N°6

COURRIER DES STATISTIQUES

Juin 2021

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE

Rédaction en chef
Odile Rascol
Contribution
Insee : Isabelle Anxionnaz,
Natacha Gualbert,
François Guillaumat-Tailliet,
Pierre Lamarche, Stéfan Lollivier,
Françoise Maurel,
Isabelle Robert-Bobée, Chloé Tavan
Cnav : Richard Merlen,
Christian Sureau
Depp : Loïc Midy
Smals : Isabelle Boydens, Gani Hamiti,
Rudy Van Eeckhout
Directeur de la publication
Jean-Luc Tavernier
Directeur de la collection
Pascal Rivière

Rédaction
Maryse Cadalanu, Pierre Glénat,
Fabienne Le Hellaye, Odile Rascol,
Pascal Rivière
Composition
Agence LATITUDE Nantes
5, rue Jacques Brel
« Les Reflets » Bâtiment A
44 800 SAINT-HERBLAIN
02 51 25 06 06
www.agence-latitude.fr
0254/21 - 0292/21
Photo de couverture
Adobe Stock®

Éditeur
Institut national de la statistique et des études économiques
88, avenue Verdier
92 541 MONTROUGE CEDEX

www.insee.fr

Courrier des statistiques N6

SOMMAIRE

PRÉSENTATION DU NUMÉRO

Relancé en 2018 dans une nouvelle formule, le *Courrier des statistiques* atteint sa troisième année d'existence et bientôt 50 articles. Nous avons pu explorer des sujets variés, méthodes, outils, mais aussi des questions institutionnelles ou juridiques posées par la statistique publique, en veillant à rester ouverts sur l'extérieur, en France ou à l'étranger, afin de se comparer et de nourrir nos réflexions.

La revue ne peut faire l'impasse sur un changement majeur de ces dernières années : si les statisticiens continuent à organiser la collecte d'informations par enquête, et à innover en la matière[1], ils doivent aussi, de plus en plus, tirer parti d'un monde où des données existent déjà, qu'ils n'ont pas construites. On pourra objecter que cela a toujours existé avec les sources administratives, mais celles-ci évoluent, s'enrichissent, comme on l'a vu dans le numéro N1, à travers la DSN[2] notamment. Plus généralement, les données externes font désormais toujours partie du paysage.

L'interrogation sur les gisements de données existants, leur mode d'obtention, leur degré d'élaboration, leur champ, leur temporalité, est systématique. Il s'agit là d'une préoccupation centrale dans le présent numéro, sixième de la nouvelle série, qui commence en présentant justement quatre sources de données, tout à fait essentielles pour des usages statistiques.

À tout seigneur tout honneur, c'est l'enquête Emploi qui ouvre le bal : enquête phare de la statistique publique en France, autour de laquelle gravitent d'autres opérations statistiques, l'enquête Emploi reste année après année une source inépuisable pour les études socio-économiques. Pour autant, celle-ci n'est pas immuable, elle s'adapte à un monde qui change. **François Guillaumat-Tailliet et Chloé Tavan** présentent les grandes lignes de la refonte de 2021 et ses motivations : l'exigence d'harmonisation au niveau européen[3], mais aussi la