la gouvernance du répertoire

Pour qu'un système aussi complexe qu'un répertoire fonctionne, il faut le piloter, et en particulier prendre des décisions concernant ses évolutions, dans le respect permanent de ses finalités. Pour cela, si les tableaux de bord sont très utiles, il faut également que des instances de décision soient mises en place, à savoir une véritable gouvernance du référentiel, associant les différentes parties prenantes.

Celle-ci est en fait de deux natures : la gouvernance structurelle (sa définition) et la gouvernance « instancielle » du répertoire (son contenu, son peuplement). Dans la première, on trouve par exemple le modèle de données (même si les modalités de validation sont difficiles), la norme d'échange[43], le suivi des réglementations concernées, et dans la seconde, la qualité des données et les conséquences concrètes de la non-qualité (retours d'usage), ou bien l'accostage du répertoire sur les systèmes d'information métier.

Dans ces instances, il faut parvenir à un consensus sur les évolutions de contenu : nouveaux types d'objets à incorporer (« intègre-t-on les associations dans le répertoire d'établissements ? »), nouvelles données, nouvelles nomenclatures, etc., en gardant en tête l'unité de sens, pour éviter d'ajouter des objets ou des variables non pertinents. Les sujets à aborder portent aussi sur les services offerts par le répertoire : extension de la palette de services et du niveau de service (ergonomie, temps de réponse, etc.).

> ***En pratique, « gouverner un répertoire » revient souvent à « rester raisonnable dans ses ambitions ».***

En pratique, « gouverner un répertoire » revient souvent à « rester raisonnable dans ses ambitions ». Ajouter ne fût-ce qu'une seule variable dans un répertoire représente un coût : introduire une nouvelle variable requiert en effet d'organiser les flux d'alimentation associés, donc de trouver les fournisseurs, de s'assurer de la fraîcheur des informations fournies, de développer les contrôles correspondants, de faire évoluer les services, etc. Cela ne se limite donc pas à une simple modification de schéma de base de données.

43 Par exemple, décider des organismes habilités à transmettre des mises à jour, du mécanisme des contrôles de flux, de la cinématique des flux. La gouvernance d'une norme d'échange est un sujet en soi, qui engendre des réunions empreintes de technicité : c'est le cas pour la norme EDI-CFE (alimentation de Sirene), pour la norme A (SNGI), pour NEODeS (norme de la Déclaration Sociale Nominative).

Un répertoire joue un rôle pivot : c'est un objet-frontière, qui « articule des perspectives d'acteurs appartenant à des mondes sociaux hétérogènes » *(Trompette et Vinck, 2009)*[44]. À ce titre, il est naturellement confronté à de multiples acteurs ayant leurs propres intérêts, leurs propres besoins, qui vont aller à l'encontre de la stabilité et de l'unité de sens. Les instances décisionnelles sont ainsi, naturellement, un lieu de rapport de forces entre d'une part des utilisateurs qui en souhaitent toujours plus, d'autre part les responsables du répertoire cherchant à garantir sa généralité, sa cohérence et sa solidité pour des usages futurs.

Il faut toujours se méfier des impressions superficielles de simplicité

Comme on l'a souligné en introduction, l'entité « répertoire » présente tous les attributs de la simplicité : on imagine une liste d'objets, une liste de courses à faire, etc. et on n'y associe pas *a priori* une quelconque complexité. Dès lors, le fait d'assurer la gestion d'un tel objet peut donner l'illusion de l'évidence. Cet *a priori* est source d'erreur et conduit parfois à de grandes difficultés, voire des échecs : *(Bizingre et alii, 2013)* consacrent ainsi un chapitre au cas du référentiel d'une organisation, décrivant « simplement » la liste de ses unités, et montre pour quelles raisons de tels référentiels sont très difficiles à élaborer. Il apparaît, en particulier, que les unités organisationnelles ne sont pas les mêmes selon les usages, et qu'il s'avère impossible pour un référentiel de satisfaire une trop grande variété de besoins, exprimés ou implicites[45].

> ***Ce que l'on stocke, échange et met à disposition, ce ne sont pas des choses inertes, mais de véritables concentrés de sens, mouvants, diversement interprétables selon les acteurs et ayant des impacts colossaux sur des processus métiers.***

Car dans un monde où la donnée ne cesse de prendre de l'importance, les répertoires s'inscrivent dans des systèmes référentiels plus larges[46], véritables infrastructures de connaissances *(Borgman, 2015)*, incluant des nomenclatures ou d'autres répertoires, une standardisation des échanges, un système documentaire, une comitologie (groupe de travail, comité décisionnel), un sous-bassement juridique, le tout étant associé à une « culture référentielle » de l'organisation[47] dont il ne faut pas sous-estimer l'importance. Pourquoi tout cela ? Ce que l'on stocke, échange et met à disposition, ce ne sont pas des choses inertes, mais de véritables concentrés de sens, mouvants, diversement interprétables selon les acteurs et ayant des impacts colossaux sur des processus métiers.

44 Cf. page 8 : il y est rappelé que l'article à l'origine du concept fécond d'objet-frontière *(Star et Griesemer, 1989)* en distingue quatre types, l'un d'entre eux étant… le répertoire, un autre étant… le format d'échange.
45 Le chapitre 5 porte explicitement sur ce sujet.
46 On peut citer, à titre d'exemple, le système d'information sur le milieu marin (SIMM). L'annexe de l'arrêté du 8 juillet 2019 approuvant le schéma national des données sur le milieu marin (SNDMM) définit le système d'information, la nature des données, explicite la gouvernance, les référentiels, les besoins fonctionnels, etc.
47 On peut aussi citer le Sandre pour les systèmes d'information de l'eau.

C'est une matière soumise à une double dynamique :

- sur le temps court, à haute fréquence, les événements qui incessamment la modifient (créations, cessations, mises à jour) ;
- sur le temps long, l'évolution de la sémantique, qui en change la texture ; ce que les données portent en elles n'est pas figé, absolu, hors contexte *(Boydens, 2000)*[48]. Les répertoires, véritables « *hubs* » de données, portent sur un réel qui n'est ni déterministe, ni facile à appréhender.

Dès lors, rien de plus normal que des réunions techniques ou décisionnelles parfois interminables, avec des consensus difficiles à trouver. Que les projets de répertoire[49] soient longs et coûteux ne doit pas non plus surprendre au vu des sujets à traiter : structuration de la base de données, tour d'horizon des sources d'alimentation, reprise des données passées, transition et bascule vers le nouveau système, élaboration et maintenance de la norme d'échanges, transformation des anciens flux d'entrée, poste de travail du gestionnaire, gamme de services à mettre en œuvre, optimisation du fonctionnement en production… sont le lot commun de ces projets (*(Alviset, 2020)* et *(Demotes-Mainard, 2019))*.

Plus généralement, et même si la dimension informatique y joue un rôle majeur en raison du haut degré d'interopérabilité généralement attendu, un projet de réalisation de répertoire ne doit pas être vu en premier lieu comme une opération technique. Car le résultat est avant toute chose un système hautement sensible caractérisé par des finalités générales certes partagées, mais qui se heurtent dans l'opérationnel à des enjeux individuels qui eux ne le sont pas. C'est donc une exigence permanente de compromis entre les acteurs du répertoire, avec la flexibilité suffisante pour satisfaire des utilisateurs, tout en ne baissant jamais la garde sur la rigueur de conception, les finalités transverses. C'est aussi naturellement un reflet d'enjeux de pouvoir, un « lieu » de confrontation d'intérêts divergents… comme ce fut le cas durant la passionnante aventure qui a conduit à élaborer le système métrique[50].

48 Pour reprendre les termes de l'auteur, on ne peut faire « l'hypothèse du monde clos ».
49 On parle ici de répertoire à statut référentiel, bien entendu.
50 *(Alder, 2005),* par exemple p. 28 ou p. 30 : « L'erreur fondamentale des utopistes est de supposer que tout le monde partage la même utopie. ».

▶ Bibliographie

- AGENCE DU NUMÉRIQUE EN SANTÉ, 2022. *CI-SIS. Cadre d'Interopérabilité des Systèmes d'Information de Santé*. [Consulté le 7 septembre 2022]. Disponible à l'adresse suivante : *https://esante.gouv.fr/produits-services/ci-sis*.
- ALDER, Ken, 2005. *Mesurer le monde : 1792-1799, l'incroyable histoire de l'invention du mètre.* 24 avril 2005. Trad. M. Devillers-Argouarc'h. Flammarion. ISBN 978-2082103282.
- 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 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497083/courstat-4-7.pdf*.
- AMOSSÉ, Thomas, 2020. La nomenclature socioprofessionnelle 2020 : Continuité et innovation, pour des usages renforcés. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee. N° N4, pp. 62-80. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497076/courstat-4-5.pdf*.
- BERNARD, Catherine, 1995. Le répertoire Sirene. In : *Courrier des statistiques*. Décembre 1995. Insee, n° 75-76.
- BIZINGRE, Joël, PAUMIER, Joseph et RIVIÈRE, Pascal, 2013. *Les référentiels du système d'information*. Juillet 2013. Dunod. Collection InfoPro. ISBN 978-2100598748.
- BORGMAN, Christine L., 2015. *Big data, little data, no data : scholarship in the networked world.* The MIT Press, 2015.
- BOWKER, Geoffrey C. et STAR, Susan Leigh, 2000. *Sorting things out. Classification and its consequences*. 25 août 2000. The MIT Press. ISBN 978-0262522953.
- BOYDENS, Isabelle, 2000. *Informatique, normes et temps.* Février 2000. Éditions Bruylant. ISBN 978-2802712688.
- BOYDENS, Isabelle, HAMITI, Gani et VAN EECKHOUT, Rudy, 2021. Un service au cœur de la qualité des bases de données. Présentation d'un prototype d'ATMS. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 100-122. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398691/courstat-6-art-6.pdf*.
- CAMUS, Benjamin, 2022. Le défi de l'élaboration d'une nomenclature statistique des infractions. In : *Courrier des statistiques.* [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 146-161. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/6035948/courstat-7-art-8.pdf*.
- CROS, Françoise et RAISKY, Claude, 2010. Autour des mots de la formation « Référentiel ». In : *Recherche et Formation*. [en ligne]. ENS de Lyon. N°64-2010, pp. 105-116. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://journals.openedition.org/rechercheformation/215#tocto1n7*.
- DÉBARBAT, Suzanne et QUINN, Terry, 2019. Les origines du système métrique en France et la Convention du mètre de 1875, qui a ouvert la voie au Système d'international d'unités et sa révision de 2018. In : *Comptes Rendus Physique*. [en ligne]. Janvier-février 2019. Elsevier. Volume 20, n° 1–2, pp. 6-21. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://doi.org/10.1016/j.crhy.2018.12.002.*
- 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 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168399/courstat-2-7.pdf*.

- DENIS, Jérôme, 2018. *Le travail invisible des données. Éléments pour une sociologie des infrastructures scripturales.* [en ligne]. Août 2018. Presses des Mines, Collection Sciences Sociales. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://doi.org/10.4000/books.pressesmines.3934*.
- DI RUOCCO, Nunzio, SCHEIWILER, Jean-Michel et SOTNYKOVA, Anastasiya, 2012. La qualité des données : concepts de base et techniques d'amélioration. In : *BERTI-ÉQUILLE, Laure, 2012. La qualité et la gouvernance des données au service de la performance des entreprises*. 18 septembre 2012. Hermes Science Publications. PP. 25-54. ISBN 978-2-7462-2510-7.
- DINSIC, 2015. *Référentiel général d'interopérabilité. Standardiser, s'aligner et se focaliser pour échanger efficacement*. [en ligne]. Décembre 2015. Direction Interministérielle du Numérique et du Système d'Information et de Communication de l'État. Version 2.0. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.numerique.gouv.fr/uploads/Referentiel_General_Interoperabilite_V2.pdf*.
- GRATIEUX, Laurent et LE GALL, Olivier, 2016. L'optimisation des échanges de données entre organismes de protection sociale. Rapport IGAS – IGF, février 2016. Disponible à l'adresse : *https://www.igas.gouv.fr/IMG/pdf/2015-090R1.pdf*.
- GUIBERT, Bernard, LAGANIER, Jean et VOLLE, Michel, 1971. Essai sur les nomenclatures industrielles . In : *Économie et statistique.* [en ligne]. Février 1971. Insee. N° 20, pp. 23-36. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://doi.org/10.3406/estat.1971.6122*.
- 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 7 septembre]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.
- LACOMBE, Anne, 1979. Histoire de l'invention et de la mise en application du système métrique. In : *The French Review*. [en ligne]. Décembre 1979. American Association of Teachers of French. Vol. 53, n° 2, pp. 246-254. [Consulté le 7 septembre]. Disponible à l'adresse : *https://www.jstor.org/stable/390566*.
- MALLARD, Alexandre, 2000. L'écriture des normes. In : *Réseaux*. [en ligne]. Volume 18, n°102, 2000. La fabrication des normes. pp. 37-61. [Consulté le 7 septembre]. Disponible à l'adresse : *https://doi.org/10.3406/reso.2000.2257*.
- OLSON, Jack E., 2003. *Data Quality – The Accuracy Dimension.* [en ligne]. Janvier 2003. Morgan Kaufmann. ISBN (1-1-55860-891-5, 978-1-55860-891-7).
- RÉGNIER-PÉCASTAING, Franck, GABASSI, Michel et FINET Jacques, 2008. *MDM – Enjeux et méthodes de la gestion des données.* Novembre 2008. Dunod. Collection InfoPro. ISBN 978-2100519101.
- RIVIÈRE, Pascal, 2005. Indicateurs de qualité en matière de production de données : quelques éléments de réflexion. In : *Courrier des statistiques*. Septembre 2005. Insee. N° 115, pp. 35-40.
- STAR, Susan Leigh et GRIESEMER, James R., 1989. Institutional Ecology, 'Translations', and Boundary Objects: Amateurs and Professionals on Berkeley's Museum of Vertebrate Zoology, 1907-39. In : *Social Studies of Science.* [en ligne]. 1er août 1989. Volume 19, n°3, pp. 387-420. [Consulté le 7 septembre]. Disponible à l'adresse : *https://doi.org/10.1177%2F030631289019003001*.

- SUREAU, Christian et MERLEN, Richard, 2021. Le Répertoire de gestion des carrières unique (RGCU). Un nouveau référentiel ouvrant des perspectives pour l'analyse sociale. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 64-81. [Consulté le 7 septembre]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398687/courstat-6-art-4.pdf*.
- TROMPETTE, Pascale et VINCK, Dominique, 2009. Retour sur la notion d'objet-frontière. In : *Revue d'anthropologie des connaissances.* [en ligne]. 2009/1 Vol. 3, n° 1, pp. 5-27. [Consulté le 7 septembre]. Disponible à l'adresse : *https://doi.org/10.3917/rac.006.0005*.
- WALLGREN, Anders et WALLGREN, Britt, 2016. Frames and Populations in a Register-based National Statistical System. In : *Journal of Mathematics and Statistical Science*. Volume 2016, pp. 208-216. [Consulté le 7 septembre]. Disponible à l'adresse : *http://www.ss-pub.org/wp-content/uploads/2016/04/JMSS15121601.pdf*.
- WIENER, Norbert, 1948. *Cybernetics – Or Control and Communication in the Animal and the Machine.* 1961, 2e édition. The MIT Press, Cambridge, Massachusetts. ISBN 978-0-262-73009-9.

Le Répertoire national d'identification des personnes physiques (RNIPP) au cœur de la vie administrative française

Lionel Espinasse et Valérie Roux***

Le répertoire national d'identification des personnes physiques comprend l'état civil de 113 millions de personnes nées ou ayant vécu en France. Un numéro d'identification, le NIR, plus connu comme le « numéro de sécurité sociale » est attribué à chacune d'elles. L'Insee gère le RNIPP à partir d'informations issues des actes d'état civil, transmises soit par les communes, pour les personnes nées en France, soit par la Caisse nationale d'assurance vieillesse, pour les personnes nées à l'étranger. Jumelé au référentiel d'identités de la Cnav, le SNGI, le RNIPP constitue une pièce maîtresse du système social français.

L'accès aux informations personnelles du RNIPP est encadré par des textes réglementaires et reste très limité. Utilisé principalement pour vérifier la conformité des identités ou pour vérifier le statut vital des personnes, le répertoire génère chaque mois plus de 42 millions de connexions, notamment via le service FranceConnect.

Les informations du RNIPP se limitent aux données d'état civil. Mais l'Insee collecte simultanément d'autres informations pour des finalités statistiques, comme la profession, l'adresse ou la situation conjugale. Diffusées uniquement sous une forme agrégée, elles alimentent les études démographiques sur la société française.

The National Register for the Identification of Individuals includes the civil status of 113 million people who were born or have lived in France. An identification number is given at each of therm, the NIR, better known as the "social security number". INSEE manages the RNIPP using information from civil status records, transmitted either by the municipalities, for people born in France, or by the main French Pension Scheme (CNAV), for people born abroad. Together with the SNGI, the CNAV's identity repository, the RNIPP is a key component of the French social security system.

Access to personal information in the RNIPP is governed by regulations and remains very limited. Mainly used to check the conformity of identities or to verify the vital status of individuals, the register generates more than 42 million connections each month, notably via the FranceConnect service.

The information in the RNIPP is limited to civil status data. But INSEE also collects other information for statistical purposes, such as occupation, address or marital status. Distributed only in aggregate form, they are used for demographic studies of French society.

* Adjoint à la cheffe du département de la Démographie, DSDS, Insee, *lionel.espinasse@insee.fr*

** À la date de la rédaction, cheffe du département de la Démographie, DSDS, Insee, *valerie.roux@insee.fr*

Durant de nombreux siècles, nos sociétés ont évolué au sein de communautés villageoises peu nombreuses : chacun se connaissait et s'identifiait facilement. Avec l'apparition des villes et les mobilités croissantes des personnes, il est devenu nécessaire de pouvoir identifier formellement et sans erreur chaque personne. Les noms de famille sont ainsi apparus au XV^{e} siècle ; ils ont été définitivement fixés au XVII^{e} siècle et inscrits sur tous les actes, d'abord religieux puis d'état civil *(Desabie et Hayoun, 1987)*. L'identité d'une personne reste par définition identique tout au long de sa vie, d'où l'idée de la lier à la naissance, avec un nom patronymique transmis par la famille et une place capitale donnée à la date et au lieu de naissance. La nécessité d'un système garantissant l'état civil de chaque personne s'est rapidement imposée dans tous les pays pour éviter les problèmes d'usurpation d'identité, par exemple pour qu'une personne ne puisse contracter deux mariages en même temps !

Avec la modernisation du monde administratif, sont apparus le besoin de compléter l'identité d'une personne par un identifiant et la nécessité de stocker ces informations de manière organisée.

Avec la modernisation du monde administratif, sont apparus le besoin de compléter l'identité d'une personne par un identifiant et la nécessité de stocker ces informations de manière organisée.

En 1941, René Carmille, alors chef du service de la Démographie puis chef du service national des Statistiques, le prédécesseur de l'Insee, initiait un premier répertoire avec un numéro unique pour chaque personne (alors appelé numéro national d'identité). Ce répertoire a été établi à l'aide de moyens mécanographiques à partir des registres d'état-civil détenus par les greffes[1].La tenue du répertoire démographique a ensuite été inscrite dans le décret fondateur de l'Insee dès sa création en 1946[2].

Durant de nombreuses années, ce répertoire a été tenu de manière décentralisée par les directions régionales de l'Insee : le travail se faisait à la main, dans de gros cahiers. L'informatisation du répertoire et sa centralisation ont eu lieu à partir de 1973 (*Lang, 2018*), en réponse à une demande de plus en plus forte des administrations liée à un contexte d'informatisation de nombreux fichiers. Ainsi, dès sa création, les administrations de sécurité sociale se sont appuyées sur le numéro Insee pour leur organisation. Mais le répertoire est aussi très utile pour les administrés puisqu'il évite les erreurs, d'homonymie par exemple, qui pourraient leur être très préjudiciables (dans le calcul de leurs droits par exemple au moment de la liquidation de la retraite).

Le répertoire national d'identification des personnes physiques (RNIPP), est depuis un décret de 1982[3] le nom officiel de ce répertoire qui sert de référence pour les identités et l'état vital des personnes en France.

C'est un répertoire vivant, évoluant chaque jour par l'enregistrement de nouveaux évènements, copie conforme de ceux enregistrés par les services d'état civil des communes

1 Cette naissance dans le contexte particulier de la Seconde Guerre mondiale a fait l'objet de travaux historiques : (*Azema, Levy-Bruhl et Touchelay, 1998).*
2 Voir les références juridiques en fin d'article.
3 Voir les références juridiques en fin d'article.

pour les personnes nées en France (naissances, décès, changements de noms ou de prénoms, parfois de sexe, etc.) et par les services de la Caisse nationale d'assurance vieillesse (Cnav) pour les personnes nées à l'étranger. Son haut niveau de qualité n'est possible que grâce à la mobilisation d'équipes en direction régionale qui veillent à l'exhaustivité, à la fiabilité et à la disponibilité du répertoire. Quel rôle joue aujourd'hui ce répertoire en France ? Comment est-il alimenté et quels mécanismes permettent d'en assurer l'exactitude ?

Une immatriculation dès la naissance

Depuis sa création, l'Insee est responsable de la gestion du RNIPP. Ce répertoire enregistre, pour toute personne née en France, quelle que soit sa nationalité, tous les événements affectant son état civil dès lors que les informations figurant sur les actes officiels ont été transmises à l'Insee ; il ne contient aucune autre information relative aux personnes (notamment ni leur nationalité ni leur lieu de résidence). Au moment de la première inscription, **l'immatriculation,** il attribue de façon univoque un identifiant pérenne à un individu : c'est le numéro d'inscription au répertoire, le NIR, plus communément appelé « numéro de sécurité sociale » ***(encadré 1)***.

Le premier service rendu par le RNIPP est donc celui de l'immatriculation. Ensuite, plusieurs évènements de la vie des personnes seront déclarés par les communes à l'Insee et enregistrés dans le répertoire. En premier lieu, le décès lorsqu'il survient, mais également les changements d'état civil (changement de nom, prénom ou de sexe), ceci afin que le répertoire reste en permanence la copie exacte de l'identité formelle de la personne. Cette identité est celle enregistrée dans les registres d'état civil des 35 000 communes de France et le RNIPP doit en être le miroir exact. Il constitue la référence pour le lien entre les traits d'identité (nom, prénoms, sexe, date et lieu de naissance) et l'identifiant NIR. Il permet ainsi de proposer un service d'identification consistant à confirmer ou non que des traits d'identité existent dans un registre d'état civil, donner une information sur le statut de la personne, vivante ou décédée, et transmettre son NIR.

> ***Le répertoire reste en permanence la copie exacte de l'identité formelle de la personne.***

En 2021, le RNIPP recense plus de 113 millions d'individus ***(encadré 2)*** ; il enregistre environ 2 millions d'événements par an.

113 millions de personnes mais moins d'une dizaine de variables

Les variables du RNIPP, définies par le décret n° 82-103 du 22 janvier 1982, sont moins d'une dizaine. En cela, le RNIPP respecte la caractéristique de stabilité d'un répertoire ayant statut de référentiel[4] : des variables structurantes, stables et fiables, donc peu nombreuses, car il serait sinon coûteux d'en assurer la qualité et la fraîcheur. Seules sont présentes les variables absolument nécessaires à sa fonction de référence pour les identités des personnes.

4 Voir l'article de Pascal Rivière sur les répertoires dans ce même numéro.

Les informations enregistrées sont ainsi les éléments d'identité de la personne tels qu'inscrits dans les actes d'état civil détenus par la mairie : nom de naissance et prénoms, sexe, date, lieu de naissance et numéro de l'acte de naissance et, le cas échéant, date et lieu de décès, numéro de l'acte de décès. S'ajoutent parfois la filiation et le nom marital lorsqu'ils sont nécessaires à l'identification (dans les cas d'homonymies notamment) ainsi que les traits d'identité précédents (en cas de changement de nom, prénom, sexe, etc.). Figure également le numéro unique généré par le répertoire au moment de l'inscription, le NIR. Ce numéro restera attaché à la personne et servira de référence tout au long de sa vie mais également après son décès. En effet, le RNIPP ne supprime aucun enregistrement, même après le décès de la personne concernée. Les cas de changement de NIR existent (par exemple en cas de changement de sexe), mais restent très marginaux.

> ***Le RNIPP ne supprime aucun enregistrement, même après le décès de la personne concernée.***

▶ Encadré 1. Le NIR, un identifiant unique des individus

Le NIR est composé de 13 chiffres et est unique pour chaque personne :

- Sexe (1 chiffre) : 1 pour les hommes, 2 pour les femmes ;
- Année de naissance (2 derniers chiffres de l'année) ;
- Mois de naissance (2 chiffres) ;
- Lieu de naissance (5 chiffres) ;
- Numéro d'ordre dans la commune (3 chiffres).

S'ajoute une clé de vérification sur 2 chiffres.

Par exemple Adrien né le 5 décembre 2021 à Coulommiers et enregistré comme 232e naissance du mois pour cette commune dans le répertoire aura comme numéro NIR : 1 21 12 77 131 232 - 50 (l'ordre d'enregistrement par la commune pouvant être différent de l'ordre de survenue de la naissance ; par ailleurs l'ordre dans le mois peut comprendre des personnes nées à 100 ans d'écart, l'année figurant uniquement sur 2 caractères).

Dans la quasi-totalité des cas, une personne se voit attribuer un NIR à la naissance et elle n'en changera jamais, mais il existe quelques exceptions à cette règle :

- lorsqu'une personne change de sexe, le numéro NIR est mis à jour avec une modification uniquement du premier numéro pour que le chiffre indiqué corresponde bien au sexe actuel de la personne ;
- un second cas concerne les personnes nées avant 1962 en Algérie alors que ce territoire était français. Ces personnes peuvent faire le choix de modifier leur NIR avec un numéro de lieu de naissance correspondant aux anciens codes département de l'Algérie ;
- enfin, il peut arriver qu'une erreur sur un NIR conduise à le corriger ultérieurement.

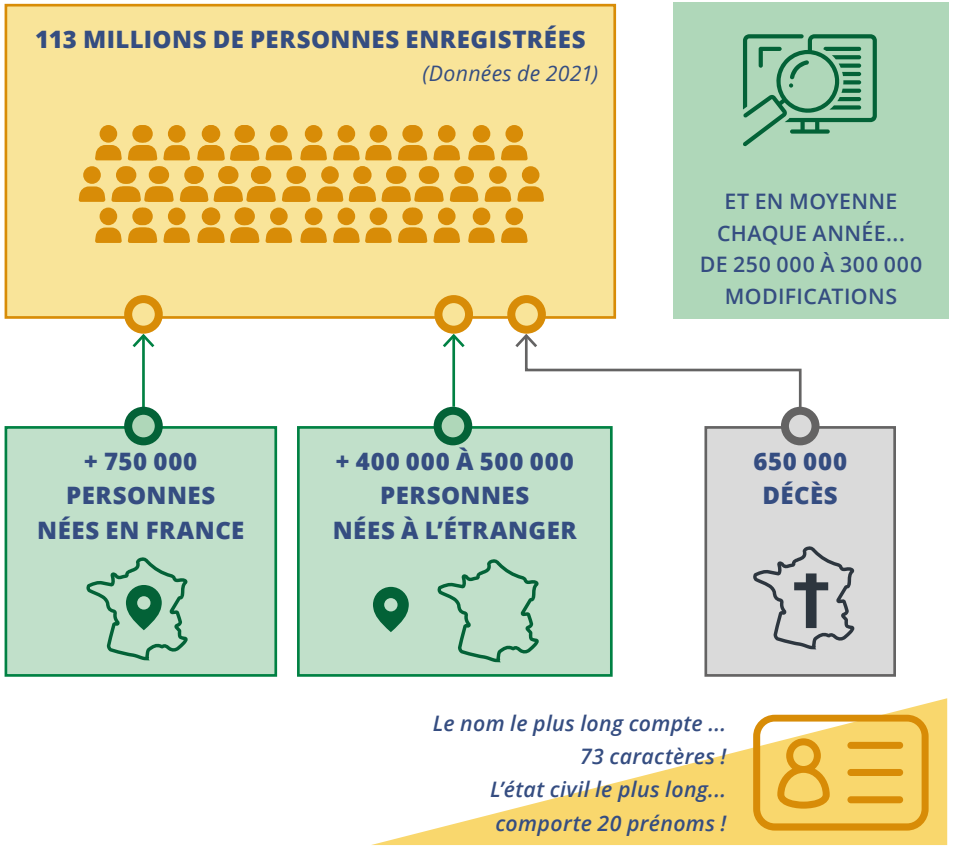
Plus généralement, la gestion du RNIPP doit tenir compte des évolutions de la géographie et des codes associés. C'est le cas en France avec les fusions de communes, les changements de département d'une commune ou la création de nouveaux codes (par exemple, Paris considérée auparavant avec le code unique 75056 et désormais avec 20 codes différents pour chaque arrondissement). C'est aussi le cas à l'étranger avec les disparitions et créations de pays, par exemple en ex-Yougoslavie.

Un système vivant qui change toutes les minutes

Un répertoire n'est pas figé. C'est un système vivant qui évolue chaque jour parce que le monde réel qu'il représente évolue. Mais les modifications apportées au RNIPP doivent toujours être fondées sur des actes officiels. Pour les personnes nées en France, la principale source d'alimentation du RNIPP vient ainsi des actes d'état civil tenus par les communes. Pour chaque évènement, les officiers d'état civil établissent des actes (par exemple des actes de naissance ou des actes de décès) et transmettent ensuite certaines de ces informations à l'Insee pour alimenter le RNIPP. Les officiers d'état civil doivent respecter des délais de transmission : 1 jour ouvré après l'établissement de l'acte pour les naissances, une semaine pour les décès et un mois pour les autres actes.

En cas d'évolution, par exemple un changement de nom ou de sexe, le RNIPP ne peut être modifié qu'à la suite de l'établissement d'un nouvel acte d'état civil qui prend en compte la modification. Ces nouveaux actes prennent en général la forme de « mentions en marge » portées sur les actes d'état civil originaux. Par exemple, si une personne change de sexe, l'indication du nouveau sexe sera portée en marge de son acte de naissance original par l'officier d'état civil de la commune de naissance. Et cette information sera transmise à l'Insee pour l'actualisation du RNIPP.

Encadré 2. Le RNIPP en chiffres



Pour établir les actes, les officiers d'état civil suivent des instructions du ministère de la Justice, rassemblées dans l'instruction générale relative à l'état civil (IGREC), sous l'autorité du Procureur de la République. L'Insee donne, pour sa part, des instructions sur les modalités de transmission des données depuis les communes, mais pas sur les informations elles-mêmes. En particulier, l'Insee n'intervient pas sur l'acceptabilité d'un prénom.

> ***Tout évènement non enregistré dans un acte d'état civil ou non transmis officiellement à l'Insee est ignoré du répertoire.***

Ce principe de conformité aux actes d'état civil présente l'avantage de fiabiliser l'information. Toutefois, cette exigence de sécurité entraîne paradoxalement une limite : tout évènement non enregistré dans un acte d'état civil ou non transmis officiellement à l'Insee est ignoré du répertoire. De ce fait, le décès d'une personne à l'étranger sans transmission de l'information vers un organisme en France ne peut être pris en compte : ceci explique que le RNIPP puisse comprendre des personnes supposées vivantes alors qu'elles sont en fait décédées. On trouve ainsi des supercentenaires dans le répertoire ! Mais pour autant, ces personnes ne reçoivent pas forcément des prestations du système social français. Un système de contrôle est mis en place par les organismes sociaux pour suspendre les versements lorsqu'il n'y a pas eu de preuve de vie récente de la personne[5].

Une double finalité, pour le RNIPP et pour les besoins statistiques

Les transmissions d'information des officiers d'état civil ne se limitent cependant pas aux seules données d'état civil nécessaires à la mise à jour du RNIPP. L'Insee profite de l'existence de ces transferts d'information pour demander d'autres variables qui, cette fois-ci, ont le statut de données statistiques et non plus de données administratives.

La collecte s'opère dans le cadre d'une enquête statistique inscrite au programme du Conseil national pour l'information statistique (Cnis) et reconnue d'intérêt général, avec caractère obligatoire. Les données collectées permettent d'enrichir la connaissance socio-démographique autour des évènements importants de la vie que sont la naissance, le décès, le mariage, le Pacs (et prochainement le divorce). La base de données associée au RNIPP mais aussi à la collecte statistique et à l'échantillon démographique permanent (EDP) s'appelle la **Base des Répertoires de Personnes Physiques (BRPP)**.

Parmi ces données, on trouve par exemple la commune de domicile de la mère, ce qui permet de comptabiliser les naissances en fonction du lieu de résidence de la mère – naissances dites « domiciliées » , et non pas seulement de l'endroit où elle a accouché – naissances dites « enregistrées ». De manière plus classique, on collecte également des informations sur la situation professionnelle, la nationalité, le statut conjugal, etc. Pour certains évènements, des informations plus spécifiques sont demandées, comme le nombre d'enfants précédents de la mère à chaque naissance, ou le nombre d'enfants en commun lors d'un mariage ou de la conclusion d'un pacte civil de solidarité (Pacs).

5 Voir à ce sujet l'article de Joseph Préveraud de Vaumas sur le SNGI, dans ce même numéro.

L'ensemble de ces informations, celles relatives à l'identité et celles d'intérêt statistique, sont collectées par le biais de formulaires appelés **bulletins d'état civil** ***(encadré 3)***, au nombre de 10. La plupart servent aux deux usages : mettre à jour le RNIPP et collecter des variables statistiques, mais certains bulletins ne servent que les usages statistiques. C'est le cas par exemple des bulletins de conclusion et dissolution de Pacs.

▶ Des équipes mobilisées pour assurer la qualité du répertoire

> ***Faire vivre le RNIPP, c'est aussi entretenir des échanges réguliers avec les communes.***

Les bulletins d'état civil existent sous forme papier, mais les transmissions d'information entre les communes et l'Insee sont largement dématérialisées. En 2021, 99 % des naissances, 94 % des décès et 81 % des mariages étaient transmis de manière dématérialisée. Pour cela, les communes peuvent utiliser, soit une application mise à disposition gratuitement par l'Insee[6], soit un logiciel proposé par un éditeur privé. Ces logiciels du marché proposent des fonctionnalités plus riches que l'application de l'Insee, notamment ils ne se limitent pas à la transmission de données à l'Insee mais incluent toute la chaîne de traitement de l'état civil, y compris la partie relative à l'établissement de l'acte original.

Faire vivre le RNIPP, c'est aussi entretenir des échanges réguliers avec les communes, notamment pour garantir son exhaustivité et sa qualité. Malgré la forte dématérialisation tout ne peut pas être automatisé. La gestion quotidienne du bon fonctionnement de ce dispositif de transmission est ainsi assurée par 70 agents de l'Insee, au sein des « sites BRPP » dans des établissements régionaux, appuyés par un pôle national à Nantes.

En effet, **l'exhaustivité** de l'information contenue dans le répertoire, tout au moins au regard de la conformité aux actes d'état civil, est garantie normalement par l'obligation réglementaire faite aux officiers d'état civil de signaler tout événement à l'Insee. Afin de vérifier qu'il ne manque aucun bulletin, l'Insee organise néanmoins un suivi d'exhaustivité, avec des alertes lorsque des transmissions de communes sont anormalement faibles, lorsque des « trous » dans les numéros d'actes ou d'ordre sont détectés, ou si des présomptions de décès non déclarés au RNIPP sont détectées (par exemple lorsque l'information sur un décès est transmise par les organismes sociaux). À côté de ces opérations courantes, sont organisées des opérations qualité plus ponctuelles, en particulier une enquête annuelle sur l'exhaustivité des remontées de mariages, ou d'autres plus ciblées sur les centenaires ou les individus enregistrés par erreur en double.

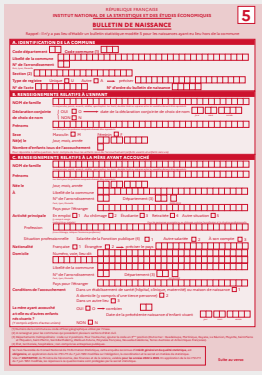
Les sites BRPP sont aussi les interlocuteurs des communes et leur apportent conseil et assistance, notamment sur tous les sujets liés à la dématérialisation des transmissions (formalisation conventionnelle et assistance technique). Ils gèrent les flux de bulletins papier vers et en retour du prestataire de saisie ; effectuent des reprises ou des corrections de données en cas d'erreur, ainsi qu'un peu de saisie résiduelle. Grâce à ces travaux, 96 % des naissances sont enregistrées en moins de 8 jours et 97 % des décès en moins de 21 jours.

6 Application Aireppnet (Alimentation informatisée du répertoire des personnes physiques par internet) dont la documentation est sur le site de l'Insee, voir *(Insee, 2021).*

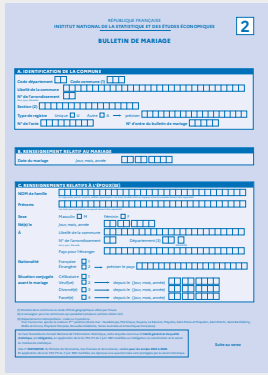
▶ Encadré 3. Dix bulletins d'état civil

Les bulletins d'état civil sont accessibles sur le site de l'Insee.fr.
(https://www.insee.fr/fr/information/1303477)

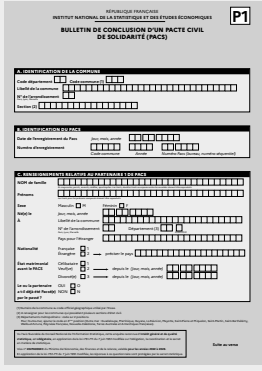
NAISSANCE


RÉPUBLIQUE FRANÇAISE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
5
BULLETIN DE NAISSANCE

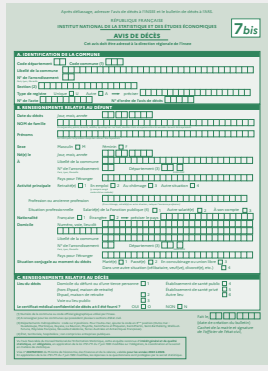
MARIAGE


RÉPUBLIQUE FRANÇAISE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
2
BULLETIN DE MARIAGE

CONCLUSION D'UN PACS


RÉPUBLIQUE FRANÇAISE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
P1
BULLETIN DE CONCLUSION D'UN PACTE CIVIL DE SOLIDARITÉ (PACS)

DÉCÈS


RÉPUBLIQUE FRANÇAISE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
7bis
AVIS DE DÉCÈS

MAIS AUSSI :

- Transcription relative à un jugement déclaratif de naissance
- Bulletin d'enfant sans vie
- Transcription relative à un jugement d'adoption plénière
- Transcription relative à un jugement déclaratif de décès ou d'absence
- Bulletin de mention en marge
- Bulletin de dissolution d'un pacte civil de solidarité (Pacs)

Les sites traitent également des demandes d'identification de personnes pour l'inscription au répertoire électoral unique (REU) : il s'agit alors de vérifier l'existence de ces personnes et de récupérer leur identité officielle, lorsque l'identification ne peut pas être automatisée et demande une analyse approfondie. Les gestionnaires analysent aussi les cas de présomptions de décès, lorsqu'une information est transmise par les organismes de sécurité sociale.

Les personnes nées à l'étranger peuvent-elles être immatriculées au RNIPP ?

Les personnes nées à l'étranger doivent également être inscrites au RNIPP lorsqu'elles viennent vivre en France même pour quelques mois, pour travailler, faire leurs études ou se faire soigner. L'employeur a besoin d'un NIR pour payer les cotisations sociales de son salarié et le NIR est aussi nécessaire à ce dernier pour bénéficier des droits sociaux, notamment ceux de l'assurance maladie.

> ***Depuis 1987, l'Insee délègue à la Cnav la responsabilité de l'immatriculation des personnes nées à l'étranger.***

Pour être inscrites au RNIPP et obtenir ainsi un « numéro de sécurité sociale », les personnes nées à l'étranger doivent en faire la demande auprès d'un organisme de sécurité sociale (caisses primaires d'assurance maladie, caisses d'allocations familiales, mutualité sociale agricole, caisses des indépendants, etc.)[7]. C'est ensuite la Caisse nationale d'assurance vieillesse (Cnav) qui traitera leur demande. En effet, depuis 1987, l'Insee délègue à la Cnav la responsabilité de l'immatriculation des personnes nées à l'étranger. C'est également la Cnav qui continuera par la suite à enregistrer toutes les modifications concernant ces personnes.

La Cnav gère ainsi le système national de gestion des identifiants (SNGI), répertoire miroir du RNIPP pour la partie état civil, qui sert de système de référence en matière d'identité pour tous les organismes de protection sociale pour toutes les personnes nées en France ou à l'étranger[8]. C'est au sein de ce système que se font en premier lieu les immatriculations des personnes nées à l'étranger et l'attribution d'un NIR. Les deux systèmes – RNIPP et SNGI –, sont synchronisés chaque nuit avec les nouvelles immatriculations ou modifications opérées dans la journée ***(figure 1)***. Cela permet au SNGI de récupérer toutes les nouvelles inscriptions ou modifications concernant les personnes nées en France et au RNIPP de faire de même pour les personnes nées à l'étranger. Ainsi les deux répertoires sont bien exhaustifs et identiques en ce qui concerne les traits d'identité, l'état vital et le NIR. L'existence du SNGI permet ainsi aux organismes de protection sociale de disposer de leur propre système pour la vérification des identités.

Des travaux qualité, opérés conjointement par l'Insee et la Cnav, ont lieu régulièrement pour vérifier la cohérence entre les deux répertoires. En effet, malgré ce système quotidien d'échange informatique, il peut y avoir divers problèmes aboutissant à des divergences.

7 Depuis peu, les employeurs peuvent aussi faire directement la demande pour leurs salariés étrangers sur un site dédié.
8 Voir l'article de Joseph Préveraud de Vaumas sur le SNGI, dans ce même numéro.

Trois grandes catégories de divergences se dégagent :

- des NIR présents uniquement dans une seule des deux bases ;
- pour un NIR commun, des divergences dans les traits d'identité ;
- et des informations différentes sur les décès.

Les travaux menés en 2021 sur un échantillon de 16 millions de personnes permettent d'estimer à 0,05 % la part des situations nécessitant une correction pour rétablir la cohérence entre les deux répertoires.

Malgré toutes ces actions de vérification de la qualité du répertoire, tant à l'Insee qu'à la Cnav, il peut arriver que des personnes soient déclarées décédées à tort. Ces cas sont très rares, mais peuvent arriver notamment en cas d'homonymie, de données d'identité très proches au sein d'une famille ou d'inversion des identités entre la personne qui déclare le décès et la personne décédée… Des correctifs sont alors appliqués le plus rapidement possible.

▶ Le premier usage de ce répertoire : certifier l'état civil et l'état vital des personnes

Le décret de 1982 ne donne pas explicitement de finalités au RNIPP mais celles-ci sont explicitées par ailleurs par la Commission nationale de l'informatique et des libertés qui le définit comme « *un instrument de vérification de l'état civil des personnes nées en France* » *(Cnil, 2009)*. Il est d'ailleurs intéressant de noter que cette commission a été créée à l'origine pour contrôler les usages d'un tel fichier ***(encadré 4)***. Le RNIPP est un répertoire qui a donc un statut de référentiel. Les données qui y sont présentes sont essentielles pour identifier sans ambiguïté une personne et le fait qu'elle soit vivante ou décédée.

En premier lieu, il permet la **certification de l'état civil** d'une personne, c'est-à-dire qu'il permet à un organisme autorisé de vérifier si une personne qui s'est déclarée sous une identité donnée existe et est en vie. Il sert ainsi pour les organismes de sécurité sociale *via* le SNGI, mais aussi pour l'administration fiscale, la Banque de France, le répertoire Sirene et pour d'autres acteurs de plus en plus nombreux.

Mais les données contenues dans ce répertoire sont très sensibles et donc protégées. Pour avoir accès aux informations qui s'y trouvent, notamment pour pouvoir vérifier un état civil en interrogeant le répertoire, il faut y être autorisé par un acte réglementaire. Un décret spécifique[9] recense depuis 2019 toutes les catégories d'acteurs et les finalités de traitement pour lesquelles l'utilisation du NIR ou l'accès au RNIPP est autorisé, dans chaque secteur d'activité (protection sociale, logement, travail, justice, financier, fiscal et douanier, statistique publique et recensement, éducation), par exemple, « *pour faire certifier par l'Insee les états civils des personnes physiques titulaires de comptes bancaires : les services de la direction générale des finances publiques* ».

9 Décret n° 2019-341 du 19 avril 2019, voir références juridiques en fin d'article.

Toute utilisation du NIR ou des données du RNIPP qui n'entre pas dans les cas d'usages visés par ce décret est interdite ou doit être prévue par un autre texte législatif ou réglementaire. L'Insee vérifie systématiquement que l'organisme demandeur et le traitement engagé sont bien référencés dans le décret. Si ce n'est pas le cas, l'Insee refuse l'accès aux données et explique au demandeur la nécessité d'obtenir une autorisation, soit par la loi, soit par une modification du décret ou d'un autre texte réglementaire.

Le RNIPP : un répertoire pivot au service d'autres répertoires

L'importance que le RNIPP a pris dans l'organisation administrative française est due à la qualité des informations d'état civil qui y sont enregistrées, à leur actualité mais aussi à l'interopérabilité de ce répertoire, au sens informatique du terme, avec de nombreux autres systèmes d'information. C'est pourquoi le RNIPP joue un rôle de pivot de plus en plus net dans le système administratif français, et singulièrement pour d'autres répertoires.

> ***Lorsqu'une personne s'inscrit comme électeur, son état civil est au préalable vérifié auprès du RNIPP.***

Prenons le cas du répertoire électoral unique (REU), en place depuis 2019 pour gérer les listes électorales (inscription, radiation, par commune, ainsi que les procurations) *(Demotes-Mainard, 2019)*. Lors de l'organisation de scrutins électoraux, les listes électorales sont extraites de ce répertoire. Lorsqu'une personne s'inscrit comme électeur, son état civil est au préalable vérifié auprès du RNIPP.

Encadré 4. La CNIL est née de l'existence d'un répertoire des identités

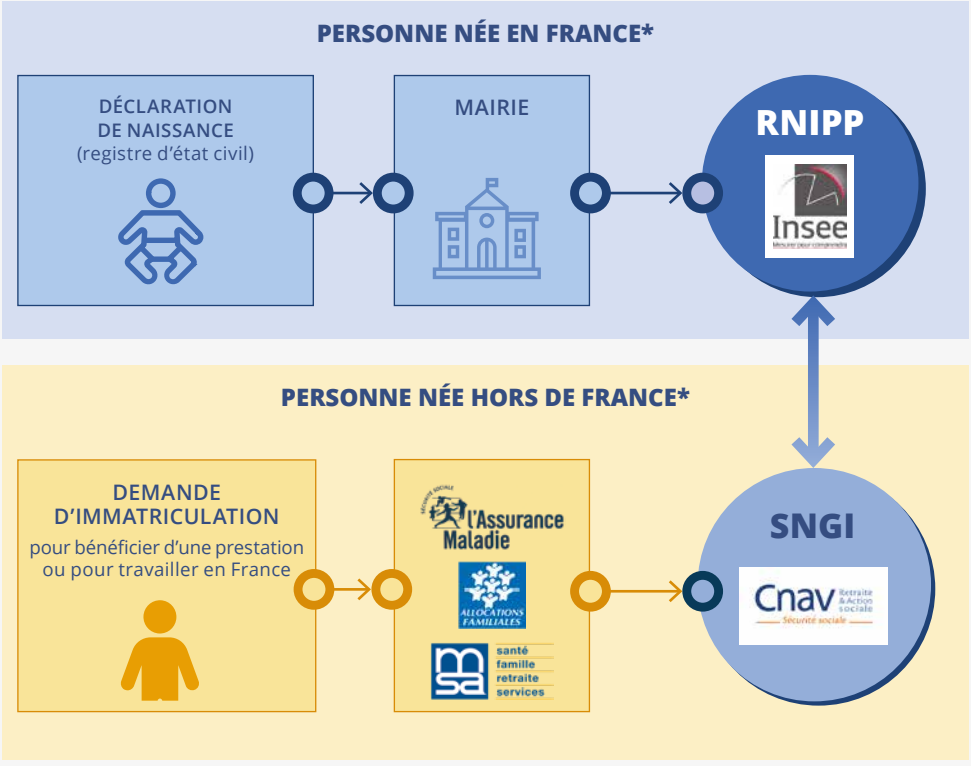
Au moment de l'informatisation du répertoire d'identification des personnes physiques, une forte polémique est née. En 1970, le directeur de la statistique générale de l'époque au sein de l'Insee, Jacques Desabie, décrit les nombreux avantages de l'informatisation du fichier et de l'usage d'un identifiant numérique pouvant devenir l'identifiant unique du citoyen dans ses relations avec l'administration (*Desabie, 1970*).

Mais le projet d'interconnexion de nombreux fichiers de l'administration sur la base de cet identifiant suscite de vives réactions et le choix fait en 1973 par le ministère de l'Intérieur de nommer ce projet « Système Automatisé pour les Fichiers Administratifs et le Répertoire d'Identification » est alors malheureux, car son acronyme Safari va être à l'origine d'une violente polémique, synthétisée par un article du journal Le Monde « Safari ou la chasse aux Français ». Le projet sera abandonné mais de cette polémique naîtra la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (*Cnil et Ina, 2022*). Cette loi prévoit en son article 18 que *« l'utilisation du répertoire national d'identification des personnes physiques en vue d'effectuer des traitements nominatifs est autorisée par décret en Conseil d'État pris après avis de la commission »*. Depuis sa création, la CNIL a ainsi toujours veillé à ce que le numéro d'identification au répertoire (connu sous le nom de numéro de sécurité sociale) soit cantonné, notamment au domaine de la sécurité sociale et ne soit pas utilisé par toutes les administrations. C'est ainsi que l'éducation nationale a mis en place pour le suivi des élèves un autre système, le répertoire national des identifiants élèves, étudiants et apprentis, doté d'un identifiant spécifique (l'INE ou identifiant national élève).

Ceci permet d'abord de vérifier que la personne existe et n'est pas décédée ; cela permet aussi de récupérer son état civil complet et officiel. C'est celui-ci qui sert ensuite de référence dans le REU (et qui est inscrit sur la carte d'électeur).

La mise en place du REU en 2019 a aussi permis à de nombreux électeurs de connaître leur état civil officiel tel qu'il existe au RNIPP, car celui-ci est apparu pour la première fois sur leur carte d'électeur. Certains ont alors parfois constaté des erreurs, sur leurs nom ou prénom. Cela pouvait provenir d'erreurs de saisie ou encore de changements non transmis par la commune de naissance à l'Insee.

Figure 1 - Deux circuits d'immatriculation pour deux répertoires complémentaires



** le terme « France » désigne ici la France métropolitaine, les DOM, les collectivités de Saint-Pierre et Miquelon, Saint-Barthélémy, Saint-Martin, la Polynésie française et Wallis-et-Futuna.*

Cette large information des électeurs a ainsi permis une grande opération qualité sur le RNIPP. Sur 85 millions de personnes en vie présentes dans le répertoire, il y a eu 45 200 changements d'état civil effectués par l'Insee pour des personnes nées en France (68 % concernant des prénoms et 30 % des noms) et 58 000 changements pour des personnes nées à l'étranger (donc effectués par la Cnav). Soit au total 103 200 modifications (0,1 % du répertoire). Un très gros travail pour les équipes de l'Insee et de la Cnav, mené en quelques mois.

Le **répertoire Sirene**[10] d'immatriculation des entreprises mobilise également le RNIPP pour vérifier l'identité des entrepreneurs individuels au moment de leur inscription. Cela permet encore une fois d'être sûr de l'identité du créateur d'entreprise et du fait qu'il soit toujours en vie. Si l'entreprise a ultérieurement des problèmes avec ses clients ou fournisseurs, cela permettra à ces derniers de pouvoir engager des démarches judiciaires sans mauvaise surprise.

▶ Un rôle nouveau avec le développement de *FranceConnect*

Initialement cantonné à la sphère sociale, le RNIPP joue un rôle de plus en plus structurant dans le système administratif français. Le nombre de « clients » est en croissance constante avec une accélération ces dernières années où l'on enregistre entre 20 et 30 millions de demandes d'identifications par an (hors *FranceConnect*). Plus récemment, l'arrivée de *FranceConnect* a encore accru les usages de ce répertoire.

Créé en 2018 par la direction interministérielle du numérique et du système d'information et de communication de l'État (Dinum), *FranceConnect* est un téléservice qui permet de sécuriser et de simplifier la connexion à plus de 1 000 services en ligne en 2022. Il offre à chaque usager la possibilité d'utiliser une identité numérique (identifiant et mot de passe) liée à un opérateur pour accéder aux services d'autres opérateurs. Ainsi, par exemple, l'usager peut employer ses identifiants Améli[11] pour se connecter au service des impôts. Plus besoin de retenir une multitude d'identifiants et de mots de passe.

Avant d'autoriser l'accès au service souhaité, le dispositif *FranceConnect* vérifie l'identité de l'usager en la comparant à celle du RNIPP. Et ce n'est que s'il y a concordance que l'accès est rendu possible. La disponibilité de l'application de consultation du RNIPP est ainsi un enjeu majeur pour le bon fonctionnement de cet écosystème. D'autant que le nombre de consultations du RNIPP par *FranceConnect* n'a cessé de croître, passant de 8 millions d'utilisateurs en 2019 à 37 millions en 2022. Les exigences de robustesse, de disponibilité, de tenue à la charge sont donc des préoccupations de premier ordre pour l'Insee, afin d'assurer le niveau de service attendu.

En cas d'erreur repérée par un citoyen sur les données le concernant, il est alors primordial que celle-ci puisse être corrigée rapidement afin que le préjudice soit le plus faible possible. C'est pourquoi l'Insee et la direction de l'Information légale et administrative (DILA) ont mis en place en 2019 une démarche en ligne permettant de demander la correction de son état civil en transmettant une copie de son acte de naissance.

10 *https://www.insee.fr/fr/metadonnees/source/serie/s1020*
11 Nom du site officiel de l'Assurance Maladie en France.

Ce service est accessible uniquement pour les personnes nées en France sur le site officiel de l'administration française. Les personnes nées à l'étranger, en attendant qu'un service analogue soit développé, peuvent s'adresser à leur organisme de protection sociale.

Certaines personnes ne peuvent pas accéder à internet, il leur reste alors possible de demander une correction de leur état civil par courrier, adressé à l'Insee pour les personnes nées en France, ou à leur organisme de sécurité sociale pour les personnes nées à l'étranger.

Un système d'information exigeant pour tenir à jour le répertoire

On le voit, le RNIPP présente deux caractéristiques fortes.

En premier lieu, les conséquences individuelles d'une erreur de gestion sur ce répertoire sont potentiellement graves. Elles peuvent entraîner des difficultés pour bénéficier de droits sociaux, comme recevoir sa pension de retraite, ou pour accéder à des services (numériques ou pas) réclamant une identification, voire pire : être déclaré décédé à tort.

> ***Cette interopérabilité doit être en permanence opérationnelle.***

En second lieu, son système d'information opère en interaction avec de nombreux autres (Cnav, *FranceConnect*, logiciels éditeurs des communes, etc.) et cette interopérabilité doit être en permanence opérationnelle.

Dans ce contexte, le bon fonctionnement de la gestion de ce répertoire impose de fortes exigences[12], de plusieurs natures :

- une **disponibilité** permanente du système d'information ;
- une **capacité d'adaptation** rapide et permanente aux évolutions légales ou réglementaires ;
- le maintien permanent de **l'exhaustivité et de l'actualité** des informations ;
- des **délais de traitement des cas individuels** respectant les règles de droit et tenant compte des conséquences sur la vie quotidienne des personnes ;
- le maintien de **l'interopérabilité** avec les systèmes d'information de nombreuses autres administrations ;
- la garantie de la **protection des données** à caractère personnel ;
- la **lutte contre les cyberattaques**.

Satisfaire ces exigences implique de mener de nombreuses actions, tant sur le contenu du répertoire que sur son système d'information. Une campagne d'envergure d'intégration dans le RNIPP de l'ensemble des Français nés à l'étranger, en lien avec la Cnav et le Service Central de l'État civil du ministère de l'Europe et des Affaires étrangères (et non uniquement de ceux qui se font connaître pour bénéficier de droits sociaux), a ainsi été menée en 2022,

12 Voir l'article de Pascal Rivière sur les répertoires, dans ce même numéro.

avec pour objectif de faciliter leurs démarches s'ils reviennent en France ou leurs démarches en ligne pour accéder à des téléservices français. De même, des actions d'amélioration de l'identification des personnes nées à l'étranger ont été lancées, afin de faciliter leur accès à *FranceConnect.*

Les exigences de maintien de l'interopérabilité interviennent dans un contexte où les interactions avec les systèmes d'information des autres administrations mettent en œuvre diverses techniques. D'abord des web services, avec la Cnav principal partenaire pour l'alimentation du répertoire, mais aussi avec les grands opérateurs de démarches administratives (*FranceConnect*, Sirene et le Guichet unique d'immatriculation des entreprises). Les échanges avec les communes *via* les logiciels d'éditeur privés transitent également par des API. Enfin, la technique du dépôt-retrait de fichiers[13] est très utilisée pour les prestations d'identification, par exemple avec les services fiscaux ou les organismes bancaires ou d'assurance ***(figure 2)***.

L'ensemble des échanges se doit de respecter des normes. Celles-ci sont publiques et diffusées sur le site internet de l'institut *(Insee, 2021)*. Pour les échanges avec les communes, ces normes portent d'abord sur le format des fichiers. Elles s'expriment aussi sous la forme de consignes de saisie des données, notamment sur les signes diacritiques autorisés, les ligatures, la gestion des tirets ou apostrophes ou encore les abréviations des noms de lieux. Ainsi, seul l'alphabet romain et les signes diacritiques connus dans la langue française sont autorisés.

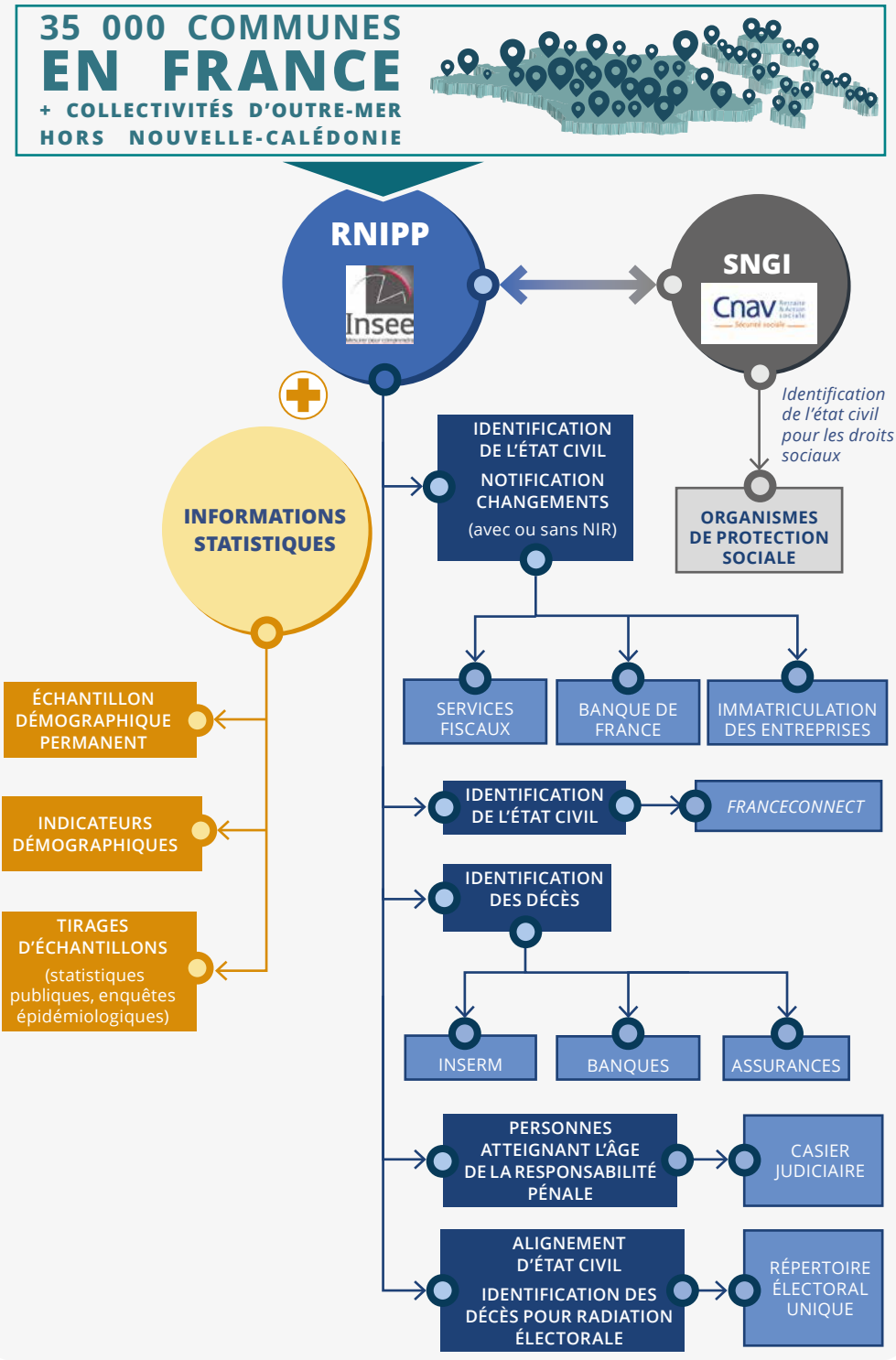
De la même manière, les échanges avec les clients du service de certification d'identité doivent respecter les formats de fichiers prescrits par l'Insee ainsi que les modes de transmission et de sécurisation des données.

> ***En 2021, le RNIPP a été disponible pour* FranceConnect *99,5 % du temps.***

Les conditions de disponibilité du système d'information sont aussi régulièrement à surveiller au regard de la charge de sollicitation. En 2021, le RNIPP a été disponible pour *FranceConnect* 99,5 % du temps. Mais le périmètre de ce service ne cesse de croître, par exemple avec l'adhésion récente des Caisses d'allocations familiales qui génère de fait de nombreux nouveaux utilisateurs. Ces évolutions de la charge de connexion peuvent conduire à des adaptations régulières du système informatique.

13 L'organisme demandeur dépose dans un espace sécurisé un fichier comprenant les traits d'identité qu'il souhaite faire vérifier au RNIPP. L'Insee récupère ce fichier, le traite et dépose la réponse dans un espace sécurisé où le demandeur pourra la récupérer.

Figure 2 - Le RNIPP, l'identification au cœur des deux sphères administrative et statistique



▶ Des données pour calculer le taux de fécondité ou suivre les décès du Covid-19

> ***Le RNIPP ne permet pas de comptabiliser la population vivant en France.***

Lorsqu'une personne est inscrite au RNIPP, elle y reste définitivement. Les personnes immigrées peuvent retourner dans leur pays de naissance, mais elles demeurent définitivement inscrites au RNIPP. Si un jour elles reviennent vivre en France, elles n'auront pas de nouvelles démarches à accomplir et pourront utiliser le NIR qui leur est donné « à vie ». Si elles décèdent hors de France, l'information ne sera vraisemblablement pas connue et elles seront maintenues en vie dans le répertoire. Cela explique pourquoi le RNIPP ne permet pas de comptabiliser la population vivant en France et n'est donc pas un registre de population (*Poulain, Herm, 2013*). En effet, l'adresse ne fait pas partie des variables enregistrées dans le répertoire, ce qui ne permet pas de distinguer le nombre de personnes présentes dans le RNIPP qui ne vivent pas en France. Ainsi, le RNIPP compte 85 millions de personnes présumées en vie alors qu'avec le recensement de la population on estime la population vivant en France à 68 millions au 1er janvier 2022[14].

Le RNIPP permet néanmoins à l'Insee de produire des statistiques démographiques. Chaque année, le bilan démographique de la France et de ses territoires s'appuie sur ces données *(Papon, 2022)*. C'est ainsi qu'on estime des indicateurs démographiques majeurs comme le nombre de naissances et de décès en France, le nombre d'enfants par femme, l'espérance de vie à la naissance, etc.

Le répertoire a également joué un rôle majeur en France pendant l'épidémie de Covid-19. L'Insee et Santé publique France ont collaboré pour mettre en place très rapidement dès le mois de mars 2020 une surveillance de la mortalité. Une rubrique du site internet de l'Insee dédiée au suivi de la mortalité toutes causes confondues a été créée et mise à jour toutes les semaines au début de l'épidémie puis avec des fréquences variables selon l'acuité de la crise *(Insee, 2022)*.

Indépendamment de cet usage conjoncturel lié à la crise de la Covid-19, le RNIPP joue aussi un rôle dans la chaîne de production statistique des causes de décès. Ces causes, identifiées par le médecin signant le certificat de décès, sont traitées et analysées par l'Inserm dans un cadre respectant le secret médical, mais la vérification d'exhaustivité est réalisée avec les données du répertoire.

Par ailleurs, les variables statistiques collectées dans les bulletins d'état civil sur la profession, la nationalité, la situation conjugale, etc. font l'objet de publications spécifiques et sont diffusées sous forme de tableaux détaillés et de fichiers détail sur le site insee.fr[15].

14 Sur le champ géographique du RNIPP : France métropolitaine, DOM, collectivités de Saint-Pierre et Miquelon, Saint-Barthélémy, Saint-Martin, Polynésie française et Wallis et Futuna

15 Fichiers détail sur les naissances, décès et mariages : https://www.insee.fr/fr/statistiques/5419788

Certaines données issues du RNIPP sont également intégrées dans un large panel démographique, l'Échantillon démographique permanent (EDP). Ce panel sociodémographique[16] rassemblant plusieurs sources de données a été mis en place en France pour étudier la fécondité, la mortalité, les parcours familiaux, les migrations géographiques au sein du territoire national, la mobilité sociale, professionnelle et résidentielle, les carrières salariales et les niveaux de vie ainsi que les interactions possibles entre ces différents aspects.

Enfin, le RNIPP est également utilisé pour tirer des échantillons et permettre la mise en place d'enquêtes sur des sujets particuliers. Ce peut être pour le suivi d'une catégorie spécifique de la population, les centenaires par exemple ou pour des besoins d'étude épidémiologique.

▶ Un répertoire qui s'adapte aux évolutions de la société

Le RNIPP et les transmissions d'informations statistiques s'adaptent aux évolutions de la société. Dans le cas du RNIPP, il s'agit essentiellement d'évolutions provoquées par des modifications législatives ; pour le domaine statistique, il s'agit plus largement de mieux rendre compte des réalités sociétales. Ainsi, la loi Bioéthique du 2 août 2021 a institué la « PMA pour toutes » (procréation médicalement assistée pour les couples de femmes) ce qui a conduit à faire évoluer le bulletin de naissance et ses bulletins associés (enfant sans vie et jugement déclaratif de naissance) pour permettre d'enregistrer deux parents de sexe féminin mais également le bulletin de mention en marge pour les reconnaissances par une deuxième mère.

Les évolutions dans le domaine de la conjugalité ont également été importantes ces dernières années. La mise en place du PACS a conduit à la création de deux nouveaux bulletins à vocation statistique (conclusion et dissolution de PACS). De même, la mise en place des divorces par consentement mutuel devant notaire (ou déjudiciarisés) en complément des divorces par procédure judiciaire a conduit à revoir les circuits de recueils d'informations. Si les informations sur les divorces par jugement sont traditionnellement transmises par les institutions judiciaires, un nouveau dispositif a dû être imaginé pour les divorces devant notaires. Ainsi à compter de 2023, l'Insee collectera les mentions de divorce portées en marge des actes de mariages, ceci afin de pouvoir établir des statistiques sur l'évolution des divorces en France car actuellement cette donnée est incomplète.

> ***La Covid-19 a également mis en exergue l'important besoin d'informations statistiques sur les décès au niveau infra-communal.***

Dans un tout autre domaine, celui des études épidémiologiques, la crise sanitaire de la Covid-19 a également mis en exergue l'important besoin d'informations statistiques sur les décès au niveau infra-communal. Jusqu'à présent, la localisation du domicile du défunt se limitait au niveau de la commune, ce qui rendait difficile les analyses par quartier, qui nécessitaient alors d'utiliser en complément d'autres sources d'information. À compter de 2023, l'adresse du domicile du défunt sera intégrée dans le bulletin de décès, ce qui permettra les exploitations au niveau infra-communal.

16 Pour plus de détail sur l'EDP, voir *(Robert-Bobée et Gualbert, 2021).*

> ***Une refonte importante des bulletins sera ainsi opérationnelle en 2023.***

Une refonte importante des bulletins sera ainsi opérationnelle en 2023. Elle a été menée après une large concertation auprès des principaux utilisateurs des données issues des bulletins d'état civil (ministère de la Justice, Inserm, Ined, Drees, Haut conseil à la famille, à l'enfance et à l'âge, Union nationale des associations familiales, etc.) et a obtenu le label d'intérêt général et de qualité statistique. Outre les évolutions déjà citées, on trouvera aussi une actualisation de la collecte des données sur la profession, sur le statut conjugal et sur les lieux d'accouchement ou de décès. Cette refonte sera aussi l'occasion de moderniser le processus d'appariement entre les causes de décès et les déclarations de décès par une utilisation plus efficace du numéro non identifiant généré par l'application informatique de transmission dématérialisée du certificat de décès. Toujours dans le domaine de la santé, le RNIPP contribuera aussi bientôt à l'alimentation du système national de données de santé.

▶ Bibliographie

- AZEMA, Jean-Pierre, LEVY-BRUHL Raymond, TOUCHELAY Béatrice. Mission d'analyse historique sur le système statistique français. *Hall open science*, juillet 1998 [Consulté le 25 novembre 2022]. Disponible à l'adresse : *https://hal.archives-ouvertes.fr/hal-02426370/document*.
- CNIL et INA, 2022. L'histoire de la CNIL en vidéo. 40 ans au service des libertés. In : *site de la Commission nationale de l'informatique et des libertés*. [en ligne]. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://sites.ina.fr/cnil/focus/chapitre/2*.
- CNIL, 2009. RNIPP : Répertoire national d'identification des personnes physiques. In : *site de la Commission nationale de l'informatique et des libertés.* [en ligne]. 19 juin 2009. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.cnil.fr/fr/rnipp-repertoire-national-didentification-des-personnes-physiques-0*.
- DEMOTES-MAINARD, Magali, 2019. Élire, un projet ambitieux au service du Répertoire électoral unique. In : *Courrier des statistiques*. 27 juin 2019. Insee. N° N2, pp. 58-71. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168399/courstat-2-7.pdf*.
- DESABIE, Jacques et HAYOUN, Michel, 1987. Le répertoire d'identification des personnes physiques. In : *Pour une histoire de la statistique*. Éditions Joëlle Affichard, Economica, Tome 2, Matériaux, pp. 57-62. ISBN 978-271781261-X.
- DESABIE, Jacques, 1970. L'Insee entreprend d'automatiser le répertoire des personnes. In : *Économie et statistique.* [en ligne]. Mars 1970. Insee. N°10, pp. 69-71. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://doi.org/10.3406/estat.1970.1930*.
- INSEE, 2021. Dématérialisation des échanges de données. In : *site de l'Insee*. [en ligne]. 13 janvier 2021. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/information/1303515*.
- INSEE, 2022. Nombre de décès quotidiens France, régions et départements. In : *site de l'Insee*. [en ligne]. 30 septembre 2022. [Consulté le 21 octobre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6433839?sommaire=4487854&q=STATISTIQUES+MORTALITE*.
- LANG, Gérard, 2018. Histoire : Éléments pour une histoire du « Numéro de sécurité sociale ». In : *Statistique et société.*15 juin 2018. Société française de Statistique (SFdS). Vol. 6, N° 1, pp. 37-45. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *http://revues-sfds.math.cnrs.fr/index.php/stat_soc/article/view/678*.
- PAPON, Sylvain, 2022. *Bilan démographique 2021. La fécondité se maintient malgré la pandémie de Covid-19.* [en ligne]. 18 janvier 2022. Insee Première n°1889. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6024136*.
- POULAIN, Michel et HERM, Anne, 2013. Le registre de population centralisé, source de statistiques démographiques en Europe. In : *Population*. [en ligne]. Avril-juin 2013. Ined. Vol. 68, 2013/2, pp. 215-247. ISBN 978-2733201695. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.cairn.info/revue-population-2013-2-page-215.htm*.
- ROBERT-BOBÉE, Isabelle et GUALBERT, Natacha, 2021. L'échantillon démographique permanent : en 50 ans, l'EDP a bien grandi ! In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 47-63. [Consulté le 7 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398685/courstat-6-art-3.pdf*.
- WALLGREN, Anders et WALLGREN, Britt, 2007. *Register Based Statistics: Administrative Data for Statistical Purposes.* 21 mai 2007. Éditions John Wiley & Sons. ISBN 978-0470027783.

▶ Fondements juridiques

- Décret n° 2019-341 du 19 avril 2019 relatif à la mise en œuvre de traitements comportant l'usage du numéro d'inscription au répertoire national d'identification des personnes physiques ou nécessitant la consultation de ce répertoire. In : *site de Légifrance.* [en ligne]. Mise à jour le 29 octobre 2022. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000038396526*.
- Décret n° 46-1432 du 14 juin 1946 portant règlement d'administration publique pour l'application des articles 32 et 33 de la loi de finances du 27 avril 1946 relatifs à l'institut national de la statistique et des études économiques pour la métropole et la France d'outre-mer. In : *site de Légifrance.* [en ligne]. Mise à jour le 12 juin 1989. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000872629*.
- Décret n° 82-103 du 22 janvier 1982 relatif au répertoire national d'identification des personnes physiques. In : *site de Légifrance.* [en ligne]. Mise à jour le 1er juillet 2021. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000520382*.
- Loi n° 2021-1017 du 2 août 2021 relative à la bioéthique (1). In : *site de Légifrance.* [en ligne]. Mise à jour le 4 août 2021. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000043884384*.
- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. In : *site de Légifrance.* [en ligne]. Mise à jour le 26 janvier 2022. [Consulté le 18 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460*.

Un référentiel des identités pour les besoins de la sphère sociale

Le système national de gestion des identifiants (SNGI)

Joseph Préveraud de Vaumas*

Le système national de gestion des identifiants (SNGI) est le référentiel des identités, pour les besoins des organismes de la protection sociale. Créé en 1988 par la Caisse nationale d'assurance vieillesse (Cnav), le SNGI traite les états civils et le NIR (numéro d'inscription au répertoire, plus connu sous le nom de numéro de sécurité sociale) des ayants droit de la sécurité sociale. Au fil du temps, il s'est imposé comme référentiel socle notamment parce qu'il permet l'attribution d'un NIR aux individus nés hors de France.

Construit à partir des fichiers des assurés du régime général de retraite de la Cnav, il a progressivement été synchronisé avec le répertoire national d'identification des personnes physiques (RNIPP) de l'Insee, et continue de s'enrichir en permanence. Outre les données nécessaires à l'identification d'une personne, il contient des informations liées à l'instruction des demandes d'immatriculation. Le système intègre des fonctionnalités de consultation et de recherche d'identité ainsi que de diffusion d'information. Il s'appuie sur un moteur d'identification performant pour retrouver une identité à partir d'informations imprécises voire inexactes. Compte-tenu des enjeux importants pour ses utilisateurs et du caractère sensible des données qu'il contient, le SNGI est très encadré sur le plan juridique.

The National Identification Management System (SNGI) is the repository of identities for the needs of social protection organizations. Created in 1988 by the main pension scheme in France (CNAV – Caisse nationale d'assurance vieillesse), the SNGI processes the civil status and NIR (National Registration Number, better known as the social security number) of Social Security beneficiaries. Over time, it has become the basic reference system, particularly because it allows a NIR to be assigned to individuals born outside France.

Built from the files of insured persons in the CNAV's general pension scheme, it has been progressively synchronized with the INSEE's National Register for the Identification of Individuals (RNIPP), and continues to be enriched on an ongoing basis. In addition to the data needed to identify a person, it contains information related to the processing of registration requests. The system integrates functions for consulting and searching for identities as well as for disseminating information. It relies on a powerful identification engine to find an identity from imprecise or even inaccurate information. Given the high stakes for its users and the sensitive nature of the data it contains, the SNGI is highly regulated by law.

* Responsable du pôle Étude et Transformation du SI, Cnav,
Joseph.Preveraud-de-Vaumas@cnav.fr

Au commencement, connaître les identités pour gérer les retraites

La Caisse nationale d'assurance vieillesse (Cnav) est l'organisme de sécurité sociale qui gère les retraites des salariés du régime général et des travailleurs indépendants. Son activité concerne près de 80 millions de Français ou d'étrangers travaillant ou ayant travaillé en France *(Cnav, 2021)*.

Au début du processus de gestion du système de retraites, l'employeur déclare les données « sociales » de ses salariés[1]. La Cnav réalise ensuite un premier traitement pour regrouper ces données par individu et reconstituer la carrière d'une personne qui travaille ou a travaillé dans différentes entreprises *(Sureau et Merlen, 2021)*. Il est alors possible de calculer le droit à la retraite. Pour passer du « salarié » à l'« individu », il faut pouvoir reconnaître une personne à travers les déclarations sociales des différents employeurs. La déclaration sociale contient des données d'identité (nom, prénom, date et lieu de naissance) ainsi que le numéro de sécurité sociale de chaque salarié. C'est à partir de ces informations que les rapprochements vont pouvoir se faire.

> ***Pour passer du « salarié » à l'« individu », il faut pouvoir reconnaître une personne à travers les déclarations sociales des différents employeurs.***

Il faut alors un système de référence fiable qui contienne les données d'identité de chaque individu et qui soit indépendant de l'employeur : **un référentiel des identités *(encadré 1)***. Ce référentiel constitue un enjeu majeur non seulement pour le calcul des retraites, mais aussi pour toutes les branches de la Sécurité sociale (maladie, famille, etc.) et pour tous les régimes qui rencontrent des besoins similaires.

La Sécurité sociale s'est donc dotée du système national de gestion des identifiants (SNGI), un référentiel informatique, mis en œuvre par la Cnav et utilisé par l'ensemble des organismes de la protection sociale française, en commençant par les régimes de retraite.

Dix ans pour passer des fichiers régionaux au référentiel national

Dans les années 70, la masse d'information issue des déclarations sociales était si importante que les capacités informatiques ne permettaient pas de les traiter à un échelon national. Chaque caisse régionale[2] disposait alors d'un Fichier Régional d'Identification (FRI), dont la Cnav pilotait la coordination. Cette organisation a vu ses limites quand les mouvements d'individus d'une région à l'autre se sont multipliés, et qu'il devenait de plus en plus difficile d'assurer la cohérence entre ces référentiels d'identités régionaux.

1 La déclaration sociale nominative (DSN) est réalisée chaque mois par tous les employeurs. Elle est centralisée puis diffusée aux organismes de protection sociale et sert à calculer le droit aux différentes prestations (assurance maladie, retraite, etc.). Voir *(Humbert-Bottin, 2018)*.

2 Les CRAM (caisses régionales d'assurance maladie), devenues au 1er juillet 2010 Carsat (caisses d'assurance retraite et de la santé au travail).

En 1988, l'amélioration des performances des ordinateurs a permis à la Cnav de centraliser les fichiers d'identification dans une base nationale : c'était la naissance du système national de gestion des identifiants (SNGI) ***(encadré 2)***. La convergence des fichiers régionaux vers le système national n'a pas été un simple transfert ou regroupement de fichiers : il a fallu plusieurs années pour progressivement analyser et résoudre les écarts qu'il pouvait y avoir entre les fichiers régionaux.

En parallèle, il devenait essentiel de renforcer la qualité des données qui enrichissent le référentiel national.

> ***C'est donc par intérêt mutuel que la Cnav et l'Insee ont mis en place des échanges informatiques.***

La Cnav a constitué en 1988 un service d'experts, appelé **Sandia**[3], qui assure la certification d'identité des personnes nées hors de France et ayant droit à une prestation sociale. Par ailleurs, l'Insee gérait un autre référentiel des identités : le RNIPP[4]. Alimenté directement à partir des états civils des mairies, le répertoire national d'identification des personnes

▶ Encadré 1. Le référentiel d'identité en questions

Le SNGI identifie, mais il n'authentifie pas ?

Le SNGI permet de retrouver une identité certifiée à partir des éléments fournis (nom, prénoms, date et lieu de naissance, etc.). Cela ne signifie pas que ces éléments correspondent bien à la bonne personne physique. Pour « authentifier » ce lien, il faudrait par exemple utiliser une photo, une empreinte digitale ou un code secret. Ce n'est pas la fonction du référentiel d'identités.

Mais alors à quoi sert le SNGI ?

Le SNGI associe des données d'identités déclarées provenant de sources incertaines (formulaire papier ou numérique) avec des identités de référence et certifiées. Cela permet ensuite de regrouper efficacement différentes sources d'information relatives à un même individu (par exemple pour reconstituer sa carrière à partir des déclarations de différents employeurs). L'identifiant, le NIR, vient alors en complément de l'identité pour fiabiliser le dispositif.

Pourquoi recourir à un système d'identification alors que l'on dispose de l'identité ?

Trois principales sources d'erreurs peuvent rendre difficile l'identification d'une personne à partir de ses traits d'identités :

- quand le salarié ou l'assuré déclare son identité en remplissant à la main un formulaire, celui-ci peut être ensuite photocopié ou numérisé : l'identité est ainsi recopiée ou saisie dans un système informatique, avec un risque d'erreur accru ;
- les traits d'identité eux-mêmes peuvent varier, par exemple quand le nom marital est renseigné à la place du nom de naissance ou lorsque la liste des prénoms est plus ou moins complète ;
- enfin, la déclaration se fait parfois à partir d'une fausse identité (fraude).

L'enjeu pour le système d'information est à la fois de construire une référence fiable des identités et de pouvoir rapprocher les éléments déclarés et l'identité de référence.

3 Le service administratif national d'identification des assurés (Sandia) est un service de la Cnav dont les équipes sont basées à Tours.

4 Le répertoire national d'identification des personnes physiques (RNIPP). Voir l'article de Lionel Espinasse et Valérie Roux dans ce même numéro.

▶ Encadré 2. Cadre et gouvernance du SNGI

Des textes…

Le SNGI est encadré par un ensemble de décrets, en premier lieu le décret n° 2018-390 du 24 mai 2018* qui précise :

- le rôle de la Cnav comme opérateur du référentiel, c'est-à-dire l'organisme qui met en œuvre le système d'information pour le compte de la sphère sociale ;
- les finalités de ce système ainsi que ses données et leurs règles de conservation ;
- les accès et les destinataires autorisés ;
- et certains aspects liés au règlement général sur la protection des données* ; ainsi, le droit d'opposition, qui permet à une personne de s'opposer à ce que ses données personnelles soient utilisées par un organisme pour un objectif précis, ne s'applique pas au SNGI.

Le décret SNGI régule ses usages en raison du caractère particulièrement sensible du référentiel, puisqu'il permet de croiser les informations personnelles venant de différentes sources.

Le cadre juridique est aussi complété par le décret « NIR » n° 2019-341 du 19 avril 2019*, et par d'autres décrets comme ceux relatifs aux systèmes adossés, tels que le RNIAM et le RNCPS.

… des directives…

Pour encadrer ce processus d'immatriculation, le ministère des Affaires sociales édite notamment un guide de l'Identification à destination des organismes de la protection sociale. Le ministère de l'Intérieur et celui des Affaires étrangères contribuent également à la définition des directives pour le contrôle des pièces justificatives.

… des instances…

La CNIL porte une attention particulière au SNGI, elle veille notamment au respect de la confidentialité des données et à leur usage dans un cadre strictement maîtrisé.

Au sein du ministère des Affaires sociales et de la Santé, la direction de la Sécurité sociale (DSS) anime le comité opérationnel de suivi de l'identification (COSI), lequel regroupe les principaux organismes de protection sociale en charge de l'identification et de l'immatriculation des individus, et l'Insee. Le COSI veille à la qualité du processus d'identification et d'immatriculation des individus, et gère le suivi des évolutions du référentiel. Il décide des actions à mener, par exemple quand des fraudes importantes sont détectées dans certains pays étrangers. Il facilite la coordination des différents organismes et apporte des précisions sur des cas particuliers d'identification.

Enfin, l'extension des usages du SNGI attire régulièrement l'attention des organes de contrôles tels que l'Inspection générale des affaires sociales, la Cour des comptes ou encore certaines enquêtes parlementaires.

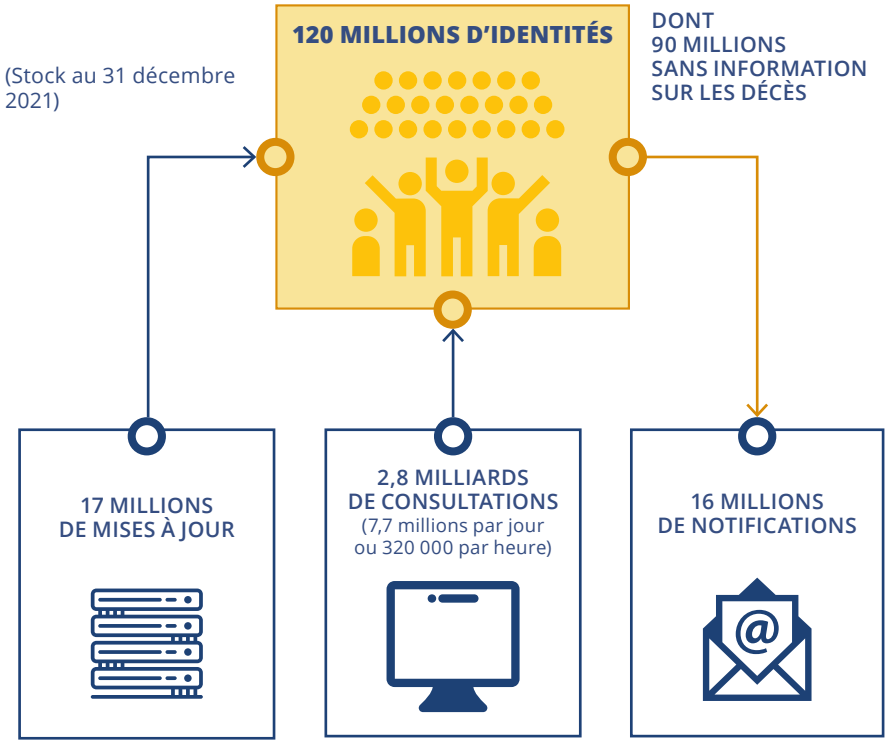
… et des cadres techniques

Le schéma stratégique des systèmes d'information (SSSI) de la Sécurité sociale encadre les grands travaux informatiques. Il est complété par la Convention d'objectifs et de gestion et le schéma directeur des systèmes d'information (SDSI) de la Cnav, qui précisent les objectifs pluriannuels de l'opérateur du SNGI.

** Voir les références juridiques en fin d'article et (CNIL, 2020).*

physiques est une source très fiable d'information pour les personnes nées en France. C'est donc par intérêt mutuel que la Cnav et l'Insee ont mis en place des échanges informatiques dès les années soixante-dix. À l'occasion de la mise en œuvre du SNGI, l'Insee a délégué à la Cnav l'immatriculation au RNIPP des personnes nées à l'étranger[5]. Le partage d'information entre la Cnav et l'Insee a commencé par les nouvelles identités (dont les naissances) qui ont enrichi simultanément les deux référentiels. Puis l'échange s'est progressivement étendu à l'ensemble des identités, notamment celles issues des fichiers régionaux de la Cnav, et a ainsi abouti à une synchronisation totale des deux référentiels en 1998. Aujourd'hui les données du SNGI proviennent d'une part du RNIPP géré par l'Insee pour toutes les personnes nées en France et d'autre part du processus de certification géré par le Sandia pour toutes les personnes nées hors de France (voir *infra*).

▶ Encadré 3. Le SNGI en chiffres



* Il y a environ 2 millions d'identité en plus au SNGI qu'au RNIPP, car le SNGI contient des identités non certifiées qui ne sont pas transmises au RNIPP.

** Le nombre important de consultations vient principalement du flux DSN et RGCU : chaque déclaration mensuelle de chaque salarié nécessite au moins un passage au SNGI. Voir *(Humbert-Bottin, 2018)* et *(Sureau et Merlen, 2021)*.

5 La délégation est en vigueur depuis la signature d'un protocole le 25 juin 1987. Elle est régie par une convention cadre régulièrement mise à jour. La délégation concerne toutes les personnes nées à l'étranger, qu'elles soient françaises ou pas, et les personnes nées dans certaines collectivités territoriales (Nouvelle-Calédonie, etc.).

Au sein de la sphère sociale, un déploiement progressif et mutuellement profitable

Si au départ le SNGI servait principalement à la gestion des retraites du régime général, les actions mises en œuvre pour garantir la qualité de ses données en ont fait un référentiel fiable, ce qui a permis de lui donner un nouvel essor.

Ainsi, en 1998, lorsque le gouvernement a décidé de déployer les cartes d'assurance maladie (appelées cartes Vitale), il fallait s'assurer que chaque individu éligible recevrait bien une carte et une seule. Un référentiel unique et partagé pour l'ensemble des caisses des différents régimes d'assurance maladie permettait de garantir le rattachement unique d'un individu à l'une de ces caisses. C'est ainsi qu'un nouveau référentiel, le RNIAM[6], a été mis en œuvre. Il a été adossé au SNGI pour la gestion des identités des individus qui le composent.

Près d'une décennie plus tard, afin de lutter contre le non-recours aux droits et contre la fraude, le gouvernement a souhaité mettre en place un nouvel outil pour faciliter le partage d'informations entre les organismes de la sécurité sociale. Le répertoire nommé RNCPS[7] déployé en 2009 est lui aussi adossé au SNGI.

> ***Cette extension de l'usage du SNGI est autant liée à sa qualité qu'elle y contribue.***

Cette extension de l'usage du SNGI est autant liée à sa qualité qu'elle y contribue. En effet, le partage d'informations issues des organismes partenaires a permis à chaque fois de détecter puis de corriger des anomalies sur certaines identités.

Désormais, le SNGI est incontournable pour la plupart des systèmes d'information de la sphère sociale. Il intervient ainsi dans la mise en œuvre de dispositifs comme la déclaration sociale nominative, le dossier médical partagé ou le calcul des allocations logements.

Il sert également à partager les informations d'identité avec d'autres administrations, par exemple pour le compte personnel de formation[8] ou le prélèvement des impôts à la source[9].

SNGI et RNIPP, deux systèmes étroitement liés et complémentaires

Le SNGI et le RNIPP ont en commun de « gérer » l'identité, mais ils ont chacun des objectifs et des usages distincts qui expliquent leur coexistence.

Le RNIPP est alimenté en partie par le SNGI mais surtout par l'ensemble des mairies de France ce qui implique beaucoup de flux en entrée. Le SNGI est alimenté par le RNIPP,

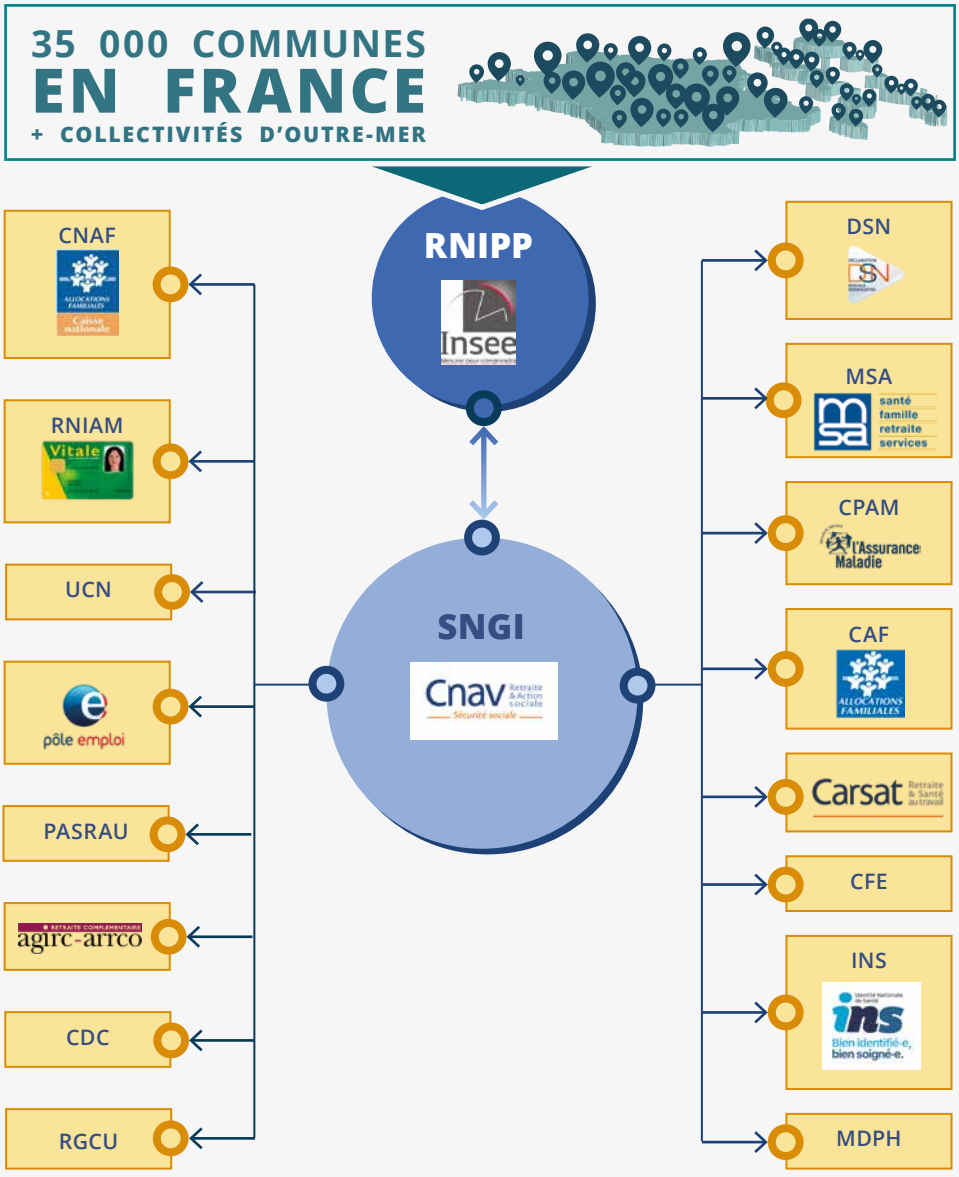
6 Le répertoire national inter-régime de l'assurance maladie (RNIAM) contient pour chaque individu, le régime et la caisse d'assurance maladie à laquelle il est rattaché. Il sert au dispositif de délivrance des cartes Vitale. *(CNIL, 2009).*

7 Le répertoire national commun de la protection sociale (RNCPS) contient les rattachements d'un individu à chaque organisme de protection sociale qui lui attribue une prestation ; il permet d'avoir les informations sur ces prestations perçues et les adresses déclarées pour les percevoir.

8 Le compte personnel de formation est un dispositif de financement de la formation continue ; il est géré par la Caisse des dépôts et consignations (CDC) (*DGEFP, 2022).*

9 Le prélèvement à la source concerne les salaires mais aussi d'autres sources de revenus dont les prestations sociales imposables. Le SNGI participe à cette partie du dispositif appelée PASRAU (prélèvement à la source des revenus autres).

Figure 1 - Le SNGI au cœur de la sphère sociale



ORGANISMES DE PROTECTION SOCIALE

- *Agirc-Arrco : caisse de retraite complémentaire*
- *CAF : caisse d'allocations familiales*
- *Carsat : caisses d'assurance retraite et de la santé au travail*
- *CDC : caisse des dépôts et consignations*
- *CFE : Caisse des Français de l'étranger*
- *CNAF : caisse nationale des allocations familiales*
- *CPAM : caisse primaire d'assurance maladie*
- *DSN : déclaration sociale nominative*
- *INS : identité nationale de santé*
- *MDPH : maison départementale pour les personnes handicapées*
- *MSA : mutualité sociale agricole*
- *PASRAU : passage des revenus autres*
- *RGCU : répertoire de gestion des carrières unique*
- *RNIAM : répertoire national interrégimes des bénéficiaires de l'assurance maladie*
- *UCN : URSSAF caisse nationale*

par quelques organismes de protection sociale et par le Sandia ***(figure 1)***, ce qui est sans commune mesure avec la masse des flux gérés par l'Insee.

En termes de restitutions, le rapport est inverse : le SNGI est très utilisé en consultation ***(encadré 3)*** pour vérifier une identité ou identifier une personne en tenant compte du NIR, alors que le RNIPP n'offre cette fonctionnalité qu'à partir des traits d'identité.

Pour la gestion des identités, le RNIPP s'appuie sur les actes de naissance délivrés par les officiers d'état civil des mairies (pour les personnes nées en France) alors que c'est le Sandia qui certifie les identités des personnes nées à l'étranger. Cette différence de gestion, spécifique à la Sécurité sociale, permet de verser des prestations sociales aux personnes qui ne sont pas nées en France.

▶ Selon le lieu de naissance, deux démarches et deux circuits différents

Le SNGI contient des données d'état civil (nom, prénom, date et lieu de naissance, etc.) associées à un identifiant : le numéro d'inscription au répertoire (NIR) plus connu sous l'appellation de « numéro de sécurité sociale ». Le référentiel est alimenté par deux flux principaux de données ***(figure 2)*** :

- le flux provenant de l'Insee, *via* le RNIPP, pour les personnes nées en France[10] ;
- et un flux provenant du Sandia pour les personnes nées hors de France.

Pour ajouter une nouvelle identité au SNGI, il faut commencer par lui attribuer un NIR ***(encadré 4)***. Cette opération s'appelle **l'immatriculation**.

Les personnes nées en France sont immatriculées *automatiquement* au RNIPP par l'Insee après la déclaration de naissance faite en mairie[11]. Mais pour les autres, l'immatriculation résulte d'une *démarche volontaire* : elles doivent au préalable avoir effectué une demande, soit pour pouvoir bénéficier d'une prestation sociale (allocation familiale, couverture maladie, etc.), soit pour que leur employeur puisse renseigner la déclaration sociale nominative. La demande d'immatriculation s'effectue auprès d'un organisme de sécurité sociale (caisse primaire d'assurance maladie, caisse d'allocation familiale, etc.). Elle doit être accompagnée de deux pièces justificatives de l'identité (pièce d'état civil et pièce d'identité) qui permettent d'une part de vérifier la cohérence des pièces entre elles et d'autre part d'avoir les données de filiation (nom et prénom des parents). Ces dernières permettent de distinguer des personnes aux identités semblables (mêmes nom, prénom, date et lieu de naissance). Le risque de confusion est d'autant plus grand que pour une naissance à l'étranger, le lieu retenu est le *pays* et non pas la *commune* comme pour les naissances en France. La demande d'immatriculation et les pièces justificatives sont enfin transmises par l'organisme de sécurité sociale au Sandia, lequel réalise les contrôles nécessaires à la **certification** de l'identité.

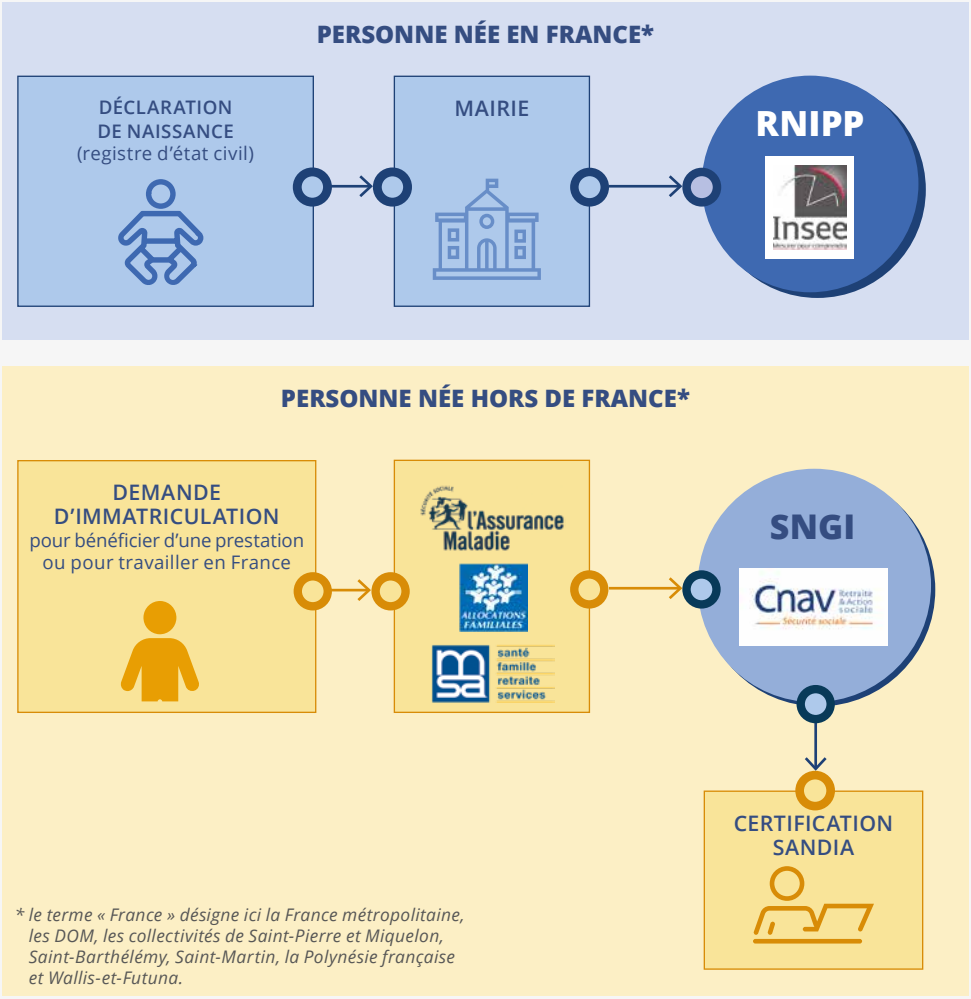
10 Et ce, quelle que soit la nationalité de ces personnes. Plus précisément, le flux concerne les personnes nées en métropole, dans les départements d'outre-mer, et dans quelques collectivités territoriales (Saint-Pierre et Miquelon, Saint-Martin, Saint-Barthélémy et la Polynésie Française). Le flux « Sandia » est le complément à ce flux.

11 Voir l'article de Lionel Espinasse et Valérie Roux sur le RNIPP, dans ce même numéro.

La suite du processus se fait en deux temps : un numéro provisoire (numéro identifiant d'attente ou NIA) est d'abord attribué à la personne sur la base d'au moins une pièce justificative, puis le numéro définitif (qui peut être le même) est attribué quand l'identité est certifiée par le Sandia.

L'attribution d'un numéro provisoire ou définitif n'ouvre pas en tant que tel des droits aux prestations sociales (l'attribution d'une prestation est gérée à part), mais il permet aux organismes de les verser. À l'inverse, si le processus de certification n'aboutit pas dans un délai de neuf mois (par exemple si la personne ne fournit pas la seconde pièce justificative), alors les organismes de sécurité sociale doivent interrompre le versement des prestations. Dans ce cas, l'identité est conservée au SNGI, pour des besoins de suivi et de traçabilité. De ce fait, le SNGI contient mécaniquement plus d'identités qu'il n'y a de personnes percevant réellement une prestation.

Figure 2 - Deux circuits d'immatriculation pour deux répertoires complémentaires



▶ Une mise à jour permanente et largement automatisée

Le SNGI est un référentiel, mais c'est aussi un système d'information dont les fonctionnalités se regroupent dans trois types de service (au sens informatique du terme) : l'alimentation, l'accès aux données et la diffusion des informations pour les partenaires de la sphère sociale.

Pour une bonne alimentation du référentiel, le SNGI s'appuie essentiellement sur des échanges automatisés.

Les échanges quotidiens entre le SNGI et le RNIPP permettent de synchroniser les mises à jour réalisées de part et d'autre. Ainsi le SNGI reçoit les informations sur les personnes nées en France et transmet celles sur les personnes nées hors de France. Certaines spécificités, telles que les identités en cours de certification au SNGI, ne sont toutefois pas incluses dans cette synchronisation.

L'immatriculation et la certification sont deux services phares du SNGI. Ils permettent d'enregistrer les demandes d'immatriculation, d'attribuer un NIR ou un NIA (voir *supra*), de gérer le processus de certification par le Sandia, d'informer les partenaires du traitement de leur demande.

Un circuit d'alimentation spécifique du SNGI existe pour les Français nés hors de France, *via* le Service central d'état civil, service du ministère de l'Europe et des Affaires étrangères (basé à Nantes). Cette source de données permet une immatriculation automatique, sans mobiliser le processus de certification de l'identité du Sandia.

▶ Encadré 4. Le NIR au cœur du SNGI

Le numéro d'inscription au répertoire est un **identifiant unique** dont la genèse est bien antérieure au SNGI *(CNIL, 2000)*. C'est un élément central, associé à chaque identité du SNGI. Il permet entre autres de référencer chaque identité de manière unique, d'indexer les données, de partager facilement avec d'autres systèmes.

Le NIR est constitué de 13 caractères :

2	57	04	35555	261
S code sexe	AA année de naissance	MM mois de naissance	LLLLL code lieu de naissance	OOO numéro d'ordre

Chacune des composantes du NIR traduit la complexité de la gestion des identités. Ainsi, le mois de naissance va de 1 à 12 mais peut aussi prendre d'autres valeurs (de 20 à 99) qui indiquent que le mois de naissance de l'individu n'est pas connu (cela peut arriver pour des personnes nées à l'étranger).

Pour une commune de métropole, le code lieu de naissance est composé des deux chiffres du code de département (ou 2A / 2B pour la Corse) puis des 3 chiffres du code de la commune. Pour un pays étranger, le code commence par 99 puis est suivi par le code à 3 chiffres de ce pays.

Le SNGI gère l'actualisation des **informations de décès** (date et lieu). Comme pour les naissances, il existe plusieurs circuits en fonction du lieu de décès. Pour une personne décédée en France, le certificat arrive en mairie qui remonte l'information à l'Insee et c'est ensuite le mécanisme de synchronisation avec le RNIPP qui fait le lien avec le SNGI. Pour une personne décédée à l'étranger, c'est le plus souvent un proche de l'assuré qui transmet l'information à l'organisme de sécurité sociale concerné, lequel la transmet ensuite directement au SNGI. Le système mémorise l'information avec un indice de certification permettant de savoir si l'information est déclarative ou fondée sur une pièce justificative. Quand c'est un Français qui décède à l'étranger, l'information peut également parvenir *via* un circuit allant des ambassades ou des consulats vers les mairies de résidence en France, puis vers le RNIPP. L'enjeu est important, notamment pour le paiement des retraites des personnes résidant à l'étranger. Sans l'information de décès, les régimes de retraite concernés continueraient à verser des pensions à tort. Pour éviter ces situations, le SNGI dispose d'un service d'échange d'informations avec certains pays, selon des conventions signées entre états. Ce dispositif permet d'avoir une information plus fiable. Un dispositif complémentaire de « contrôle d'existence » permet aux organismes d'interroger les bénéficiaires de prestations qui résident à l'étranger pour s'assurer qu'ils sont toujours en vie. En cas de non-réponse, l'organisme peut suspendre la prestation même si aucune information de décès n'est présente au SNGI[12].

> ***Sans l'information de décès, les régimes de retraite concernés continueraient à verser des pensions à tort.***

Le SNGI intègre également des services pour **mettre à jour une identité.** Une mise à jour est nécessaire soit lorsqu'une identité est mal enregistrée dans le système (il s'agit donc d'une correction), soit lorsque les traits d'identité évoluent. C'est par exemple le cas lors d'une adoption (le nom de famille change), lors de l'ajout ou du changement d'un nom d'usage (mariage, divorce, etc.), lors de l'ajout d'un accent pour les identités accentuées (historiquement, les accents n'étaient pas gérés au SNGI).

▶ Deux modes d'accès aux données : vérification ou identification

Outre les services d'alimentation ou de mise à jour, le système prévoit des fonctionnalités permettant un accès contrôlé aux données. Ces services sont utilisés par les organismes de la sphère sociale soit *via* des échanges automatisés entre systèmes d'information, soit par des agents de ces organismes. Les assurés (le grand public) ne peuvent pas accéder directement au SNGI. **La vérification** s'effectue à partir du NIR et du nom de l'individu, et donne alors accès à l'ensemble des autres données d'état civil (prénoms, date et lieu de naissance, information « décès », etc.). Avec cette double clé de consultation, le système vérifie la cohérence des informations *avant* de fournir les éléments d'identité. Ce contrôle permet à la fois d'éviter de renvoyer une mauvaise identité à la suite d'une erreur de saisie et d'assurer une protection des données personnelles en limitant l'accès aux seules personnes disposant

12 Dans la sphère sociale, on parle de personnes « présumées » vivantes puisque l'absence d'information sur le décès ne signifie pas toujours que la personne est en vie. Dans les statistiques tirées du SNGI, on considère en général qu'un individu est présumé vivant quand il n'y a pas d'information sur un décès et que l'âge est inférieur à 110 ans.

de ces deux informations. Ce service permet ainsi de vérifier que le NIR connu du demandeur est correct et d'obtenir l'identité certifiée par le SNGI.

L'identification permet quant à elle d'effectuer une recherche d'identité. Elle sert notamment lorsque le NIR n'est pas connu ou que les traits d'identité sont incomplets ou inexacts. Cette fonction recherche dans le référentiel les individus ayant les traits d'identité les plus proches (voir *infra* la description du moteur d'identification). Elle implique donc de contrôler *a posteriori* que le résultat correspond bien à l'identité recherchée. Pour limiter les risques d'erreur, le SNGI ne renvoie un résultat que s'il est jugé suffisamment proche de la demande et qu'il n'y a pas d'autres identités approchantes. Cette opération est très utilisée dans le traitement des déclarations sociales qui peuvent contenir des fautes de frappe ou des imprécisions (par exemple le salarié donne un prénom qui n'est pas son premier prénom à l'état civil ou un nom d'usage à la place d'un nom de naissance).

> ***Cette opération est très utilisée dans le traitement des déclarations sociales qui peuvent contenir des fautes de frappe ou des imprécisions.***

▶ Un traitement central : le moteur d'identification

Le service d'identification permet de retrouver une identité à partir d'éléments plus ou moins complets et plus ou moins exacts. Il est au cœur des traitements du SNGI. Il intervient par exemple quand un opérateur saisit une identité au clavier et fait une faute de frappe. Il est très utile pour retrouver une identité parmi plusieurs qui seraient très proches, ou quand le lieu mentionné correspond à un ancien libellé. Il peut rectifier le cas d'une personne qui communiquerait son propre NIR en l'associant à l'identité de son enfant. Ce service s'appuie sur un composant technique essentiel : le **moteur d'identification**.

Ce « moteur » intègre l'ensemble des règles spécifiques à l'identification d'une personne à partir d'informations de son état civil (par exemple un nom marital communiqué à la place du nom de naissance, une liste incomplète de prénoms, etc). Cela permet de reconnaître au mieux une identité parmi celles connues du SNGI tout en tenant compte des imprécisions ou erreurs possibles dans la demande.

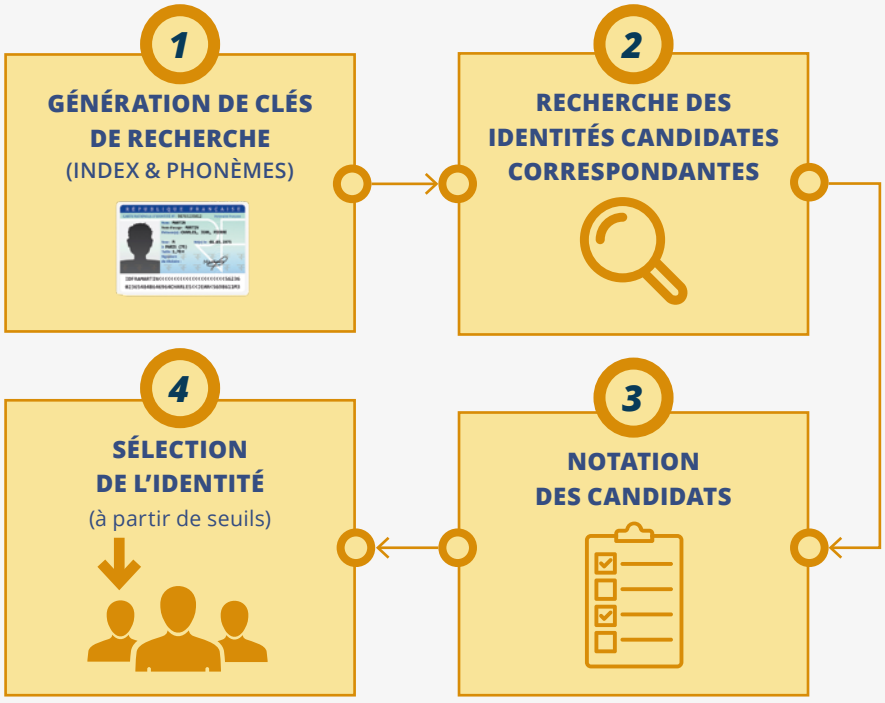
Le moteur intègre un autre ensemble de règles pour ne pas retourner d'identités s'il y a trop de risque d'erreur sur le résultat, soit parce que l'identité trouvée est trop éloignée de la demande, soit parce que trop d'identités pourraient correspondre à la demande (auquel cas il faudra préciser celle-ci).

Le moteur d'identification procède en quatre étapes ***(figure 3)*** :

- tout d'abord il génère des clés de recherche pour chaque élément disponible pour la recherche (nom, prénom, etc.), en s'appuyant notamment sur des règles phonétiques qui permettent de retrouver une identité même lorsqu'elle est mal orthographiée ;
- puis le moteur sélectionne toutes les identités du SNGI correspondantes aux clés constituées à l'étape précédente. Celles-ci permettent de cibler la recherche et contribuent donc à l'optimisation du traitement ;
- chaque identité sélectionnée est ensuite évaluée : une note est d'abord attribuée pour chacun des critères renseignés dans la demande. Ces notes sont pondérées en fonction de leur importance pour calculer une note globale associée à chaque identité trouvée. Les résultats sont ainsi classés en fonction de leur pertinence par rapport aux éléments fournis dans la recherche ;
- enfin, les notations sont étalonnées selon différents seuils pour ne sélectionner que l'identité la plus pertinente ou aucune (s'il n'y a pas de correspondance trouvée ou au contraire s'il en existe plusieurs mais sans possibilité de discerner la bonne).

La qualité de l'identification dépend de plusieurs facteurs : la qualité des données du SNGI, la présence de l'identité recherchée dans le référentiel, la complétude et l'exactitude des éléments fournis dans la recherche. Si une identité est recherchée au SNGI alors qu'elle n'existe pas, le moteur pourrait tout de même renvoyer un résultat approchant. Des contrôles automatiques permettent de limiter ces situations, mais *in fine* seul un contrôle manuel peut permettre de conclure.

Figure 3 - Le moteur d'identification en 4 étapes



La qualité des données de la demande joue un rôle fondamental. On peut comparer le moteur d'identification à un moteur de recherche internet : plus les critères sont inexacts ou imprécis, plus le risque d'obtenir un résultat non pertinent est élevé.

Les caractéristiques de l'identité recherchée influent aussi fortement : le risque de se tromper est plus important si on recherche une identité très répandue (« Monsieur *Dupont* né à Paris ») que si on recherche une identité rare (« Monsieur *DubateauKiflotte* né dans un petit village »).

Dans la DSN, l'employeur renseigne le NIR et l'état civil de ses salariés. Quand l'une des informations est manquante ou inexacte, c'est le moteur d'identification qui permet de retrouver la bonne identité avec les données d'état civil certifiées. Là encore, des contrôles automatiques ou manuels sont lancés si les informations déclarées diffèrent de celles connues du SNGI. Par exemple, un « bilan d'identification des salariés » est transmis à l'employeur, ce qui lui permet, le cas échéant, de réinterroger ceux dont l'identité est mal libellée.

Le service est régulièrement éprouvé pour couvrir certaines situations particulières. Il est par exemple arrivé de devoir identifier des triplées (donc avec même nom, même date et lieu de naissance) ayant des prénoms très proches. L'exemple de l'***encadré 5*** illustre la complexité de l'exercice. Si un organisme concerné dispose de peu d'éléments, alors le moteur d'identification trouvera plusieurs identités susceptibles de correspondre à ces éléments. L'analyse des divergences par un expert de l'identification permet de sélectionner l'identité la plus vraisemblable pour mener ensuite les actions nécessaires à la certification de l'identité.

▶ La diffusion d'information vers les partenaires

Les organismes qui utilisent le SNGI gèrent eux-mêmes une grande quantité d'identités ; ils possèdent leur propre référentiel d'identité : le SNGI constitue donc en quelque sorte un « référentiel-socle »[13] auquel des référentiels « secondaires » sont adossés.

Pour leur éviter de réinterroger en permanence le SNGI, le système offre à ses partenaires un service d'« abonnement » qui les notifie quand un changement survient sur une identité. Comme il s'agit d'informations sensibles, seuls les organismes qui « connaissent » un individu reçoivent les mises à jour le concernant. Pour cela, chaque organisme doit « s'abonner » aux seuls individus qui l'intéressent.

> ***Les notifications concernent tous les types d'actualisation : ajout d'un nom d'usage, d'une information « décès », correction d'une donnée, etc.***

Les notifications concernent tous les types d'actualisation : ajout d'un nom d'usage, d'une information « décès », correction d'une donnée, etc. Certaines mises à jour sont plus délicates, notamment quand elles impactent le NIR de l'individu. C'est par exemple le cas quand on corrige le mois ou l'année de naissance qui font partie intégrante de l'identifiant. Il arrive aussi (rarement) qu'une personne soit connue sous plusieurs NIR au SNGI et, dans ce cas, une opération de fusion des identités est réalisée[14]. Ce type de correction

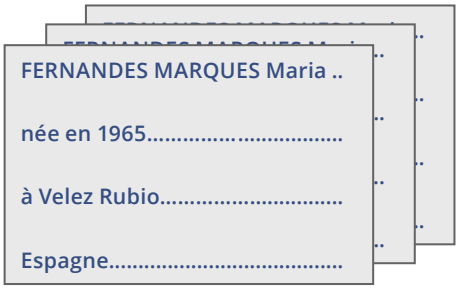
13 Sur ces notions, voir l'article de Pascal Rivière dans ce même numéro.
14 Il y a environ 10 000 recompositions de NIR par an et autant de fusions d'identités.

Encadré 5. Un exemple (fictif) pour illustrer la complexité de l'identification

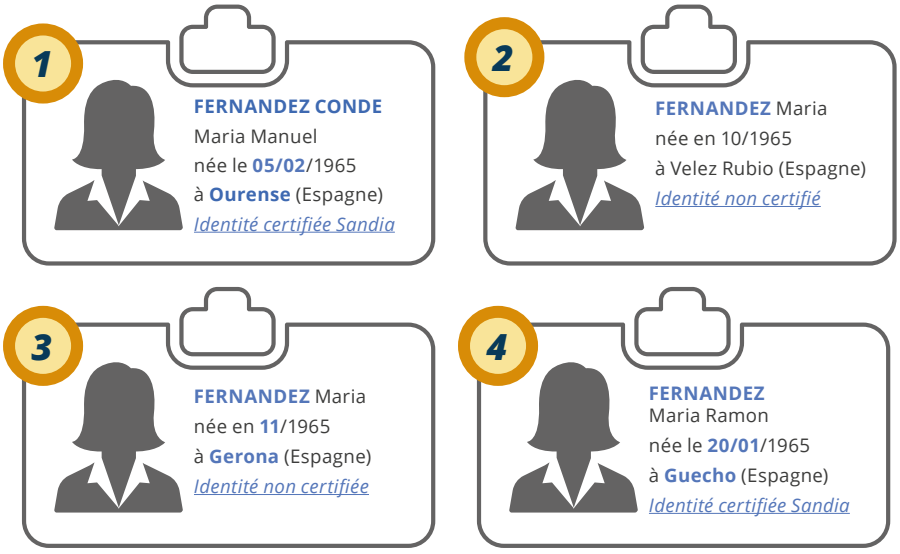
Supposons que l'on cherche à identifier :

- Maria FERNANDES
- née le 14 octobre 1965
- à Velez Rubio en Espagne.

...à partir d'informations issues de saisies partielles et transmises sous cette forme au SNGI :



L'étape d'identification va conduire à proposer quatre résultats avec plus ou moins d'éléments divergents :*



Il existe une forte probabilité que le résultat n°2 concerne l'individu recherché ; cette conclusion reste à confirmer, en vérifiant la filiation ou d'autres éléments du dossier. Si le NIR est confirmé, il faudra ensuite transmettre les pièces pour la certification de l'identité.

** Le service d'identification renvoie au plus une identité. Cependant il se décline dans un mode « expert » qui permet de renvoyer jusqu'à 10 identités quand les identités sélectionnées sont voisines (semblables). Ce mode d'identification est réservée à des agents spécialement formés pour traiter ces situations particulières.*

d'anomalie implique également des traitements spécifiques au niveau de chaque organisme partenaire, pour revoir les éléments associés à chacune des deux identités fusionnées.

Le SNGI offre également un service de notification spécifique pour le compte personnel de formation (voir *supra*). Un compte est créé automatiquement par la Caisse des dépôts et consignations pour toute personne âgée de 16 ans ou plus. Le SNGI envoie donc à ce partenaire des notifications automatiques basées sur un critère d'âge et non plus sur un critère d'abonnement individuel.

Le volume d'échanges et le nombre important de partenaires imposent de normaliser les échanges : c'est le rôle de la « norme A » du SNGI[15]. Elle offre une grande flexibilité tout en permettant une optimisation des traitements. Conçue dans les années 80, elle intègre bon nombre de principes avant-gardistes pour l'époque bien qu'aujourd'hui généralisés à travers des formats standard comme Xml[16]. Cette norme sert à la fois pour les échanges par fichiers pour les traitements de masse et elle est aussi le socle des échanges par *Web* Service pour les traitements unitaires.

▶ Pour pouvoir fournir les données, il faut… d'autres données

Pour rendre tous ses services, le SNGI s'appuie sur des informations que l'on peut regrouper en différentes catégories :

- les **données d'identification** qui définissent une identité (nom, prénoms, etc.) et l'identifiant, le **NIR** ; ces données sont elles-mêmes réparties en deux types : celles qui font l'objet d'une certification par l'Insee ou le Sandia[17] et celles qui peuvent compléter l'identité sans pour autant être certifiées sur la base de documents officiels[18] ;
- les données du dossier d'instruction **nécessaires à l'immatriculation d'un individu et à la certification de son identité** (état du dossier, pièces justificatives, etc.). Elles sont conservées au SNGI, car elles servent de preuves et garantissent la traçabilité des traitements ;
- les **données pour la diffusion**, notamment les abonnements des organismes aux notifications sur les identités ;
- la **nomenclature des lieux** qui permet d'identifier un lieu de naissance ou de décès ;
- les **données de gestion interne** nécessaires au bon fonctionnement du référentiel (historique et suivi des mises à jour, gestion des incohérences, données de paramétrages, données techniques, données pour l'optimisation des traitements d'identification, etc.).

Le SNGI en tant que référentiel des identités ne contient que les données nécessaires à ses services et ne contient pas de données connexes sur un individu telles que son adresse, ses coordonnées de contact, etc.

Alors que l'immatriculation se base sur une connaissance de l'identité à un instant donné, l'identification est d'autant plus performante qu'elle peut rapprocher d'anciennes

15 Cette norme décrit différents types de messages : les messages « Aller » pour interroger le SNGI, les messages « Retour » pour obtenir la réponse du SNGI, et les messages « Notifications » qui sont émis par le SNGI.
16 On retrouve ainsi dans la norme A, une structuration par blocs, sous blocs et données, la notion d'élément obligatoire ou facultatif, des mécanismes de contrôles de conformité, etc.
17 Les informations d'état civil « certifiables » sont : le nom patronymique et la liste des prénoms, la date et le lieu de naissance, les données de filiation (nom et prénoms des parents), la date et le lieu de décès (certifiés par l'Insee).
18 Les informations « non certifiables » sont par exemple les date et lieu de décès reportés par un partenaire de la sphère sociale ou les identités secondaires (noms maritaux, noms d'usage).

informations (un ancien nom marital, un ancien nom de commune ou de pays de naissance, un ancien NIR quand celui-ci a été corrigé) avec une identité actuelle. Ainsi, au SNGI, les données sont conservées jusqu'à l'extinction des prestations des ayants droit. Du fait de l'usage possible des données *post-mortem* notamment pour attribuer une pension de réversion, les données ne sont actuellement pas purgées.

L'indispensable nomenclature des lieux et la question de leur codification

L'identité comprend notamment le lieu de naissance et éventuellement le lieu de décès de l'individu. Pour que tous les partenaires du SNGI utilisent la même manière de décrire ces lieux, il faut pouvoir s'appuyer sur une norme qui fait référence. Par ailleurs, il faut aussi définir des conventions sur son usage. Par exemple, pour une personne née avant 1992 en URSS et qui voudrait se faire immatriculer de nos jours, quel pays de naissance utilisera-t-on ? L'URSS ou l'une des républiques actuelles issues de son éclatement ? Dans ce genre de situation, c'est le nom de lieu valide *à la date d'immatriculation* qui est utilisé.

Pour gérer les lieux, le SNGI s'appuie sur le code officiel géographique établi par l'Insee. Ce code suit les évolutions de la géographie française et mondiale et tient compte par exemple des changements de nom de pays, de fusion ou d'éclatement de territoire, etc. Il traite aussi les évolutions du type d'un lieu : par exemple pour les anciens départements français qui sont devenus des pays au moment de la décolonisation, ou les fusions de communes, les transformations des collectivités en département (Mayotte), etc.

> ***Le SNGI dispose des données permettant de faire le lien entre un ancien lieu et sa correspondance actuelle.***

Ainsi, le SNGI dispose des données permettant de faire le lien entre un ancien lieu et sa correspondance actuelle. Il dispose des noms officiels et des noms « admis » pour les différents lieux. Cela permet de faciliter l'identification en retrouvant un individu soit à partir d'un nom de lieu actuel soit à partir du nom de lieu valide au moment de sa naissance.

Le code officiel géographique permet, au-delà de la normalisation des lieux, de leur associer un identifiant unique qui sert entre autres à la définition du NIR.

L'omniprésence des cas particuliers

Quand on parle de données d'identité, on n'imagine pas tous les cas particuliers que cela peut recouvrir ! Par exemple, le nom de naissance, aussi appelé nom patronymique, répond classiquement à des règles précises notamment sur les caractères admis, sur la présence possible d'espace, d'apostrophe, de tiret ou de signes diacritiques, mais dans la pratique, il peut présenter des spécificités peu connues. Ainsi il existe des individus sans nom, dont l'état civil ne contient qu'un ou plusieurs prénoms. Ce cas particulier (et même rare) a pour conséquence que le nom de naissance n'est pas une donnée obligatoire lorsque l'on recherche une identité au SNGI !

Comme dans tout système informatique, des contrôles ont été spécifiés pour éviter les erreurs de saisie, mais certains ne sont tout simplement pas possibles. Par exemple, on pourrait vouloir contrôler que le nom ne contient pas les termes de civilité « Madame » ou « Monsieur » ; mais comme le SNGI gère des identités étrangères, « Madame » pourrait très bien se dire « Madamé » et être un nom tout à fait valide dans certains pays. Il existe même quelques identités contenant des titres honorifiques comme « *Son Altesse Sérénissime* »…

On trouve aussi des noms très courts (1 ou 2 caractères) ou très longs (plus de 50 caractères), des dates de naissance ou de décès partielles (avec le mois ou le jour inconnu), des personnes nées dans les anciennes colonies françaises et qui veulent être reconnues comme nées en France et celles qui veulent être considérées comme nées à l'étranger[19], etc. Chaque spécificité nécessite un traitement particulier.

▶ Les enjeux de la qualité des données renforcent les exigences de contrôle

L'usage du SNGI par un nombre croissant de systèmes informatiques de la sphère sociale et son implication dans une grande variété de dispositifs (prélèvement à la source, compte personnel de formation, répertoire électoral unique, *FranceConnect*, etc.) renforce sensiblement l'exigence sur la qualité des données et des traitements.

Une erreur sur une identité (une mauvaise orthographe par exemple) peut entraîner des conséquences bien plus graves qu'un nom mal écrit sur un courrier. Un individu qui serait victime d'une telle erreur pourrait être confondu avec un autre, être à tort considéré comme mort ou encore voir ses prestations sociales attribuées à un autre. La qualité des données est aussi un élément indispensable à la lutte contre la fraude puisqu'elle permet de limiter les fausses identités et de renforcer les contrôles *via* un meilleur croisement des informations[20].

Cet enjeu fort a conduit à concevoir différents mécanismes de contrôle qui portent sur divers aspects du référentiel :

Tout d'abord, les circuits de contrôle et de certification des identités *via* les mairies, l'Insee, le Sandia et plus largement tous les organismes de la sphère sociale, contribuent par essence à éviter les erreurs sur les données d'état civil. La qualité passe aussi par des contrôles de cohérence en continu sur les flux d'alimentation du SNGI (par exemple, un individu ne peut avoir une date de décès antérieure à sa date de naissance, le code lieu de naissance ou de décès doit correspondre au libellé associé, etc.). D'autres contrôles sur les flux assurent que les mises à jour ne sont effectuées que par des personnes ou organismes habilités. *A posteriori*, des analyses statistiques et des outils de *Data mining* permettent de surveiller plus finement la qualité des données ***(encadré 6)***. Dans le cadre du RGPD[21], les individus eux-mêmes peuvent demander la rectification[22] des informations les concernant et deviennent ainsi acteurs du contrôle de la qualité des données. Enfin, à l'occasion

19 Voir l'article de Lionel Espinasse et Valérie Roux dans ce même numéro.
20 Le croisement d'informations entre organismes de la sphère sociale est autorisé et encadré par les différents textes juridiques qui régissent les systèmes d'information concernés.
21 Règlement général sur la protection des données, voir les fondements juridiques en fin d'article.
22 En revanche ils ne peuvent pas demander la suppression des informations les concernant, car ils ne peuvent pas exercer de droit d'opposition au traitement .

> ***La non-qualité des données est souvent constatée sur des identités « anciennes ».***

de croisements avec les fichiers des organismes partenaires, la détection et la correction des écarts contribuent aussi à l'amélioration continue de la qualité des informations.

Ce contrôle de la qualité a permis de repérer des causes fréquentes d'anomalie sur l'identité. Ainsi, la non-qualité des données est souvent constatée sur des identités « anciennes » : moins contrôlées en amont de leur intégration dans le référentiel, celles-ci sont issues pour beaucoup de saisies manuelles et donc avec un risque d'erreur accru. Le dispositif rencontre aussi parfois des difficultés pour récupérer des informations ou des justificatifs fiables pour corriger les erreurs. C'est notamment le cas lorsque l'assuré est très âgé ou lorsqu'il vient d'un pays instable ou ne disposant pas d'un registre d'état civil fiable. Les personnes qui viennent provisoirement travailler en France et qui quittent le territoire avant que leur identité ne soit certifiée posent encore un autre type de difficulté.

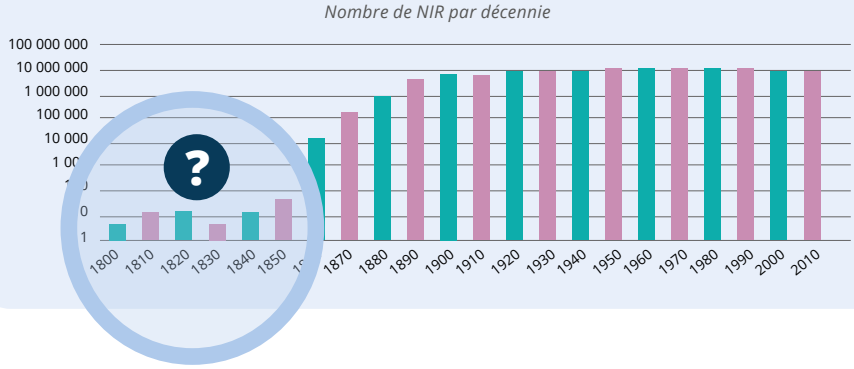
Dans ce contexte, il est nécessaire d'assurer une surveillance et un suivi permanent de la qualité des données du SNGI. Cette activité qui vient compléter les contrôles automatisés est aussi appelée « **administration du référentiel** ».

▶ Encadré 6. Un exemple d'analyse de la qualité des données

Le NIR contient une information sur l'année de naissance sur deux positions, donc sans faire figurer le siècle de naissance.

La variable « année de naissance » du SNGI est une variable calculée, à partir de la donnée « siècle de naissance » enregistrée dans le référentiel et des deux caractères constitutifs du NIR.

Une analyse réalisée en 2017 a permis de détecter de probables erreurs sur le siècle de naissance :



L'analyse de la cohérence entre année de naissance et de décès s'appuie sur le calcul de l'âge de l'individu au moment du décès. Elle a permis de détecter des situations anormales comme ces personnes censément décédées bien avant leur naissance ou à un âge au-delà de 118 ans.

Ce rôle majeur dans les démarches de qualité est assuré par de nombreux acteurs : l'Insee et le Sandia qui surveillent les flux d'immatriculations, mais également les équipes informatiques de la Cnav qui surveillent tous les flux entrants ou sortants du SNGI et qui réalisent, *via* des outils dédiés, des analyses statistiques des données.

Comme les contrôles en entrée du dispositif, les analyses *a posteriori* sont régulièrement confrontées à la diversité même des états civils dans la population. On pourrait par exemple vouloir détecter des erreurs de code sexe en s'appuyant sur le prénom d'une personne. Mais un prénom considéré comme féminin en France peut être utilisé également au masculin pour une identité étrangère : par exemple « *Rose* » qui est parfois utilisé avec la prononciation « *Rosé* ».

▶ Différents systèmes pour différents usages : une explication s'impose

On l'a vu à plusieurs reprises : le SNGI n'est pas le seul référentiel qui traite de l'identité et chaque système a ses spécificités et des objectifs qui lui sont propres. Il faut faire preuve d'une certaine expertise pour bien comprendre les nuances et être prudent quant à l'interprétation des statistiques issues de ces systèmes.

On peut par exemple constater des écarts de plusieurs millions d'individus entre le nombre de cartes Vitale émises (58 millions de cartes actives comptabilisées fin 2019), la population recensée en France par l'Insee (67,8 millions au 1er janvier 2022) et la population connue du SNGI (115 millions d'identités fin 2020, dont 86 millions « présumés vivants »). Si ces écarts peuvent paraître *a priori* étonnants, ils s'expliquent simplement et ne sont pas le reflet de dysfonctionnements : le recensement de l'Insee compte le nombre de personnes résidant en France, alors qu'une carte Vitale peut être attribuée à une personne qui a ensuite quitté le pays. Pour le SNGI, comme expliqué précédemment, la présence d'une identité même sans information de décès ne signifie pas que la personne est toujours vivante, ni qu'elle vit en France ou qu'elle bénéficie d'une prestation sociale.

Une autre source de confusion fréquente concerne la nationalité : le SNGI détient des informations sur le lieu de naissance, ce qui n'est pas nécessairement la même chose que la nationalité.

Ces exemples illustrent l'importance de comprendre le rôle de chaque référentiel, de leurs données et de leur positionnement vis-à-vis d'autres sources de données pour les comparer de manière pertinente.

Les perspectives sur une extension du SNGI sont nombreuses. Tout d'abord pour l'intégrer directement ou indirectement dans les chaînes de traitements de plus en plus d'organismes de la sphère sociale ou pour d'autres partenaires comme les collectivités territoriales. De nouveaux usages et services se profilent également notamment pour fiabiliser les données, améliorer la lutte contre la fraude, et peut-être envisager des extensions au niveau européen.

▶ Bibliographie

- CNAV, 2021. *L'Assurance retraite - Missions et chiffres clés 2020.* [en ligne]. Édition juillet 2021. [Consulté le 17 mai 2022]. Disponible à l'adresse : *https://www.lassuranceretraite.fr/portail-info/files/live/sites/pub/files/PDF/cnav-missions-chiffres-cles-2020.pdf*.
- CNIL, 2000. Le NIR, un numéro pas comme les autres. In : *20e rapport d'activité 1999.* [en ligne]. Édition 2000. La Documentation française. Chapitre 2, pp. 61-98. [Consulté le 17 mai 2022]. Disponible à l'adresse : *https://www.cnil.fr/sites/default/files/atoms/files/20171116_rapport_annuel_cnil_-_rapport_dactivite_1999_vd.pdf*.
- CNIL, 2009. RNIAM : Répertoire national interrégimes des bénéficiaires de l'assurance maladie. In : *site de la CNIL*. [en ligne]. 22 juin 2009. [Consulté le 17 mai 2022]. Disponible à l'adresse : *https://www.cnil.fr/fr/rniam-repertoire-national-interregimes-des-beneficiaires-de-lassurance-maladie-0*.
- CNIL, 2020. Tout savoir sur le décret « cadre NIR » dans le champ de la protection sociale. In : *site de la CNIL.* [en ligne]. 14 mai 2020. Commission nationale de l'Informatique et des Libertés. [Consulté le 17 mai 2022]. Disponible à l'adresse : *https://www.cnil.fr/fr/tout-savoir-sur-le-decret-cadre-nir-dans-le-champ-de-la-protection-sociale*.
- DGEFP, 2022. *Mon compte formation.* [en ligne]. Délégation Générale à l'Emploi et à la Formation Professionnelle du Ministère du Travail, de l'emploi, de la formation professionnelle et du dialogue social. [Consulté le 17 mai 2022]. Disponible à l'adresse : *https://www.moncompteformation.gouv.fr*.
- 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 17 mai 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.
- SUREAU, Christian et MERLEN, Richard, 2021. Le Répertoire de gestion des carrières unique (RGCU). Un nouveau référentiel ouvrant des perspectives pour l'analyse sociale. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 64-81. [Consulté le 17 mai 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398687/courstat-6-art-4.pdf*.

▶ 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]. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32016R0679&from=FR*.
- Article L114-12-1 du Code de la sécurité sociale. In : *site de Légifrance*. [en ligne]. 1er janvier 2019. Modifié par la loi n°2018-771 du 5 septembre 2018 et par la loi n°2014-288 du 5 mars 2014. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000037385319/*.
- Décret n° 85-1343 du 16 décembre 1985 instituant un système de transfert de données sociales. In : *site de Légifrance*. [en ligne]. Version abrogée depuis le 17 juin 2013. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000866619*.
- Décret n° 96-630 du 16 juillet 1996 relatif à l'utilisation du numéro d'inscription au Répertoire national d'identification des personnes physiques pour les traitements nominatifs concernant le contrôle des ressources des allocataires du revenu minimum d'insertion. In : *site de Légifrance*. [en ligne]. Version abrogée depuis le 26 octobre 2004. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000173690*.
- Décret n° 2018-390 du 24 octobre 2018 relatif à un traitement de données à caractère personnel dénommé « système national de gestion des identifiants ». In : *site de Légifrance*. [en ligne]. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000036940288*.
- Décret n° 2019-341 du 19 avril 2019 relatif à la mise en œuvre de traitements comportant l'usage du numéro d'inscription au répertoire national d'identification des personnes physiques ou nécessitant la consultation de ce répertoire. In : *site de Légifrance*. [en ligne]. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000038396526/*.
- Délibération (CNIL) n° 83-58 du 29 novembre 1983 portant adoption d'une recommandation concernant la consultation du Répertoire National d'Identification des Personnes Physiques (RNIPP) et l'utilisation du numéro d'inscription au répertoire (NIR). In : *site de Légifrance*. [en ligne]. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000017654414/*.
- Circulaire DSS/SD4C n° 2012-213 du 1er juin 2012 relative à l'attribution d'un numéro identifiant d'attente aux demandeurs ou aux bénéficiaires de prestations de protection sociale. In : *site du ministère des Affaires sociales et de la Santé*. [en ligne]. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://solidarites-sante.gouv.fr/fichiers/bo/2012/12-06/ste_20120006_0100_0064.pdf*.

Sirus, le répertoire d'entreprises au service du statisticien

Ali Hachid [*] *et Marie Leclair* [**]

Tout statisticien intéressé par la collecte et l'analyse de données économiques a besoin de s'appuyer sur un répertoire d'entreprises, par exemple pour constituer une base de sondage ou confronter ses données d'enquête ou administratives à des valeurs de référence. En France, depuis plus de cinquante ans, la statistique publique s'est appuyée sur le répertoire Sirene géré par l'Insee pour produire de l'information économique. Conçu pour répondre à des besoins administratifs, Sirene ne permet pas de satisfaire pleinement les attentes du statisticien d'entreprise d'autant plus qu'il n'est pas souhaitable d'en alourdir la charge de gestion.

À partir de 2008, l'Insee a choisi de développer un répertoire statistique d'entreprises : Sirus, répertoire des unités statistiques, est adossé à Sirene et répond aux exigences du statisticien : il prend en compte la notion d'entreprise au sens économique, il intègre de nouvelles unités comme les groupes de sociétés, il met en œuvre des traitements spécifiques comme la cessation statistique. L'ambition est de doter le système statistique public d'un outil partagé, servant de référence à l'ensemble des productions. L'enjeu est également de renforcer la cohérence d'ensemble des statistiques diffusées dans le domaine économique.

Any statistician interested in collecting and analysing economic data needs to rely on a business register, for example to build a sampling frame or to compare survey or administrative data with reference values. In France, for over fifty years, official statistics have relied on the SIRENE register managed by INSEE to produce economic information. Designed to meet administrative needs, SIRENE does not fully meet the expectations of business statisticians, especially as it is not desirable to increase its management burden.

Since 2008, INSEE has chosen to develop a new register: SIRUS, the statistical business register, is linked to SIRENE and meets the requirements of the statistician: it takes into account the notion of enterprise in the economic sense, it integrates new units such as groups of companies, and it implements specific processing such as statistical cessation. The ambition is to endow the official statistical system with a shared tool, used as a reference for all productions. The challenge is also to strengthen the overall consistency of business statistics.

* Responsable des répertoires statistiques Sirus et Citrus, division Infrastructures et répertoires statistiques, Insee, *ali.hachid@insee.fr*

** Cheffe du département Répertoires, infrastructures et statistiques structurelles, Insee, *marie.leclair@insee.fr*

Une question centrale pour le statisticien est de savoir si la statistique qu'il produit représente bien la population qu'il souhaite couvrir. Pour cela, il a besoin de données de référence, qui décrivent cette population couverte et auxquelles il confronte ses propres données. Les répertoires jouent ce rôle en recensant l'ensemble des unités statistiques d'intérêt.

Pour la production des statistiques d'entreprise, le répertoire inter-administratif **Sirene** a longtemps été la référence en France *(https://www.insee.fr/fr/information/1972043)*. À partir de 2008, un répertoire complémentaire dédié spécifiquement aux usages des statisticiens publics est entré dans le paysage : **Sirus** (Système d'immatriculation au répertoire des unités statistiques) a, de ce fait, un champ et des unités d'observation différents de Sirene. Référentiel pour la statistique d'entreprise, il centralise de nombreuses informations provenant de différents processus statistiques. Il offre une large palette de services, aux responsables d'enquêtes comme aux systèmes d'information utilisateurs.

Répertorier les entreprises pour produire des statistiques représentatives

Recueillir une information *via* un questionnaire adapté, redresser la non-réponse ou ventiler ses statistiques selon des critères variés ne sera pas pertinent si la population sur laquelle ces statistiques sont calculées est trop particulière pour qu'on en tire une quelconque généralité. Pour éviter cet écueil, le responsable d'enquête tire aléatoirement un échantillon dans une base de sondage. Celle-ci doit comprendre l'ensemble des unités d'intérêt pour l'enquête : ce sera, par exemple, l'ensemble des unités ayant une activité industrielle pour **l'enquête annuelle de production** (EAP) ou encore l'ensemble des entreprises créées un semestre donné pour le **système d'information sur les nouvelles entreprises** (Sine).

> ***Si chaque enquête mobilise une base de sondage(…) il faut, pour la constituer, s'appuyer sur un référentiel recensant exhaustivement les unités statistiques : un répertoire.***

Si chaque enquête mobilise une base de sondage qui est fonction de son champ d'intérêt, il faut, pour la constituer, s'appuyer sur un référentiel recensant exhaustivement les unités statistiques : un répertoire.

Pour répondre au besoin de constitution de bases de sondage, le répertoire doit toutefois aller plus loin que le seul recensement des unités statistiques ; il est important de cumuler d'autres informations pour affiner le champ des bases de sondage à élaborer. Pour l'enquête annuelle de production citée précédemment, il est nécessaire par exemple de savoir quelles entreprises ont une activité industrielle et ce même si ce n'est pas leur activité principale ; pour le système d'information sur les nouvelles entreprises, il est fondamental de connaître la date de création de l'entreprise.

Les informations ainsi centralisées dans le répertoire peuvent être nombreuses : variables permettant de définir les entreprises productives marchandes (i.e. le champ de la statistique d'entreprises), mais également le statut juridique de l'entreprise, le secteur d'activité,

la catégorie d'entreprise au sens de la loi de modernisation de l'économie[1], le secteur institutionnel, le statut marchand, productif, exploitant ou employeur, le chiffre d'affaires, les effectifs salariés, etc.

Ce recensement des unités statistiques est indispensable aux statistiques produites à partir d'enquêtes, mais pas seulement : cela est tout autant nécessaire pour produire une statistique sur les entreprises à partir de données administratives. Le champ des données administratives ne couvre que les unités sujettes à l'obligation administrative qui a généré la source, et il faut pouvoir le confronter à la population que l'on souhaite représenter : il s'agit alors de corriger, si nécessaire, les biais de couverture de la source administrative *(Rivière, 2018)*. Par exemple, toutes les entreprises n'établissent pas de liasses fiscales[2] ; c'est le cas des **microentreprises**. De ce fait, toute statistique d'entreprise s'appuyant sur ces liasses fiscales doit être redressée pour prendre en compte le chiffre d'affaires de ces microentreprises.

▶ Comment connaître exhaustivement la population des entreprises ?

Dans le cas français, le statisticien a l'opportunité de pouvoir s'appuyer sur le répertoire inter-administratif **Sirene** (Système informatique pour le répertoire des entreprises et des établissements) qui est géré par l'Insee depuis 1973 : chaque entreprise doit, selon le code du commerce, s'y immatriculer, signaler toute modification de sa situation (un changement d'adresse par exemple ou la création d'une nouvelle implantation). Une attention particulière est notamment apportée dans sa gestion pour éviter un doublon d'immatriculation. Depuis 1997, l'utilisation du numéro **Siren** est obligatoire pour toutes les déclarations administratives et les formalités des entreprises, ce qui garantit la couverture du répertoire mais également sa mise à jour et la justesse des informations. Le champ de Sirene est même plus large que celui de la statistique d'entreprise c'est-à-dire l'ensemble des entreprises ayant une **production marchande** de biens ou de services : depuis 1983, les entreprises publiques mais aussi les administrations doivent s'immatriculer au répertoire Sirene. Les associations, lorsqu'elles sont employeuses, qu'elles demandent des aides ou paient des impôts, sont également enregistrées.

Ce répertoire Sirene est donc une ressource essentielle pour construire un répertoire à des fins statistiques. La plupart des pays peuvent s'adosser à un répertoire administratif, mais celui-ci est rarement géré par l'institut national statistique comme c'est le cas en France. Cette particularité permet de bien prendre en compte les besoins statistiques : la possibilité de mener des opérations d'amélioration de la qualité sur les adresses, de s'assurer de la bonne codification de l'activité, etc. En l'absence d'un tel répertoire comme base de départ, il faut entreprendre un travail conséquent pour initialiser un répertoire statistique d'entreprises : l'appariement et la synthèse de différentes sources administratives permet par confrontation d'essayer de dépasser les limites de couverture

1 La loi LME -article 51- publiée le 4 août 2008 et le décret 2008-1354 du 18 décembre 2008 « relatif aux critères permettant de déterminer la catégorie d'appartenance d'une entreprise pour les besoins de l'*analyse statistique et économique* ». Voir les références juridiques en fin d'article.

2 Une liasse fiscale est un ensemble de déclarations fiscales remises par les professionnels (commerçants, indépendants, professions libérales, au régime réel normal) ou les sociétés soumises à l'impôt sur les sociétés.

de chaque source ; on peut également recenser les entreprises soit méthodiquement, comme un recensement de population, soit en moissonnant par exemple les sites *web* ou les annuaires privés *(Unece, 2015)*.

▶ Vers la création d'un répertoire statistique des entreprises

Grâce à sa fraîcheur et à sa couverture large, Sirene a longtemps été exploité par les statisticiens qui ont directement adossé leurs processus statistiques à ce répertoire.

Toutefois, utiliser le même répertoire pour des usages administratifs et statistiques pose plusieurs difficultés. La gestion d'un répertoire administratif n'est pas la même, en effet, que celle d'un répertoire statistique. Par exemple, il n'est pas possible de changer la valeur de la variable pour un usage statistique ; toute modification de Sirene peut en effet avoir des incidences légales pour l'entreprise et il importe d'être conforme à la déclaration de l'entreprise. Cependant, le statisticien pourra exclure une entreprise de sa base de sondage (notamment en cas de cessation d'activité) pour éviter à l'avenir toute réinterrogation inutile. Mais ceci serait bien sûr inenvisageable dans le cadre d'un répertoire administratif, compte tenu des conséquences pour l'entreprise en cas de cessation erronée.

De même, exiger de Sirene l'intégration de nouvelles variables à finalité statistique accroît sa charge de gestion et ajoute un flou sur le statut de ses variables statistiques et de leur usage. Prenons l'exemple du code d'activité principale exercée de l'entreprise (**APE**) : cette variable figure dans le répertoire Sirene mais sa finalité est avant tout statistique[3]. Pour autant, sa présence dans le répertoire Sirene en fait une information, publique, facilement mobilisable par tous, pour obtenir par exemple des aides ou des subventions, ouvrir des droits ou des obligations. De ce fait, cet usage « administratif » et ses conséquences pratiques amènent beaucoup d'entreprises à contester la codification par l'Insee de leur activité principale. Cette information demeure dans le répertoire inter-administratif Sirene, ainsi que la **catégorie d'entreprise**, autre variable statistique ; mais toutes deux montrent l'équilibre difficile à établir au quotidien lorsqu'on diffuse des variables statistiques au niveau individuel.

> ***L'Insee a voulu dissocier les usages statistiques et administratifs des répertoires.***

Du fait de ces difficultés, l'Insee a voulu dissocier les usages statistiques et administratifs des répertoires. Un répertoire statistique, Sirus, a été ainsi créé à partir de 2008 en sus du répertoire Sirene. La séparation se parachèvera avec le passage à Sirene 4 prévu en 2023, dont l'usage sera purement administratif *(Alviset, 2020)*. Toutes les variables ayant une finalité statistique seront répertoriées dans Sirus, qui deviendra le point d'entrée pour tous les besoins statistiques ; *a contrario*, Sirus ne sert qu'aux besoins statistiques. C'est ainsi un outil interne au service statistique public qui n'a pas vocation à être diffusé à l'extérieur ***(encadré 1)***.

3 « *L'attribution par* [...] *l'Insee, à des fins statistiques, d'un code caractérisant l'activité principale exercée (APE) en référence à la nomenclature d'activités ne saurait suffire à créer des droits ou des obligations en faveur ou à charge des unités concernées* » (décret n° 2007-1888). Voir références juridiques en fin d'article.

Un champ de la statistique d'entreprise différent du répertoire Sirene

Deux autres raisons ont justifié la séparation entre répertoire administratif et répertoire statistique. La première est la volonté de mieux définir et contrôler le champ de la statistique d'entreprise. Du fait de son usage administratif, Sirene est très adhérent aux besoins d'identification d'unités concernées par le recouvrement de l'impôt ou des cotisations sociales.

Encadré 1. Sirus, un large spectre de services pour les statisticiens du service statistique public

Le répertoire Sirus n'est accessible qu'au sein du service statistique public. Les différents besoins des statisticiens vis-à-vis d'un répertoire ont préalablement été recensés : Sirus offre ainsi une palette de services personnalisés, tenant compte du mode d'accès souhaité et du champ d'intérêt du client (par exemple les entreprises agricoles, les filiales de groupe, etc.). Ces différents services sont :

- **l'extraction des caractéristiques d'une liste d'unités** : à partir de leur identifiant, la liste d'unités est enrichie d'informations (chiffre d'affaires, activité économique principale, statut d'activité, catégorie juridique, etc.) en valeur courante ou à une date de valeur précise ;
- **la constitution de référentiel et de base de sondage :** pour les clients « *premium** » ayant des besoins récurrents, le champ d'intérêt du client est codé directement dans le répertoire. Par exemple, il permet de fournir la liste complète des unités agricoles actives au moins un jour pendant une année donnée, accompagnées de leurs caractéristiques en fin d'année ;
- **un service de recherche** dans le répertoire notamment en vue d'identifier des entreprises ou des établissements à partir de leur dénomination ou de leur adresse d'implantation ;
- **un service d'abonnement personnalisé** permettant à chaque application cliente de recevoir hebdomadairement l'ensemble des mises à jour au répertoire impactant les unités de leur champ d'intérêt ;
- **des *web services* inter-applicatifs** permettant à des applications ou des processus clients de dialoguer de manière interactive avec le répertoire Sirus, par exemple pour récupérer à la volée des données sur une entreprise à afficher sur une interface métier ;
- **un service de mise à jour du répertoire** permettant à chaque processus statistique de faire remonter une information sur une unité, comme une cessation observée dans le cadre de la collecte d'une enquête auprès d'une entreprise. Le répertoire Sirus joue alors le rôle de rediffuseur de l'information à l'ensemble du réseau des statisticiens d'entreprise ;
- **une interface *web*** dédiée permettant aux statisticiens d'explorer le répertoire Sirus dans toutes ses dimensions : navigation dans la structure des groupes de société, consultation de l'historique des évènements affectant une entreprise, etc. ;
- **des bases de données archivées** et mises à disposition des statisticiens 4 fois par an, offrant un instantané du répertoire, pour un usage en *self* à partir d'un logiciel de traitement statistique.

* Il s'agit des producteurs des enquêtes structurelles (Esane), du recensement de la population, de l'indice de chiffre d'affaires, du SSM Agriculture, etc.

> ***Un certain nombre d'unités légales sont immatriculées pour remplir des obligations administratives sans pour autant avoir une réelle consistance économique.***

Un certain nombre d'unités légales *(voir ci-après)* sont immatriculées pour remplir des obligations administratives sans pour autant avoir une réelle consistance économique (on peut par exemple citer les sociétés civiles de moyens pour la gestion de ressources partagées dans les professions libérales).

De ce fait, le champ utile de Sirus est plus restreint que celui de Sirene. On en exclut par exemple les associés gérants, les loueurs personnes physiques, les sociétés civiles de moyens précédemment citées, la majorité des sociétés civiles immobilières, les associations non employeuses, etc. Alors que Sirene comptabilisait en juillet 2020 plus de 14 millions d'unités légales actives, Sirus en dénombre environ 8,4 millions ***(figure 1)***.

Le champ de Sirus est cependant un peu plus large que celui de la statistique d'entreprise : Sirus recense ainsi les unités employeuses, y compris quand elles sont non marchandes, y compris publiques ou associatives, etc. Ainsi parmi les 8,4 millions d'unités légales actives dans Sirus, un peu plus d'un million correspondent à des unités hors champ de la statistique d'entreprise, unités légales non marchandes et non productives.

▶ Une unité légale n'est pas toujours une unité statistique

La seconde raison à l'origine de la création de Sirus est la définition de l'unité d'intérêt du répertoire : ce que recense Sirene n'est pas toujours ce qui intéresse le statisticien. En effet, Sirene immatricule des **unités légales** et leurs **établissements**. Par exemple, une entreprise de vente de chaussures constituera une unité légale avec plusieurs établissements (ses différents magasins) qui seront tous enregistrés dans Sirene.

> ***Ce que recense Sirene n'est pas toujours ce qui intéresse le statisticien.***

Or, cette unité légale, définie d'un point de vue juridique, n'est pas forcément la plus pertinente pour analyser la vie des entreprises, leurs décisions, leurs stratégies : avec le développement des groupes de sociétés, les unités légales détenues par d'autres peuvent perdre tout ou partie de leur autonomie de décision. Dans notre exemple précédent, imaginons que les magasins soient possédés par un groupe qui dispose également d'une centrale d'achat et d'une entité assurant la publicité de la marque : il sera plus intéressant pour comprendre les décisions économiques d'analyser ces unités légales (magasins, centrale d'achat et entreprise de publicité) toutes ensemble. Pour définir le contour de ce groupe de sociétés, on se fonde sur les liaisons financières entre les unités légales. Le groupe est une nouvelle unité statistique, mais il n'a pas de véritable existence juridique et on ne le trouvera pas de ce fait dans Sirene.

Documenter le lien entre « entreprises au sens économique » et unités légales

À cette notion de groupe (définie par les liaisons financières), les économistes et statisticiens ont ajouté la notion d'entreprise, au sens économique, définie comme « *la plus petite combinaison d'unités légales qui constitue une unité organisationnelle de production de biens et de services, jouissant d'une certaine autonomie de décision, notamment pour l'affectation de ses ressources courantes* »[4]. En reprenant toujours notre exemple, imaginons que le groupe de vente de chaussures possède également, pour des raisons financières, des salles de sport mais gérées de manière complètement autonome des magasins de chaussures. Il sera plus intéressant d'un point de vue économique de considérer au sein du groupe d'une part l'entreprise en charge des ventes de chaussures (comprenant les magasins, la centrale d'achat et l'entreprise de publicité) et d'autre part une entreprise comprenant le réseau de salles de sport ***(encadré 2)***. Cette entreprise, au sens économique, est une convention : elle n'a pas de réalité juridique, mais du point de vue de l'économiste cette convention

Figure 1 - Le répertoire Sirus en chiffres standardisés pour les élèves français

** depuis l'initialisation de la base en septembre 2011, dont 8 millions d'entreprise actives représentant 8,5 millions d'unités légales.*

4 Règlement européen 696/93 du 15 mars 1993, voir les références juridiques en fin d'article.

est la plus apte à décrire de manière pertinente les décisions de l'entreprise. D'un point de vue pratique, le contour de ces entreprises au sens économique est construit par les statisticiens à partir des groupes de société, en mobilisant des méthodes dites de profilage, manuelles ou automatiques *(Haag, 2019)*.

En France, les statistiques structurelles d'entreprise sont construites depuis l'exercice 2017 à partir de l'entreprise au sens économique, et non plus des unités légales. Cette évolution a eu pour conséquence une réévaluation de certains agrégats économiques, comme la production car les flux monétaires sans réelle consistance économique entre unités légales du même groupe sont annulés *(Chanteloup et Haag, 2019)* : ainsi la prise en compte des entreprises conduit à une « baisse » d'environ 7 % du chiffre d'affaires total (les ventes au sein de l'entreprise ne sont plus comptabilisées avec le passage à la notion d'entreprise). De plus, la répartition sectorielle est modifiée en faveur de l'industrie, car les entreprises industrielles rassemblent en leur sein des unités légales industrielles, mais aussi du commerce et des services, activités « support » de l'activité principale industrielle de l'entreprise.

> ***Sirus répertorie ainsi ces différentes unités statistiques et identifie les liens entre elles.***

Par rapport au répertoire inter-administratif Sirene qui recense les unités légales, Sirus répertorie ainsi ces différentes unités statistiques et identifie les liens entre elles : les établissements d'une unité légale, les unités légales, les entreprises au sens économique, leur appartenance à un groupe.

Afin d'en faciliter la maniabilité et la maintenance, seules les informations couramment utilisées par le statisticien sont répertoriées dans Sirus, soit les principales caractéristiques des unités (date de création, activité principale, effectifs salariés, etc.) ainsi que les liens entre ces unités. Les informations plus détaillées doivent être recherchées ailleurs : un répertoire de groupes[5] ainsi que des données sur les restructurations[6] complètent Sirus.

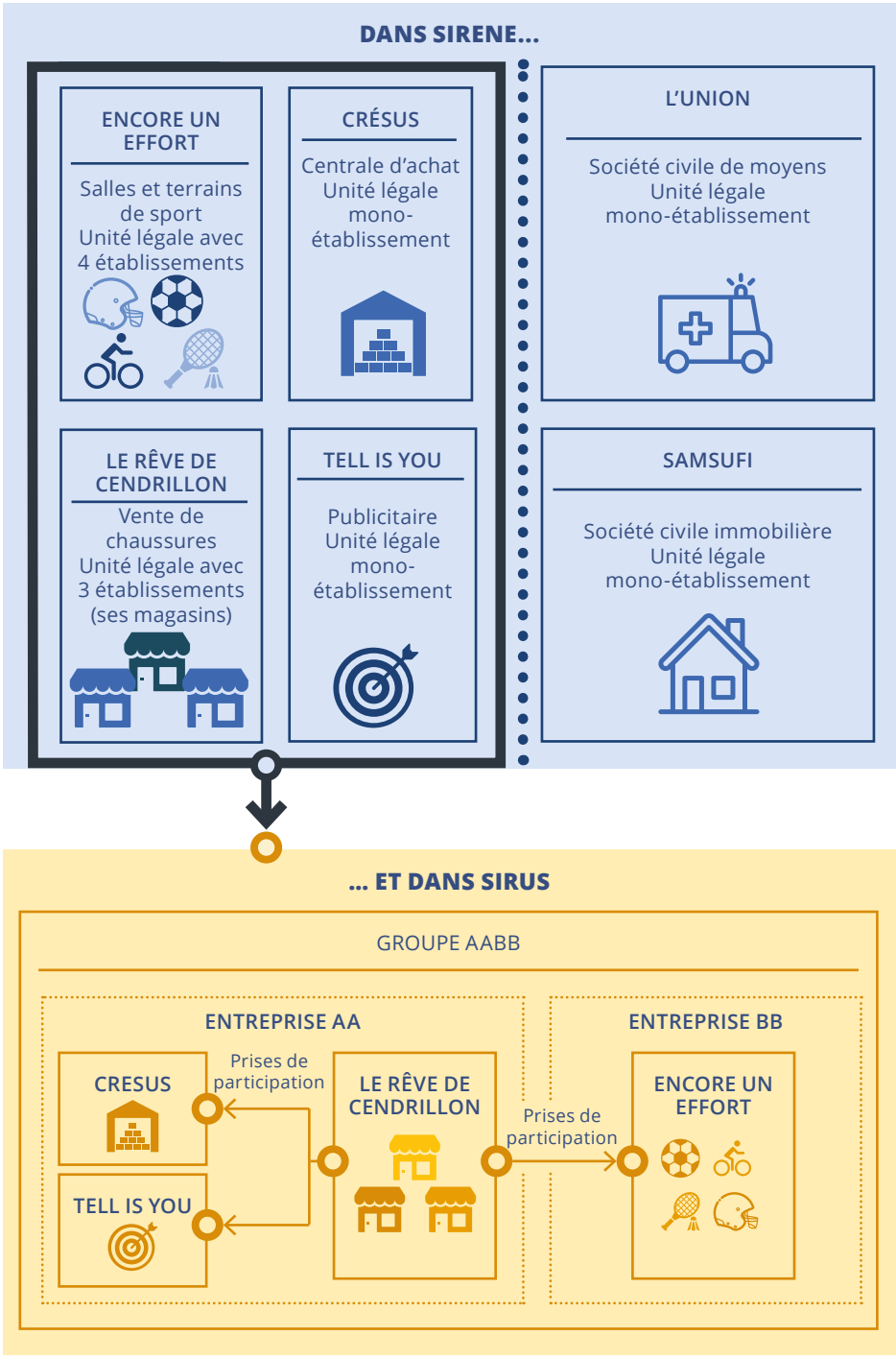
▶ Au-delà de la constitution de bases de sondage

Le recensement exhaustif des unités statistiques fait aujourd'hui de Sirus le répertoire de référence pour constituer des bases de sondage sur les entreprises. Mais la centralisation de toutes ces informations sur les entreprises permet d'autres usages de Sirus : on y calcule par exemple la notion de cessation statistique des entreprises.

Qu'une entreprise ne soit plus présente dans les sources administratives mobilisées par la mise à jour de Sirus est un indice qui laisse penser que l'entreprise a cessé son activité, de même si les données de cette entreprise n'ont pas été actualisées depuis un certain temps. Cette information est cruciale pour la précision des enquêtes qui utilisent Sirus : une telle entreprise, cessée statistiquement, ne répondra pas à une enquête. Elle ne devrait même pas donner lieu au redressement de sa non-réponse (par exemple à une extrapolation de son chiffre d'affaires à partir des entreprises répondantes). Identifier les cessations statistiques améliore la qualité des résultats d'enquête, et permet évidemment d'épargner des coûts de collecte inutiles.

5 Le répertoire Lifi (liaisons financières).
6 Le répertoire Citrus (Coordination des informations et des traitements sur les restructurations d'unités statistiques).

Encadré 2. Sirene et Sirus : des unités en moins… et des unités en plus



> ***Identifier les cessations statistiques améliore la qualité des résultats d'enquête, et permet évidemment d'épargner des coûts de collecte inutiles.***

La mise à jour des informations, en continu, dans Sirus en fait également la source privilégiée pour étudier la démographie d'entreprises : le nombre de créations, mais aussi les stocks et les cessations d'entreprises. Sirus est utilisé pour la production de ces statistiques depuis 2022[7] (Insee, 2022). Ainsi, la démographie d'entreprise est désormais calculée en concept entreprise et bénéficie des traitements de cessation statistique. De plus cela permet d'avoir un champ cohérent avec les autres productions de la statistique d'entreprise.

En France, les données d'enquêtes et les données administratives se fondent presque toujours sur le numéro Siren ; il est ainsi aisé de les apparier pour en combiner les informations. Cependant, il existe encore de rares cas où cette information est indisponible (un exemple parmi d'autres est l'enquête sur le transport routier de marchandises, qui échantillonne des véhicules), voire erronée. Dans ce cas, il faut réussir à retrouver l'identité de l'entreprise (on dit aussi « sireniser » les informations), en se fondant sur les caractéristiques connues de l'entreprise dans la source mobilisée, comme la dénomination de l'entreprise ou son adresse d'implantation. Sirus, qui regroupe en son sein de nombreuses informations sur chaque entreprise, est l'outil idéal pour les identifier et restituer un ensemble large de données les caractérisant.

Enfin, un nouvel usage a été confié à Sirus : la mesure de la charge statistique[8] des entreprises. Sirus compile pour chaque entreprise concernée, les différentes enquêtes de la statistique publique qui l'ont interrogée ainsi que le temps qu'elle a passé à répondre ; la question est en effet obligatoire à la fin de chaque enquête, depuis la commission Warsmann sur l'allègement des démarches administratives pour les entreprises (Assemblée nationale, 2011). La charge statistique de réponse aux enquêtes pour les entreprises constitue une information précieuse si on veut coordonner négativement les échantillons, c'est-à-dire choisir, quand on tire ces échantillons, avec une probabilité plus élevée les entreprises pour lesquelles la charge est la plus faible. Elle permet de répartir la charge d'enquêtes entre entreprises de manière plus équitable *(Insee, 2011)*.

▶ Sirus, au centre de nombreux processus statistiques

Pour construire Sirus, on s'adosse au répertoire Sirene, dont on complète les informations par de nombreux autres processus statistiques, qui eux-mêmes se fondent sur Sirus (pour tirer des échantillons, identifier des unités, définir leur champ, etc.). Sirus est ainsi au cœur d'un système complexe.

7 Ces séries statistiques étaient déjà produites auparavant, mais à partir de Sirene.
8 La charge statistique correspond au temps total que consacrent les entreprises à répondre aux enquêtes du service statistique public.

> ***Sirus exploite les informations provenant de différentes sources administratives pour enrichir la connaissance de ses unités.***

En effet Sirus exploite les informations provenant de différentes sources administratives pour enrichir la connaissance de ses unités : les données sur la TVA, les liasses fiscales, les données sur l'emploi, les données de l'URSSAF Caisse Nationale (UCN)[9] permettant de repérer les microentrepreneurs, les données des Douanes pour repérer les entreprises importatrices ou exportatrices ***(figure 2)***.

La centralisation de toutes ces informations permet en retour de pouvoir décrire finement les unités qui composent le répertoire et de s'assurer de leur complétude ; elle permet également de fournir un référentiel commun à l'ensemble des statistiques d'entreprise, ce qui est un enjeu pour que celles-ci, bien que produites à partir de processus différents, demeurent comparables. Se fondant toutes sur Sirus, la définition de l'activité, de la catégorie d'entreprise, de ce qui est dans ou hors du champ devrait être identique d'une enquête à l'autre.

Historisation des informations, cohérence interne : deux facteurs de complexité

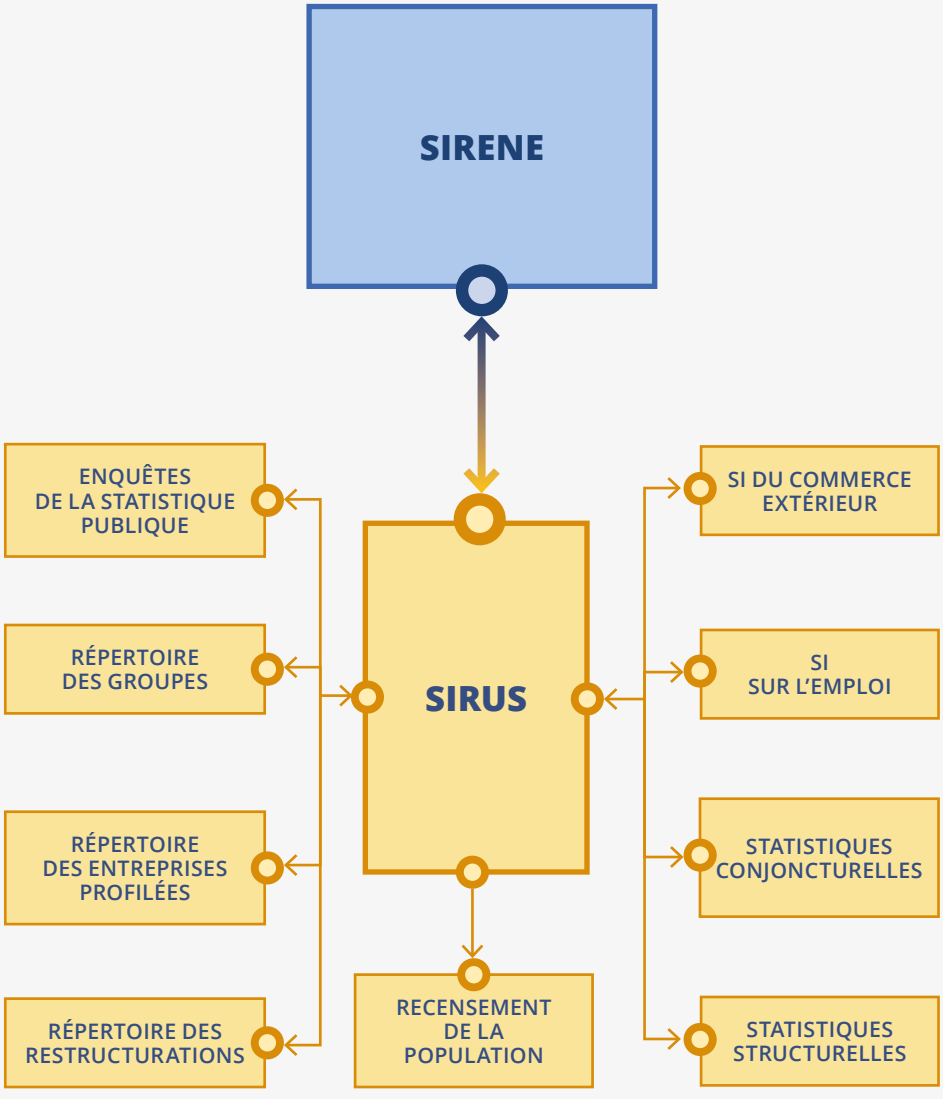
Mais derrière cette ambition, il y a bien sûr des difficultés : Sirus est mis à jour en continu par différentes sources d'information. Ses évolutions reflètent tout à la fois une réalité économique (par exemple, en décembre 2021, l'ensemble des unités actives n'est pas le même qu'un an auparavant) mais aussi la disponibilité de l'information, parfois avec retard (la cessation d'une entreprise en décembre 2021 n'est connu qu'en mars 2022). Sirus historise ainsi l'information pour prendre en compte au cours du temps les changements qui interviennent réellement sur les unités statistiques mais aussi les mises à jour des sources : les modifications ont ainsi une date de traitement (ou une date de prise en compte dans le répertoire) et une date d'effet (ou de validité dans la réalité économique). Cependant, compte tenu du nombre de sources impliquées pour la mise à jour de Sirus (de l'ordre d'une dizaine), toutes les variables et tous les changements ne peuvent pas être historisés, de manière à conférer à Sirus une certaine maniabilité.

Une autre difficulté tient également à la cohérence interne du répertoire : certaines variables, comme la catégorie d'entreprise ou encore le champ de la statistique d'entreprise, sont calculées par Sirus sur la base des différentes informations que lui apporte chaque source. Les interdépendances entre variables du répertoire Sirus peuvent être telles qu'il n'est pas toujours aisé de répercuter une correction sur l'ensemble du processus. Dès lors, une correction peut créer un dilemme entre l'incohérence potentielle qu'elle peut créer au sein du répertoire et le souci d'exactitude et de fraîcheur (avoir la dernière information d'une source de données). Et il arrive que l'on soit contraint en dernier recours de restaurer la base à partir d'une version antérieure à la prise en compte d'une information erronée.

9 Issues de l'obligation de déclarations du chiffre d'affaires aux Urssaf pour les microentrepreneurs.

Enfin, toutes ces sources d'information peuvent modifier des unités pour une même caractéristique (une enquête structurelle annuelle peut modifier le code d'activité principale exercée alors que l'entreprise aura demandé au répertoire Sirene de modifier ce code pour une raison administrative). Dès lors, il convient d'arbitrer entre ces différentes mises à jour pour savoir à laquelle Sirus donne la préséance : faut-il choisir l'information la plus récente ? Ou considérer que certaines sources sont prioritaires sur certains champs ?

Figure 2 - Sirus, au cœur des systèmes d'information statistique



> ***Pour gérer ces conflits, un ensemble de règles ont été établies visant pour chaque variable à déterminer la priorité de chaque source.***

Pour gérer ces conflits, un ensemble de règles ont été établies visant pour chaque variable à déterminer la priorité de chaque source en fonction de la date d'effet proposée de la mise à jour (Sirene ne peut pas, par exemple, mettre à jour l'activité d'une entreprise pour l'année N si cette entreprise a été enquêtée la même année par une enquête annuelle structurelle).

Les enjeux sont importants car pour être utilisé, Sirus doit apporter à ses utilisateurs une information la plus fraîche possible, la plus riche mais aussi cohérente, facilement mobilisable et traçable (l'utilisateur voudra pouvoir retrouver la base de sondage tel qu'il l'aura défini à un instant T à partir de Sirus).

▶ Des répertoires nationaux aux répertoires supranationaux

Bien que les *business registers* (répertoires d'entreprise) ne soient pas une « statistique » fournie par les instituts nationaux statistiques à Eurostat, ils font l'objet de beaucoup d'attentions dans les travaux européens et internationaux, compte tenu de leur rôle central dans la production de toutes les autres statistiques d'entreprise : manuels et groupes de travail visent à harmoniser et à améliorer la qualité de ces répertoires *(Unece, 2015 ; Eurostat, 2021)*.

Eurostat ou la commission statistique des Nations Unies incitent ainsi les instituts statistiques à réaliser des autoévaluations de la maturité de leur *business register* : la couverture de ces répertoires, en termes de **secteurs institutionnels**, mais aussi d'unités statistiques est un élément de cette évaluation, ainsi que le nombre de sources prises en compte, la fréquence de leur mise à jour, leur cohérence interne et externe, le détail géographique, les supports légaux les fondant, leur interopérabilité[10] *(Collet et Iskrenova, 2021 ; Al Kafri et Hermans, 2021)*.

> ***Si ces répertoires sont un enjeu au niveau européen et international, c'est aussi par ce que la pertinence de leur dimension nationale est remise en cause par la mondialisation.***

Si ces répertoires sont un enjeu au niveau européen et international, c'est aussi par ce que la pertinence de leur dimension nationale est remise en cause par la mondialisation *(Sturm, 2021)*. Par définition, les répertoires nationaux d'entreprise couvrent le territoire national : on trouve donc dans le répertoire français les unités légales françaises et ce que l'on appelle la « trace française » des groupes et des entreprises au sens économique. En effet, certaines entreprises sont transnationales et une partie des unités légales sur le territoire français peut appartenir à une entreprise dont le centre de décision se trouve à l'étranger. Dans ce cas, la statistique française, et les répertoires, ne suivent que la trace de l'entreprise qui se trouve sur le territoire français mais ne capte ainsi qu'une partie des choix économiques qu'opère l'entreprise.

10 Voir également l'article de Pascal Rivière dans ce même numéro.

Parce qu'il peut être plus pertinent pour comprendre et analyser les décisions de ces entreprises de se situer à un niveau supranational, un répertoire des groupes européens (*European Group Register*, EGR) est construit par les différents instituts nationaux statistiques européens sous l'égide d'Eurostat ; des tentatives similaires sont faites au niveau mondial par l'OCDE (*Analytical Database on Individual Multinationals and Affiliates*, Adima) et la commission statistique des Nations Unies (*global enterprise group register*), prenant en compte le poids considérable des entreprises multinationales dans l'économie mondiale *(Snyder, Di Matteo, Pilgrim et Ostolaza, 2021)*.

▶ Quelles évolutions à l'avenir pour Sirus ?

Le répertoire européen des groupes se construit en se fondant sur les répertoires nationaux et permet de définir les enveloppes des groupes européens. À terme, une interopérabilité sera exigée entre ces répertoires nationaux, dont Sirus, et le répertoire européen des groupes, c'est-à-dire que Sirus devra pouvoir marquer les groupes européens avec les identifiants du répertoire européen et assurer la cohérence des données des deux répertoires sur le champ des entreprises multinationales.

Ces évolutions sont prévues par le nouveau règlement européen sur les statistiques d'entreprise *Structural Business Statistics (Colin, 2019)* qui a également relevé le niveau d'exigence sur les *business registers* nationaux en termes de variables à couvrir (de manière obligatoire ou recommandée, comme le marquage des *Special Purpose Entity*[11]) et a réaffirmé le rôle central de ces répertoires, véritable colonne vertébrale de la statistique d'entreprise, pour garantir la cohérence des données dans le temps, entre les États membres et entre les différents domaines (statistiques de court terme, statistiques structurelles d'entreprises, démographie d'entreprises, statistiques du commerce selon les caractéristiques des entreprises, etc.).

Poussé par ces règlements, Sirus devrait donc s'enrichir de nouvelles variables dans les années à venir : le numéro de TVA intracommunautaire, l'identifiant LEI (Legal Entity Identifier), l'identifiant des entreprises et des groupes du répertoire européen (EGR) afin de faciliter les appariements avec d'autres sources.

Un autre enjeu important pour Sirus est la refonte du répertoire inter-administratif Sirene auquel il est adossé : à partir de janvier 2023, l'écosystème des formalités d'entreprise va être profondément modifié avec l'arrivée d'un guichet unique d'entreprises qui va remplacer tous les centres de formalité d'entreprise et amener une dématérialisation complète de ces formalités. Le répertoire Sirene va être profondément refondu pour prendre en compte ce nouveau contexte *(Alviset, 2020)*. L'enjeu pour Sirus est de continuer à produire les mêmes services alors que sa principale source sera grandement modifiée.

11 Les SPE sont des entités légales sans réelle consistance économique, créées pour donner à leur propriétaire des avantages spécifiques fournis par la juridiction d'accueil.

▶ Bibliographie

- AL KAFRI, Saleh et HERMANS, Hank, 2021. The Maturity Model for Statistical Business Registers. In : *27th Meeting of the Wiesbaden Group on Business Registers*. [en ligne]. 20-24th September 2021. Inegi, Mexico, Session n°7. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.inegi.org.mx/eventos/2021/wiesbaden/doc/7_5_PAP_HERMANS_AL_KAFRI.pdf*.
- 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 2 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497083/courstat-4-7.pdf*.
- ASSEMBLÉE NATIONALE, 2011. *Rapport n°3787 fait sur la proposition de loi n°3706 de M. Jean-Luc Warsmann relative à la simplification du droit et à l'allègement des démarches administratives*. [en ligne]. 5 octobre 2011. Tome I de M. Étienne Blanc. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.assemblee-nationale.fr/13/rapports/r3787-tI.asp*.
- CHANTELOUP, Guillaume et HAAG, Olivier, 2019. De la définition juridique à la définition économique de l'entreprise : méthode et mode d'emploi. In : *Les entreprises en France. Édition 2019*. [en ligne]. 3 décembre 2019. Insee Références, pp. 45-58. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4255789*.
- COLIN, Christel, 2019. FRIBS : un nouveau cadre commun pour les statistiques d'entreprises européennes. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 110-124. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254231/courstat-3-9.pdf*.
- COLLET, Isabelle et ISKRENOVA, Iliyana, 2021. Data-quality indicators for European statistical business registers. In : *27th Meeting of the Wiesbaden Group on Business Registers*. [en ligne]. 20-24 septembre 2021. Inegi, Mexico, Session n°7. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.inegi.org.mx/eventos/2021/wiesbaden/doc/7_2_PAP_ISKRENOVA_COLLET.pdf*.
- EUROSTAT, 2021. *European business statistics methodological manual for statistical business registers*. [en ligne]. 16 février 2021. Manuals and Guidelines. [Consulté le 2 septembre 2022]. 2021 edition. Disponible à l'adresse : *https://ec.europa.eu/eurostat/fr/web/products-manuals-and-guidelines/-/ks-gq-20-006*.
- HAAG, Olivier, 2019. Le profilage à l'Insee : une identification plus pertinente des acteurs économiques. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 86-102. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/version-html/4168409/courstat-2-9.pdf*.
- INSEE, 2022. *Rebond des créations d'entreprises en juin 2022*. [en ligne]. 19 juillet 2022. Informations Rapides n°185. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6473986*.
- RIVIÈRE, Pascal, 2018. Utiliser les déclarations administratives à des fins statistiques. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 14-24. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647013/courstat-1-5.pdf*.
- SCHUHL, Pierrette, 2018. Le Legal Entity Identifier - Contexte international et rôle de l'Insee. In : Courrier des statistiques. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 58-68. [Consulté le 14 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3646983?sommaire=3647035*.

- SNYDER, Nancy, DI MATTEO, Ilaria, PILGRIM, Graham et OSTOLAZA, Rodolfo, 2021. Future Improvements and Collaboration on the GGR and ADIMA. In : *27th Meeting of the Wiesbaden Group on Business Registers*. [en ligne]. 20-24 septembre 2021. Inegi, Mexico, Session n°6. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.inegi.org.mx/eventos/2021/wiesbaden/doc/6_4_PAP_SNYDER_OSTOLAZA.pdf*.
- STURM, Roland, 2021. Units unchanged or units unchained: How to react to globalisation from a statistical units' perspective ? In : *27th Meeting of the Wiesbaden Group on Business Registers*. [en ligne]. 20-24 septembre 2021. Inegi, Mexico, Session n°6. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://www.inegi.org.mx/eventos/2021/wiesbaden/doc/5_1_PAP_STURM.pdf*.
- UNECE, 2015. Data sources for SBR. In : *Guidelines on statistical business register.* [en ligne]. Juillet 2015. Chapitre n°6. [Consulté le 2 septembre 2022]. Disponible à l'adresse : *https://unece.org/DAM/stats/publications/2015/ECE_CES_39_WEB.pdf*.

▶ Fondements juridiques

- Règlement (CEE) n° 696/93 du Conseil, du 15 mars 1993, relatif aux unités statistiques d'observation et d'analyse du système productif dans la Communauté. In : *Journal officiel des Communautés européennes*. [en ligne]. Mis à jour le 31 décembre 2007. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:31993R0696*.
- Loi n° 2008-776 du 4 août 2008 de modernisation de l'économie (1). In : *site de Légifrance*. [en ligne]. Mise à jour le 26 août 2021. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000019283050/*.
- Décret n° 2007-1888 du 26 décembre 2007, portant approbation des nomenclatures d'activités et de produits françaises. In : *site de Légifrance*. [en ligne]. Mis à jour le 31 décembre 2007. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000017765090/*.
- Décret n° 2008-1354 du 18 décembre 2008 relatif aux critères permettant de déterminer la catégorie d'appartenance d'une entreprise pour les besoins de l'analyse statistique et économique. In : *site de Légifrance*. [en ligne]. [Consulté le 17 octobre 2022]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000019961059/*.

La base permanente des équipements (BPE)

Une source statistique singulière et constamment en mouvement

Xavier Helfenstein *

Singularité française, la statistique publique produit chaque année une base permanente des équipements – la BPE – qui recense et géolocalise les équipements, services et infrastructures accessibles à la population sur l'ensemble de la France.

Son intérêt principal est de mettre à disposition de tous un outil simple qui fournit des informations sur le niveau d'équipement, services et infrastructures d'un territoire (présence d'équipements scolaires, médicaux ou encore sportifs, etc.).

Ingrédients clefs de nombreux diagnostics ou analyses de territoire, ses produits de diffusion sont multiples et largement utilisés à l'intérieur comme en dehors de la statistique publique. Héritière de l'inventaire communal, la constitution de cette base s'appuie aujourd'hui sur la collecte de sources, principalement administratives. Bien moins onéreuse à produire et à mettre à jour, elle permet à l'Insee d'entretenir des relations fructueuses tant avec ses utilisateurs que ses fournisseurs de données. L'objectif désormais est de mieux prendre en compte encore les demandes des utilisateurs et de faciliter les échanges à la fois avec les producteurs de données et les utilisateurs, en gérant un système complet de métadonnées.

A unique feature of French official statistics is the annual production of a permanent equipment database – the BPE – which lists and geo-locates the equipments, services and infrastructures accessible to the population throughout France.

Its main interest is to make available to everyone a simple tool that provides information on the level of equipment, services and infrastructure of a territory (presence of school, medical or sports facilities, etc.).

As a key ingredient in many territorial analyses, its dissemination products are numerous and widely used both inside and outside official statistics.Heir to the "communal inventory" survey, the constitution of this database is now based on a collection of sources, mainly administrative. Cheaper to produce and update, it enables INSEE to maintain fruitful relations with both its users and its data suppliers.

The main objective now is to rationalise the work, to take better account of users' requests and to facilitate exchanges with both data producers and users by managing a complete metadata system.

* Chef du pôle Base permanente des équipements, Insee Nouvelle Aquitaine, *xavier.helfenstein@insee.fr*

Je viens d'emménager, quels sont les équipements disponibles dans mon village ? Y a-t-il un lycée à proximité de la commune où je suis muté ? Aurais-je accès facilement à un centre de dialyse autour de chez moi ? Y a-t-il un terrain de tennis pas trop éloigné ? Tout un chacun peut être amené à se poser ces questions et bien d'autres sur les équipements accessibles et leur localisation.

Les autorités publiques, qu'elles soient nationales ou locales, s'intéressent elles aussi tout naturellement à l'accès de la population aux équipements : quelles sont les zones rurales dans lesquelles les équipements de la vie courante sont difficilement accessibles ? Le vieillissement de la population accentue-t-il l'inégalité d'accès aux équipements ? L'offre de places en crèche est-elle suffisante ? L'équilibre entre habitat et équipement est-il assuré ? *(Hautdidier et Kuentz, 2011)*

L'intérêt de **la base permanente des équipements (BPE)** est de mettre à disposition un outil simple permettant de répondre à ces différentes questions. Plus précisément, la BPE fournit des informations quantitatives sur le niveau d'équipement d'un territoire. Ce faisant, elle permet de produire les données souhaitées, comme la présence ou l'absence d'un équipement, la densité d'un type d'équipement, etc. *(Barbier, Toutin, Levy, 2016)*, et plus globalement d'analyser la structuration du territoire et notamment d'identifier des zones sous-équipées.

L'Insee utilise d'ailleurs la BPE pour déterminer le zonage des bassins de vie. Pour connaître les zones d'attraction des équipements en milieu rural, des distances entre les communes et les équipements peuvent être calculées. Les données géolocalisées de la BPE conduisent à des analyses territoriales à un niveau très fin. Elles sont souvent associées à d'autres sources, comme le recensement de la population et le dispositif sur les revenus localisés sociaux et fiscaux (Filosofi) pour s'enrichir d'informations relatives à l'offre d'équipements, à l'attractivité des territoires, etc.

▶ La BPE, une exception française

> ***Au niveau européen, la BPE fait figure d'exception.***

Au niveau européen, la BPE fait figure d'exception. Si certains instituts nationaux de statistique mettent à disposition des données géolocalisées sur les équipements, il s'agit le plus souvent de fichiers organisés par thématique[1]. Quand la Commission européenne[2] cherche à mesurer l'accessibilité des services, elle ne dispose pas de données fines comparables sur l'ensemble des États membres[3], car la notion d'équipement n'est pas tout à fait cohérente avec le concept qui supporte les implantations des unités légales des répertoires d'entreprises. Elle n'est pas non plus totalement cohérente avec celui des fichiers de bâtis. Dans cet entre-deux, la France a choisi d'insérer un système d'information spécifique, la BPE : il faut sûrement y voir la spécificité française de s'intéresser de longue date aux disparités territoriales.

1 Voir par exemple les fichiers de l'ONS suisse sur la localisation des commerces, sur le site *www.data.europa.eu.*
2 En l'occurrence la direction générale de la Politique régionale et urbaine appelée DG Regio.
3 Voir l'introduction de l'article de *(Kompil et alii, 2018)* dont les auteurs utilisent par ailleurs des données privées pour valider leurs hypothèses et leur modèle.

▶ De quels équipements parle-t-on ?

> ***Pour la BPE, un équipement est un service accessible à la population, qu'il soit marchand ou non.***

Définir un équipement n'est pas chose facile : pour la BPE, un équipement est un service accessible à la population, qu'il soit marchand ou non. Le terme de « service » est à prendre dans une acception très large, car il recouvre une grande variété de cas. Il peut s'agir d'un aménagement, tel qu'un jardin remarquable ouvert au public, un lieu de baignade aménagé, une boucle de randonnée. L'équipement peut également être une infrastructure : un gymnase, une piscine, une gare, un établissement de transfusion sanguine ou encore un centre d'accueil de demandeurs d'asile. Les établissements ouverts au public et offrant des services, comme les commerces, les banques, les tribunaux, les écoles, etc., font aussi partie du champ. Plus exactement, la notion d'équipement se réfère au service rendu au sein d'un établissement et non à l'établissement lui-même. Ce sont ainsi la maternité et le service des urgences qui sont répertoriés et non le centre hospitalier auquel ces services appartiennent. Toutefois, dans la majorité des cas, le service et l'établissement se confondent.

La base permanente des équipements (BPE) répertorie ces équipements au 1er janvier de chaque année. Ils sont classés par type, au nombre de 188 en 2021 ***(encadré 1)*** et regroupés en sept domaines : les services aux particuliers, les commerces, l'enseignement, la santé et le social, les transports et déplacements, les sports-loisirs et la culture, le tourisme.

Comme souvent en production statistique, le dispositif actuel est le fruit d'une lente évolution sur une longue période.

▶ Aux origines de la BPE : l'inventaire communal

Historiquement, la BPE succède à l'inventaire communal qui avait lieu tous les 8 à 10 ans. Le dernier inventaire a été réalisé en 1998. Il s'agissait d'une enquête auprès des communes de moins de 30 000 habitants pour la France entière, y compris les DOM, à l'exception de certains départements d'Île-de-France. Une commission communale réalisait l'inventaire des équipements, commerces et services présents et, pour ceux qui étaient absents, les services de remplacement existants (tournée, permanence, rayon dans un commerce multi-services, service identique dans une commune proche, etc.).

Ce dispositif permettait ainsi de disposer de résultats communaux sur la densité d'implantation, la fréquentation des équipements et l'attractivité des communes. Cet outil avait l'avantage d'être une source homogène et facilement mobilisable. Les utilisations locales ou régionales de l'inventaire communal consistaient à décrire le degré d'équipement d'une commune, à le comparer à des communes proches ou de même importance ou encore à dresser la carte d'implantation d'un type d'équipement sur une zone. Les inventaires communaux étaient ainsi mobilisés lors de la mise au point des schémas de services publics en milieu rural, à la demande des collectivités locales ou des régions, ou pour définir des zonages tels les bassins de vie et les pays[4]. Les données disponibles permettaient

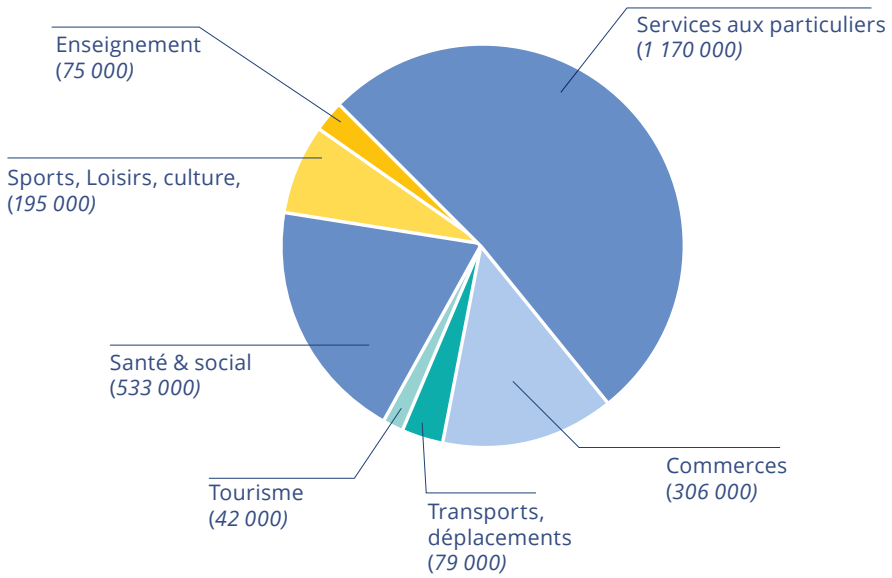
4 Le « pays » est une catégorie administrative datant de 1995 ; ce territoire présente une « *cohésion géographique, économique, culturelle ou sociale, à l'échelle d'un bassin de vie ou d'emploi* » afin d'exprimer « *la communauté d'intérêts économiques, culturels et sociaux de ses membres* » et de permettre l'étude et la réalisation de projets de développement.

également de repérer des zones rurales en difficulté et de déterminer des pôles à conforter. Certains services des collectivités territoriales et de l'État ont largement utilisé ces référentiels à l'occasion. La même source répondait à de nombreuses demandes d'entreprises pour éclairer leurs études de marché.

L'inventaire communal était cependant contesté par certains acteurs, notamment en raison d'une qualité des réponses à l'enquête jugée insuffisante et de l'absence de procédures homogènes de validation. Les résultats, produits tous les dix ans avec des coûts de collecte élevés, étaient en outre rapidement obsolètes. Enfin, l'enquête ne permettait pas de mesurer les conditions d'accessibilité de la population aux équipements, en termes d'horaires d'ouverture ou de desserte par un transport en commun.

En 1994, une enquête « équipements urbains » a été réalisée sur le champ des communes de plus de 30 000 habitants. Elle avait pour but de suppléer aux lacunes de l'inventaire communal en milieu urbain. L'absence de localisation à l'adresse, la constitution de zones fixes et un processus de diffusion insuffisant ont nui à la publicité de l'enquête et à son utilisation.

▶ Encadré 1. Une base de 2,4 millions d'équipements*



Les équipements répertoriés relèvent de 7 domaines, qui ont été subdivisés en 27 sous-domaines et 188 catégories. 85 % des équipements sont géolocalisés avec une précision de moins de 100 m, seuls 0,4 % d'entre eux ne sont pas géolocalisés.

** millésime 2021*

L'introduction des sources administratives : des gains multiples

L'Insee décide donc au début des années deux-mille de lancer le projet de production de la base permanente des équipements, reposant sur des sources administratives en lieu et place d'enquêtes. Ces données proviennent soit du service statistique public, soit d'organismes effectuant des missions de service public (La Poste, ANPE puis Pôle emploi, etc.) qui les transmettent dans le cadre de conventions établies avec l'Insee.

Les gains apportés par ce choix sont multiples et visibles dès la première édition. En utilisant au maximum les sources administratives, les coûts sont bien inférieurs à ceux que nécessitaient les inventaires communaux. L'éventail des catégories d'équipements recensés s'accroît nettement : l'inventaire communal de 1998 en comptait une trentaine, elles sont au nombre de 137 dès la BPE 2008 et de 188 dans la BPE 2021. La périodicité des résultats est annuelle, elle était quasi décennale auparavant. Les données millésimées au 1er janvier de chaque année sont mises à disposition au bout de 18 mois. Les données géolocalisées depuis 2013 ouvrent des perspectives d'utilisation bien supérieures, surtout en les combinant avec d'autres sources, elles aussi géolocalisées (Filosofi, Flores, etc.). Sous l'impulsion donnée par la loi pour une République numérique de 2016, le nombre de nouvelles sources mises à disposition s'accroît, laissant en germe la possibilité pour la BPE d'accroître encore le champ des équipements couverts.

> ***En utilisant au maximum les sources administratives, les coûts sont bien inférieurs à ceux que nécessitaient les inventaires communaux.***

Un autre gain, non des moindres, est celui de la capitalisation de l'expérience acquise au fil des ans. La pérennité du savoir-faire et l'amélioration de la qualité ne peuvent qu'être favorisées par la régularité des travaux et des relations avec les différents partenaires. Des conventions sont ainsi établies formalisant un cadre, définissant la nature de la collaboration, le contenu des données échangées, le calendrier, etc.

Mais comment procède-t-on au juste pour produire annuellement une base d'équipements aussi vaste ?

Figure 1 - Des fournisseurs de données variés et nombreux

Le dispositif mobilise 6 répertoires, 17 fichiers administratifs, 3 sources en open data, et s'appuie sur des conventions avec 11 fournisseurs.

▶ Un processus simple… en apparence

Chaque année, le processus de production commence par la collecte des données administratives auprès des partenaires ***(figure 1)***. Trois catégories peuvent être distinguées :

- en premier lieu, on utilise des sources internes à l'Insee. C'est le cas de Sirus qui est une base statistique découlant de Sirene[5]. La base de sondage de l'enquête de fréquentation dans les hébergements touristiques permet par ailleurs d'obtenir des données à l'adresse sur l'ensemble des hôtels et des campings de France ;
- la BPE se fournit également auprès de partenaires externes. Ce sont majoritairement des services statistiques ministériels. Ils couvrent des domaines variés : la santé, le sport, la culture, l'agriculture, la justice, l'éducation nationale, l'intérieur, le développement durable et les collectivités locales. D'autres acteurs institutionnels, comme Pôle emploi, La Poste et l'Agence nationale pour la cohésion des territoires (ANCT), transmettent annuellement des données permettant d'élargir la couverture de la BPE ;
- enfin, certaines données sont collectées en *open data*, telles celles sur les stations-service ou sur les établissements d'accueil auprès de jeunes enfants. Avant d'être utilisées, ces sources sont soumises à expertises amenant souvent à des contacts auprès des producteurs.

Au total, la BPE est alimentée par 6 répertoires et 17 fichiers administratifs.

▶ Mais le diable se cache dans les détails

La qualité et la nature des informations des différentes sources transmises sont cependant très hétérogènes. Par exemple, certaines sources ne contiennent pas d'informations permettant d'identifier sans ambiguïté la commune de localisation du service[6]. Les adresses des équipements sont par ailleurs présentées de manière très différente d'une source à une autre. Si certaines décomposent chaque élément d'adressage (numéro de voie, indice de répétition, type de voie, libellé de voie, code de la commune, libellé de commune), d'autres les regroupent dans une même variable, en y ajoutant même les noms ou raisons sociales des équipements ainsi que les compléments d'adresse, les boîtes postales et les cedex. Or, ce sont précisément ces éléments d'adressage qui prédéterminent la qualité de la géolocalisation.

> ***« L'architecture informatique comporte un ensemble de traitements mutualisés à toutes les sources. »***

Il est donc nécessaire d'harmoniser les données en entrée, par des traitements automatiques effectués au fur et à mesure de leur réception. L'architecture informatique comporte un ensemble de traitements mutualisés à toutes les sources : la création de variables d'adressage, la génération de codes communes ou l'attribution du millésime souhaité du code officiel géographique, la suppression des doublons, etc. Des traitements spécifiques les complètent. Ainsi, certaines sources

5 Pour illustrer la différence entre Sirene, répertoire inter-administratif des entreprises et Sirus, répertoire statistique, voir l'article d'Ali Hachid et Marie Leclair dans ce même numéro.

6 En particulier, il arrive que la source ne contienne pas le code de la commune issu du code officiel géographique, mais seulement le code postal : or le code postal peut désigner plusieurs communes différentes.

contiennent des coordonnées GPS ; elles sont alors projetées dans les systèmes adéquats[7] (Lapaine, Miljenko et Usery, 2014). De même, des caractéristiques complémentaires sont fournies pour les types d'équipements des domaines de l'enseignement et du sport, loisir et culture : présence de cantine, d'internat, équipement sportif couvert, éclairé, nombre d'aires de pratique, etc.

Une fois les sources harmonisées, elles sont appariées avec les sources de l'année précédente selon différentes clés de fusion qui permettent de repérer les équipements déjà existants, les nouveaux, ceux ayant disparu et ceux pour lesquels l'appariement est présumé. En effet, parmi les non appariés, certains présentent des éléments d'adressage et d'activité proches d'équipements présents l'année précédente. Des gestionnaires peuvent ensuite confirmer ou infirmer manuellement ces identifications présumées. Ces appariements permettent d'attribuer à chaque équipement des identifiants spécifiques à la BPE, ou de récupérer ceux de l'année précédente, et préparent ainsi la phase suivante des contrôles.

▶ Un contrôle des données nécessaire

Qu'une boulangerie cesse son activité, cela n'est pas étonnant en soi. La fermeture d'un service des urgences ou la disparition d'un aéroport interroge davantage. Ces cas sont plus fréquents qu'on ne pourrait le croire *a priori.* Dans la gestion interne des sources mobilisées par la BPE, des règles particulières peuvent induire des enregistrements tardifs. Par exemple, s'agissant des aéroports, le fournisseur de données ne peut transmettre que la liste de ceux ayant eu plus de 1 000 passagers au départ ou à l'arrivée (hors transit) au cours de l'année précédente : quelques petits aéroports ne franchissent pas cette barre certaines années. Il devient alors nécessaire d'identifier les **équipements structurants,** c'est-à-dire ceux qui sont assimilés à des équipements *a priori* pérennes, comme les hypermarchés, les lycées, les établissements de santé, les ports de plaisance, les zones de mouillage, les aéroports, etc. Quelle qu'en soit la cause, l'apparition ou la disparition d'une année à l'autre d'un équipement structurant dans une commune interpelle au point que cela justifie d'être contrôlé par un gestionnaire.

En complément et jusqu'en 2019, une opération de mesure de la qualité sur le terrain était réalisée chaque année. Elle bénéficiait de la collaboration des superviseurs du recensement de la population. Dans les communes de moins de 10 000 habitants, ceux-ci demandaient aux coordonnateurs communaux du recensement de valider une liste d'équipements, de la modifier et de la compléter le cas échéant. Des taux d'excédent et de déficit entre la base et le terrain étaient ainsi calculés, donnant une idée de la qualité de la couverture de la BPE sur les types d'équipements contrôlés. Cette opération est actuellement en suspens dans l'attente de sa refonte qui permettra d'accroître son efficience et d'améliorer la représentativité des résultats.

7 Le RGF93 pour la métropole, l'UTM40 Sud pour La Réunion, l'UTM20 Nord pour la Martinique et la Guadeloupe, et l'UTM22 Nord pour la Guyane.

> ***Chaque année, une base en évolution est diffusée.***

Afin de répondre aux besoins des utilisateurs de la BPE, chaque année, une base en évolution est diffusée. Elle porte sur deux années espacées d'un pas quinquennal et s'effectue en géographie de l'année la plus récente[8]. Deux millésimes de la BPE ne sont pas immédiatement comparables : estimer dans les comparaisons temporelles ce qui relève d'un changement de nomenclature, d'un changement de source, d'un changement de qualité dans la source ou enfin d'une évolution réelle des services accessibles sur le terrain s'avère particulièrement complexe. Des expertises sont menées pour déterminer quels types d'équipement peuvent être inclus dans le champ de diffusion de la BPE en évolution. La BPE en évolution 2016-2021 en géographie 2021 ; elle porte sur 119 types d'équipements (sur les 188 que contient la BPE 2021).

Des coordonnées géographiques complètent les informations collectées

Après la phase des contrôles en bureau, vient celle de la géolocalisation des équipements. Elle est réalisée en deux parties, par une géolocalisation automatique d'abord, puis par des reprises manuelles. Celles-ci sont effectuées pour des équipements dont la qualité de géolocalisation est jugée insuffisante[9]. Au préalable, on réutilise le résultat des recherches manuelles des années passées pour réduire le volume des recherches. Des zonages sont aussi ajoutés lors de cette phase : les arrondissements, les zones d'emploi, les bassins de vie, etc.

Une fois ces opérations terminées, démarre l'aval du processus de production : la préparation des produits de diffusion. Les tâches sont nombreuses et sont coordonnées avec les différents acteurs impliqués au sein de l'Insee, qui parfois les mettent à disposition en l'état, parfois les valorisent en les transformant en d'autres produits (des indicateurs cartographiques, des tableaux de dénombrement, etc.). La mise à jour des métadonnées associées à la source BPE fait partie intégrante de cette phase. Sa volumétrie et les multiples formes qu'elle revêt nécessitent une attention particulière (voir *infra*).

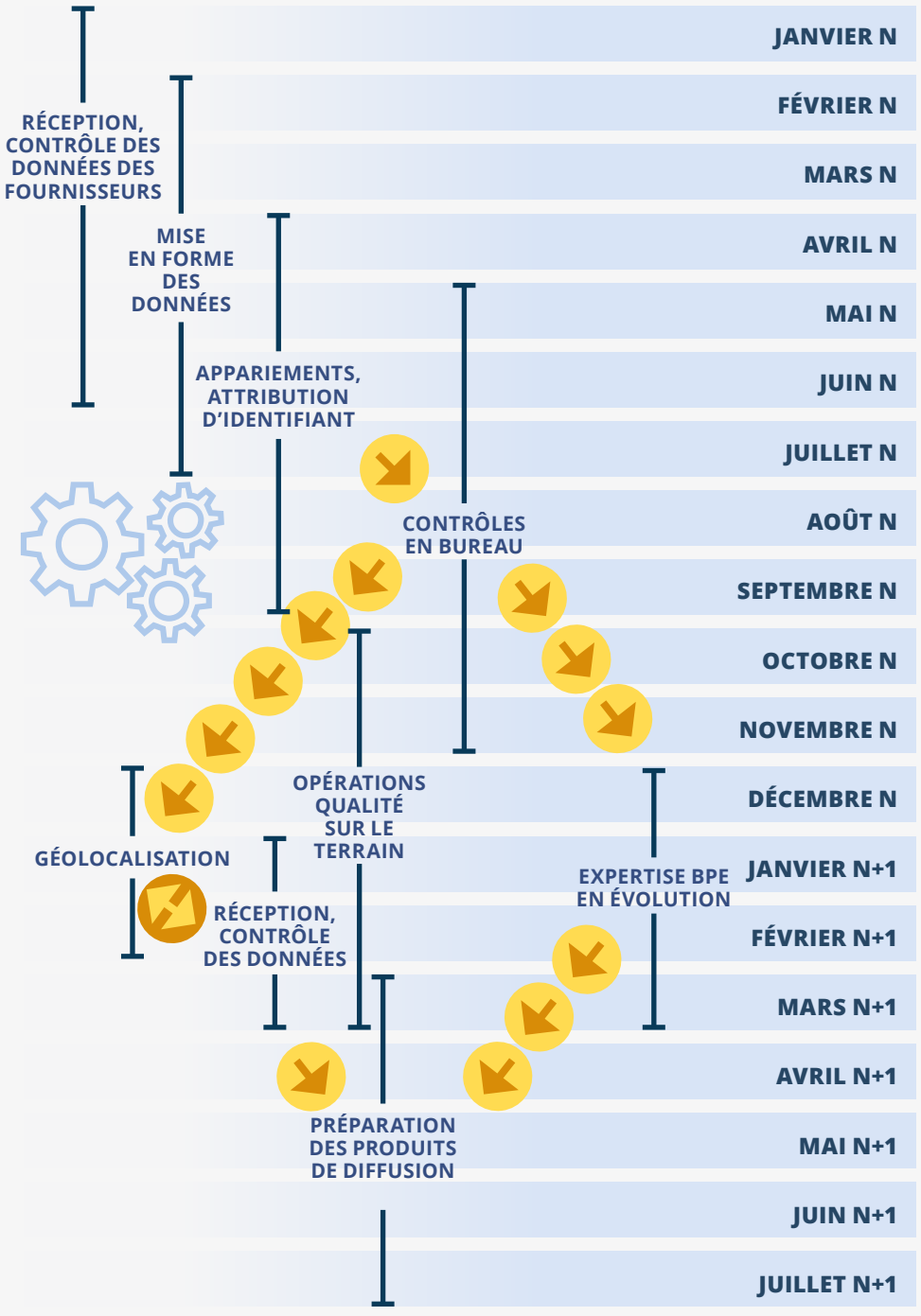
> ***Au total, le processus de production d'un millésime de la BPE dure 18 mois.***

Au total, le processus de production d'un millésime de la BPE dure 18 mois ***(figure 2)***. Les différents produits d'un millésime dont les données sont référencées au 1er janvier de l'année n sont diffusés au plus tard au début du mois de juillet de l'année n+1. Cette production permet de délivrer une offre de services diversifiée à différents utilisateurs ***(figure 3)***.

8 En cinq ans, le périmètre des communes peut évoluer, par suite de fusions de communes, de scission de portions de territoires, etc. Il importe donc d'avoir des « géographies » parfaitement comparables.

9 Les choix des types d'équipements faisant l'objet de reprises manuelles sont décidés en concertation entre le pôle BPE, la division Statistiques et analyses urbaines et le groupe Production localisée géographique (PLG).

Figure 2 - Une production cadencée sur 18 mois



▶ Une offre qui s'adapte aux utilisateurs

Trois types de publics sont identifiés pour trois niveaux de diffusion[10] :

- le grand public (incluant notamment les particuliers, les journalistes et les entreprises) ;
- les acteurs publics nationaux et régionaux[11] ;
- et les fournisseurs des données qui alimentent la BPE.

Les fichiers tous publics sont mis en ligne sur le site de l'Insee avec des fichiers d'ensemble et des bases de données thématiques. Il s'agit de dénombrements par type d'équipement, allant jusqu'au niveau de l'Iris[12], sur l'ensemble du territoire y compris les cinq départements d'outre-mer. Des fichiers détail sont également mis en ligne contenant les équipements avec leurs coordonnées géographiques et, le cas échéant, des caractéristiques complémentaires sur les domaines de l'enseignement (présence d'une cantine, d'un internat, secteur public ou privé, etc.) et des sports-loisirs (équipement couvert, éclairé, nombre de salles par structure culturelle, etc.). La mise en ligne d'un millésime annule et remplace le millésime précédent pour éviter des comparaisons temporelles inappropriées. Par ailleurs, à l'instar d'autres données, les données géolocalisées de la BPE sont utilisables sur l'espace « statistiques locales » pour concevoir des cartes à façon.

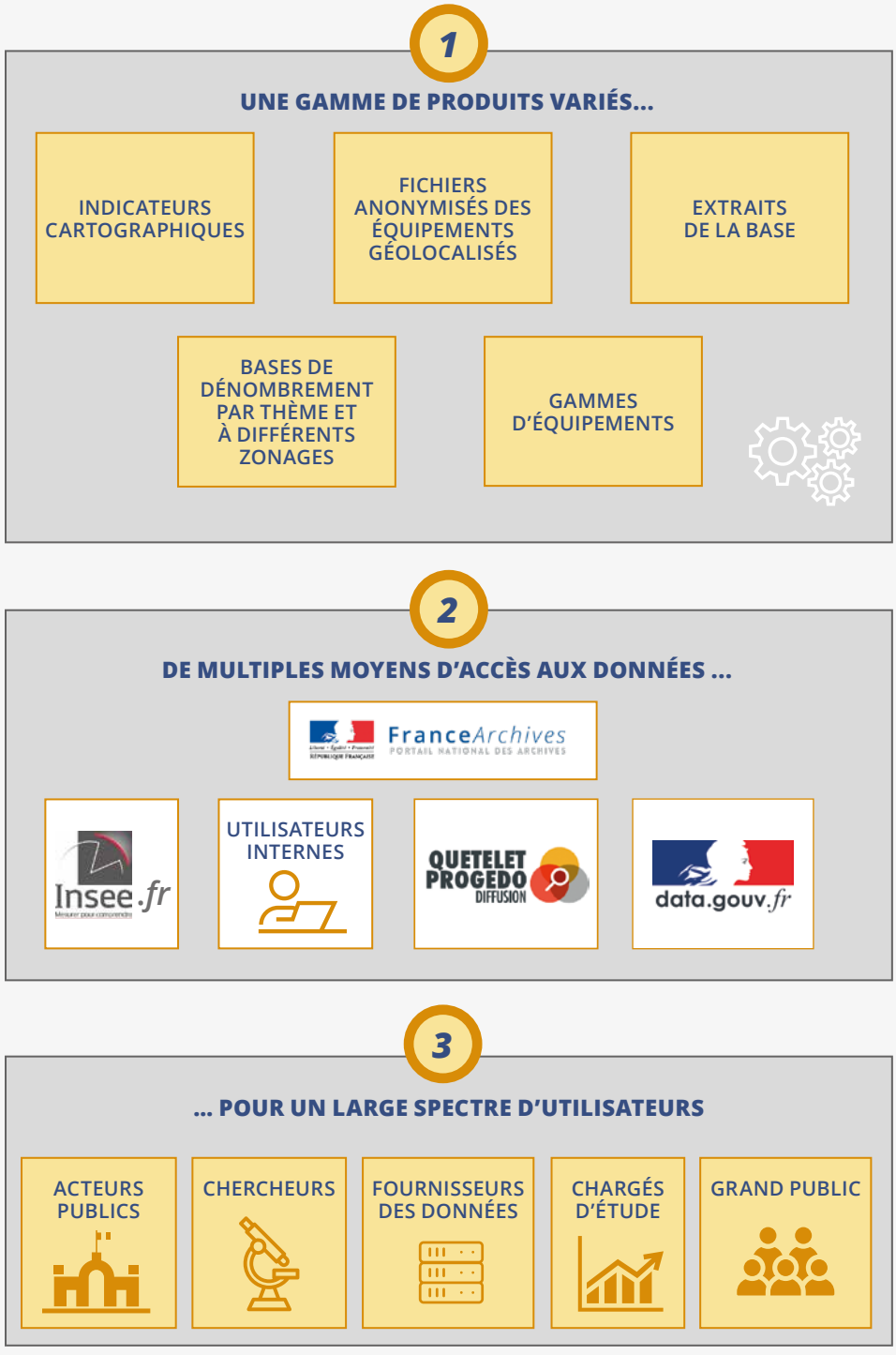
Certaines demandes sont adressées directement à l'Insee, émanant de journalistes, d'administrations, d'entreprises ou encore de particuliers et portant sur des sujets très divers : des précisions de concepts, des mises à disposition de données complémentaires, des données en évolution, etc. Les sites Insee Contact prennent en charge les réponses et font appel au pôle BPE le cas échéant. Toute structure chargée d'une mission de service public peut demander un extrait de la base sur sa zone de compétence et comportant des informations supplémentaires. Sous réserve de la signature d'une licence d'usage, cet extrait est diffusé gratuitement en application de l'arrêté du 20 juillet 2012 relatif aux modalités de diffusion de la BPE. Outre la dénomination et la localisation de l'équipement à l'adresse, sont fournies des variables de gestion (par exemple la date de création de l'équipement ou sa date de dernière mise à jour) et les variables spécifiques sur les champs de l'enseignement et des sports-loisirs (voir *supra*).

Enfin, la base est fournie chaque année gratuitement aux contributeurs, c'est-à-dire les fournisseurs alimentant la BPE, dans le cadre des conventions. Ceux-ci disposent ainsi de l'intégralité des données de la BPE, y compris les noms/raisons sociales des équipements et leur adresse.

Les produits mis à disposition sont majoritairement des produits intermédiaires adressés aux services de l'Insee chargés de les transformer en produits finaux. Ils sont formatés pour intégrer un entrepôt de données locales, dans lequel les services et les pôles d'études puisent pour réaliser leurs analyses ou actualiser leurs produits, le plus souvent en complémentarité avec d'autres sources. Ils sont transformés en indicateurs cartographiques permettant de réaliser des cartes à façon.

10 Voir l'arrêté du 20 juillet 2012 modifiant l'arrêté du 11 janvier 2008 portant création d'une BPE.
11 Les personnes de droit public et les personnes de droit privé chargées d'une mission de service public (agence d'urbanisme, agence régionale de la santé, conseil départemental, Communauté d'agglomération, etc.).
12 Îlots regroupés pour l'information statistique : quartier d'environ 2 000 habitants pour les communes de 5 000 habitants et plus, l'Iris constitue la brique de base en matière de diffusion de données infra-communales.

Figure 3 - Que peut-on faire avec les produits de la BPE ?



Ils deviennent des tableaux de comptage ou d'indicateurs de présence/absence à différents niveaux géographiques. Les produits finaux élaborés sont les bases fournies aux acteurs publics régionaux, aux fournisseurs et aux chercheurs[13] et celles transmises pour archivage. Ces bases sont aussi accessibles en *open data* sur le site *www.data.gouv.fr*.

La BPE utilisée pour enrichir les analyses territoriales

Favoriser l'accès des populations aux équipements et services publics est une problématique récurrente de nombreuses politiques publiques territoriales. Les plans d'action associés s'appuient sur un diagnostic territorial partagé à l'échelle du territoire. Un bilan de l'offre existante sur le territoire est alors réalisé, pour mettre en exergue d'éventuels déficits territoriaux et les rapporter aux populations concernées. La BPE est la source incontournable pour ce type de diagnostic. Elle permet ainsi de présenter les disparités locales en termes de taux d'équipement mais surtout en termes de temps d'accès de la population à ces équipements. Elle est pour cela couplée avec l'utilisation du distancier Metric[14] développé par l'Insee. Les territoires éloignés des équipements sont ainsi repérés et la population concernée est estimée.

Par exemple, la BPE a été utilisée à la suite de la mise en place fin 2020 des conférences régionales du sport chargées « *d'instaurer une stratégie de développement du sport à l'échelle de la région à travers un projet sportif territorial* ». Ont été mis à disposition des acteurs locaux concernés des éléments tels que l'offre d'équipements sportifs, avec des comparaisons appropriées avec d'autres équipements, des cartes par établissement public de coopération intercommunale (taux d'équipement et accès), des éléments sur les disparités d'accès en milieu urbain et en zone rurale ou encore sur les temps d'accès des établissements scolaires aux équipements sportifs, et notamment à la piscine dans le cadre du plan « aisance aquatique ». Des dossiers thématiques et des analyses ont été livrés à l'image de celui de Normandie *(Chandavoine, Guérente, Hurard, Merel, Mura et Nadaud, 2021)*.

La BPE est également utile pour mener des travaux sur le logement étudiant, avec l'approche fine souhaitée par les observatoires territoriaux du logement étudiant (OTLE). Lors de leur création en 2021, ces structures cherchaient à identifier les communes ayant sur leur territoire un établissement d'enseignement supérieur, pour affiner le champ des étudiants du supérieur tel que connu à partir du recensement de la population. Ces études ont conduit à diverses publications régionales, notamment en Auvergne-Rhône-Alpes (*Aude et Bianco, 2021*). La base a été mobilisée également pour les travaux sur l'école en milieu rural, en lien avec les conventions ruralité, pour aider à l'identification des territoires « fragilisés » où le temps d'accès des enfants devant quitter leur commune pour aller à l'école est potentiellement plus élevé qu'ailleurs. Il en est de même pour l'accès des jeunes non insérés à Pôle emploi ou aux établissements d'enseignement et de formation dans le cadre des pactes régionaux d'investissement dans les compétences.

13 *Via* Quetelet-Progedo-Diffusion .
14 Le distancier Metric (Mesure des trajets inter-communes / Carreaux) calcule des distances et des temps de parcours d'une commune à une autre ou d'un point à un autre lorsque les données sont géolocalisées. Il est mis à disposition *via Quetelet-Progedo Diffusion*.

▶ Des paniers d'équipements pour des catégories de population

> ***Le panier généraliste « Vie courante », regroupe des besoins de tous les jours, les paniers « Seniors », « Jeunes » et « Familles » ciblent les besoins des populations plus âgées, plus jeunes et des familles avec enfants.***

Pour des études plus fines, quatre **paniers thématiques** de commerces et services ont été définis par l'Insee, avec l'Agence nationale pour la cohésion des territoires et l'Institut d'aménagement et d'urbanisme d'Île-de-France. Ces paniers regroupent des équipements structurants pour certains types de population : le panier généraliste « Vie courante », regroupe des besoins de tous les jours, les paniers « Seniors », « Jeunes » et « Familles » ciblent les besoins des populations plus âgées, plus jeunes et des familles avec enfants.

Ainsi, les cartes d'accès au panier « Vie courante » et la part des populations éloignées constituent une base de travail utile pour les travaux d'élaboration de schémas d'aménagement du territoire, tels que les SDAASP[15] et les SRADDET[16]. Il en est de même de l'accès au panier « Jeunes » pour les politiques territoriales dédiées à la jeunesse : une étude menée à La Réunion s'est ainsi intéressée à l'accès aux équipements favorisant l'insertion sociale et professionnelle des jeunes sur l'île *(Grangé et Merceron, 2020)*. Le panier « Seniors » a été utilisé dans le cadre des schémas départementaux d'autonomie. À l'instar des travaux menés dans les Pays de la Loire *(Barre et Chesnel, 2019)*, ces derniers permettent aux décideurs publics d'anticiper l'impact du vieillissement de la population dans les années à venir avec l'arrivée aux grands âges des générations issues du baby-boom, accentué par l'attractivité des territoires.

▶ Des métadonnées volumineuses et rationalisées

Dans toute production statistique, il importe que le producteur fournisse aux utilisateurs une documentation adaptée à chaque produit et à chaque usage. Cette étape importante du processus statistique s'appuie sur ce qu'on appelle les métadonnées. À l'Insee, ces informations sont gérées dans le référentiel de métadonnées statistiques RMéS *(Bonnans, 2019)*.

Dans le cas de la BPE, la diversité des produits livrés, intermédiaires et finaux et la multiplicité des destinataires directs en aval du processus de production induisent de produire – et de mettre à jour – une documentation particulièrement abondante, sous des formats variés.

De nombreuses métadonnées sont ainsi associées à chaque produit : un dictionnaire des variables, un dessin de fichier, une note de mise à disposition, une note explicative complémentaire le cas échéant, la définition des types d'équipements, etc. Pour un même produit, ces métadonnées diffèrent selon les destinataires et sont à actualiser chaque année.

15 Schéma départemental d'amélioration de l'accessibilité des services au public.
16 Schéma régional d'aménagement, de développement durable et d'égalité des territoires.

Lors de leur refonte, ces métadonnées ont été structurées selon les normes adoptées par Eurostat : le SIMS (*Single integrated metadata structure*[17]) et le standard DDI (*Data Documentation Initiative*). Pour ce faire, les producteurs de la BPE ont bénéficié des services du référentiel RMéS et de l'appui des experts « qualité » de l'Insee. Au final, dans RMéS, la BPE est dotée d'un système de métadonnées original, qui lui est propre, conçu pour optimiser les mises à jour et en alléger la charge induite.

Les métadonnées en sortie de RMéS ont été construites pour répondre aux besoins en documentation des utilisateurs de la BPE, mais également pour être utilisées en *input* de leur propre processus. Par exemple, les notes de mise à disposition ou les notes explicatives complémentaires sont constituées de « briques de métadonnées » venant de RMéS. Certains outputs de la BPE, destinés à préparer des produits finaux de diffusion, sont accompagnés de fichiers *XML* qui peuvent être injectés en entrée d'un processus de conception d'indicateurs cartographiques ou de tableaux thématiques, et mettre à jour automatiquement la documentation qui les accompagne.

Avec la mise au point de ce système de métadonnées, la BPE a également gagné en cohérence et en homogénéité des documents et des fichiers produits en aval.

Une démarche qualité comme fil conducteur des évolutions

La rationalisation entreprise avec succès sur les métadonnées, a succédé aux gains déjà constatés sur la qualité des données elles-mêmes, lors de la bascule de SAS® en R des programmes qui harmonisent les différentes sources (voir *supra*).

Encadré 2. Un nécessaire comité des utilisateurs de la BPE

Pour pouvoir dresser le bilan des utilisations passées de la BPE et recueillir les nouveaux besoins des utilisateurs, qu'ils proviennent de l'Insee ou de l'externe, un comité des utilisateurs est une instance précieuse. Cette instance décrit et priorise les évolutions attendues et qui seront soumises au comité de pilotage de la BPE.

Le comité se réunit une fois par an. Il est constitué de représentants de l'Insee, de représentants de la Fédération nationale des agences d'urbanisme et de l'Agence nationale de la cohésion des territoires. Les demandes récurrentes des internautes sont également discutées, celles-ci parvenant régulièrement au pôle BPE *via* le service Insee contact. Le comité se réunit une fois par an.

Cette instance très utile est à l'origine de la diffusion de la BPE en évolution sur un pas quinquennal glissant (2015-2020, 2016-2021, etc.) en complément de la base millésimée, à la suite de nombreux besoins exprimés. Elle permet aussi de cibler les investigations quant à l'extension du champ couvert par la BPE. De nouveaux types d'équipement ont ainsi été intégrés : les mairies, les théâtres, les salles de concert, les implantations France services, etc. Des remarques sur la présence de faux actifs dans la BPE formulées lors de ce comité ont aussi eu pour conséquence le remplacement de la source la plus volumineuse (Sirene) par la source Sirus, répertoire statistique des entreprises (voir l'article 6 « Sirus » dans ce numéro du *Courrier des statistiques*).

17 Voir *https://ec.europa.eu/eurostat/fr/data/metadata/metadata-structure* : Eurostat distingue les métadonnées de référence (concepts, méthodologies utilisées, qualité des données) et les métadonnées structurelles (pour identifier les données statistiques).

La modernisation des autres étapes du processus nécessitait une analyse préalable et des outils adéquats. Une démarche qualité a donc été menée en 2021 par le pôle de production de la BPE, avec la collaboration de l'Unité Qualité de l'Insee : il a fallu remettre à plat l'existant, identifier les principaux risques du dispositif, ainsi que ses forces et faiblesses. Classiquement, l'enjeu pour la BPE était double : il s'agissait d'une part de rationaliser les procédures et d'en alléger la charge, et d'autre part de gagner en souplesse pour faciliter la prise en compte des demandes des utilisateurs ***(encadré 2)***.

En premier lieu, il s'agissait de situer le processus de production dans son environnement immédiat, en décrivant les échanges qui se déroulent avant et après le processus, selon l'approche du *Sipoc*[18]. Une analyse détaillée du processus a ensuite permis de le segmenter en briques appuyées sur le phasage du GSBPM[19] et d'identifier plus précisément, à chaque phase, les *inputs*, les traitements opérés, les *outputs*, les acteurs et leur rôle. S'en est suivie une analyse des forces et faiblesses selon les critères du Code des bonnes pratiques de la statistique européenne. La mise en exergue des risques de tout ordre (techniques, organisationnels, managériaux, etc.) pesant sur le bon déroulement des opérations a permis d'établir un plan de maîtrise des risques, identifiant trente actions pour les réduire. À titre expérimental, une analyse de la validation des données avant diffusion a aussi été testée.

Au final, un premier plan d'action qualité a été établi et validé par la maîtrise d'ouvrage comme fil conducteur de la modernisation de la BPE. Il comporte 18 actions s'étalant de juin 2021 à juin 2023. Certaines portent sur la rénovation de la chaîne de traitement, d'autres sur la description de l'opération conformément aux standards de RMéS. Le plan aborde aussi la révision du processus de contrôle et le partage des connaissances pour assurer la continuité de service.

▶ Un processus en constante évolution

L'analyse menée a notamment pointé le caractère extrêmement mouvant du processus de production de la BPE. Par nature, celui-ci est basé sur une évolution permanente des sources en entrée ; chaque année, à la différence des répertoires comme Sirene ou le RNIPP[20], de nouvelles sources sont intégrées à la base. Par ailleurs, compte tenu du concept même d'équipement, l'exhaustivité est impossible à atteindre. Si l'exhaustivité est visée, ce n'est pas sur l'ensemble des équipements offerts à la population, mais plutôt sur le champ couvert par les sources. Ainsi, certains services offerts en activités secondaires de commerces (par exemple des points de retraits de commerces en ligne) ou des infrastructures mises à disposition de tous (comme des aires de covoiturage ou des bornes de recharge de véhicules électriques) sont encore mal couverts par la statistique publique. Il est de ce fait difficile de parler de *répertoire* des équipements, encore moins de référentiel[21].

18 *Supplier input process output customer* : « fournisseur /intrant /processus /extrant /client ».
19 *General statistical business process model* ou modèle générique de description des processus de production statistique. Voir *(Unece, 2019) et (Erikson, 2020)*. Les phases de ce modèle offrent une grille d'analyse utile.
20 Voir l'article de Lionel Espinasse et Valérie Roux dans ce même numéro.
21 Voir l'article de Pascal Rivière dans ce même numéro, pour retrouver la définition d'un répertoire ayant statut de référentiel.

Cependant, le pôle BPE recherche constamment à intégrer de nouveaux types d'équipement. Régulièrement, de nouveaux fournisseurs sont associés au processus, par exemple en 2019 le Commissariat général à l'égalité des territoires, devenu Agence nationale pour la cohésion des territoires. Par ailleurs, les sources évoluent au fil des ans : les fournisseurs adoptent de nouveaux systèmes d'information, sont soumis à de nouvelles règles de gestion et de transmission d'information, etc. Ainsi, la classification *ad hoc* des types d'équipement évolue-t-elle chaque année pour intégrer les nouveautés.

L'adaptabilité des traitements et la réactivité des personnes en charge de la production de la BPE sont importantes. Cette nécessité s'illustre notamment dans le développement des traitements informatiques. Certaines opérations sont réalisées en interne, au moyen de programmes en R. Ces programmes viennent en complément d'une application développée et maintenue par l'Insee, qui permet de bénéficier d'une interface utile pour structurer les phases et faciliter la visualisation des opérations. La collaboration entre les agents du pôle BPE et les informaticiens permet de sécuriser la production de la base et de répondre aux normes de sécurité informatique de l'Insee.

Cependant, le caractère fortement évolutif des sources, du champ couvert, des informations disponibles imposent une réactivité forte aussi bien de l'informatique que du pôle BPE. Or, les modifications de la structure des sources sont parfois constatées une fois la source transmise par le fournisseur. De même, les informations nouvelles disponibles qui permettraient d'enrichir la BPE ne sont pas signalées dans des délais permettant de les prendre en compte. Elles peuvent en effet nécessiter de réaliser des développements informatiques pour faire évoluer l'application.

Depuis 2020, des travaux sont en cours pour passer d'un processus centré autour d'une application monolithe à une chaîne multi-outils. En segmentant le processus et en ciblant les *inputs*, les *outputs* et les traitements à effectuer, il est possible d'envisager des outils simples, dédiés à un nombre limité de tâches. Ce changement de paradigme permettra de simplifier la maintenabilité de la chaîne de traitement, de faciliter la prise en compte de nouvelles données disponibles et donc d'enrichir plus facilement la BPE. Il permettra d'être plus réactif aux demandes des utilisateurs. L'occasion se présente aussi de développer un nouveau support : une API[22] dont l'ambition serait de permettre aux fournisseurs de données d'insérer automatiquement dans leur système d'information les améliorations de la BPE (sur la qualité de l'adressage, sur l'ajout de zonages, etc.). L'année suivante, ce serait la BPE qui en serait bénéficiaire : un cercle vertueux se mettrait en place à chaque fourniture de données.

22 *Application Programming Interface* ou interface de programmation applicative en français. Il s'agit d'un ensemble normalisé de classes, de méthodes, de fonctions qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels.

▶ Bibliographie

- BARBIER, Max, TOUTIN, Gilles, LEVY, David, 2016. *L'accès aux services, une question de densité des territoires*. [en ligne]. 6 janvier 2016. Insee Première, n°1579. [Consulté le 4 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1908098*.
- BARRÉ, Martine et CHESNEL, Hélène, 2019. *La hausse du nombre de seniors dépendants accélérerait à partir de 2023*. [en ligne]. 13 juin 2019. Insee Analyses Pays-de-la-Loire, n°75. [Consulté le 4 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4169347*.
- 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 4 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf*.
- CHANDAVOINE, Bruno, GUÉRENTE, Sylvie, HURARD, Camille, MEREL, Aubin, MURA, Bruno et NADAUD, Laurence, 2021. *Le sport en Normandie : pratiques, équipements et emplois*. [en ligne]. 7 octobre 2021. Insee Dossier Normandie, n°19. [Consulté le 4 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5435270*.
- 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 4 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497085/courstat-4-8.pdf*.
- GRANGÉ, Claire et MERCERON, Sébastien, 2020. *Une proximité qui ne suffit pas à réduire les difficultés de formation et d'emploi. Équipements pour l'insertion sociale et professionnelle des jeunes Réunionnais*. [en ligne].19 novembre 2020. Insee Analyses Réunion, n°50. [Consulté le 4 septembre 2022]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4964208*.
- HAUTDIDIER, Baptiste et KUENTZ, V., 2011. Quelles invariances d'échelle entre densité de la population et des équipements ? Une approche empirique sur données communales françaises.In : *Rencontres de Théo Quant 2011*. Février 2011. Besançon, France. pp.32.
- KOMPIL, Mert, JACOBS-CRISIONIA, Chris, DIJKSTRA, Lewis et LAVALLE, Carlo, 2018. Mapping accessibility to generic services in Europe: *A market-potential based approach*. In : *Sustainable Cities and Society*. [en ligne]. 5 décembre 2018. n°47 (May, 2019) 101372. [Consulté le 4 septembre 2022]. Disponible à l'adresse : *https://doi.org/10.1016/j.scs.2018.11.047*.
- LAPAINE, Miljenko et USERY, E. Lynn, 2014. Projections cartographiques et systèmes de références. In : *Comité français de cartographie (CFC)*. [en ligne]. Septembre 2014. N°221, Chapitre 9, pp.87-101. Traduction par Didier Halter et Jean-François Hangouët. [Consulté le 4 septembre 2022]. Disponible à l'adresse : *https://www.lecfc.fr/new/articles/221-article-11.pdf*.
- UNECE, 2019. *Generic Statistical Business Process Model*. [en ligne]. Janvier 2019. Version 5.1. [Consulté le 4 septembre 2022]. Disponible à l'adresse : *https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1*.

Présentation du numéro N8

Avec cette nouvelle édition, le Courrier des statistiques livre son huitième numéro. La revue se donne une fois de plus pour ambition d'aborder, avec une tonalité qui se veut pédagogique quelques grandes problématiques auxquelles se confronte la statistique publique.
Le Courrier s'arrête en ouverture de ce numéro 8 sur l'enquête TeO qui explore de manière singulière comment les origines des immigrés ou des enfants d'immigrés influent sur leurs trajectoires et conditions de vie. Le second article propose d'analyser l'univers des statistiques dédiées aux collectivités locales.
Les répertoires sont à l'honneur dans les cinq articles qui suivent. Après avoir défini les répertoires, ces « référentiels indispensables et pourtant méconnus » comme des systèmes d'information normalisés et vivants, les deux articles suivants nous font pénétrer dans les constellations mêlées du Répertoire national d'identification des personnes physiques (RNIPP) et du système national de gestion des identifiants (SNGI). Puis on quitte le domaine des individus pour s'intéresser aux entreprises, avec le répertoire d'unités statistiques Sirus, outil indispensable au statisticien d'entreprises. Enfin, le dernier article nous plonge dans une singularité de l'appareil statistique français à travers la présentation de la base permanente des équipements (BPE).

ISSN 2107-0903
ISBN 978-2-11-162356-9

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE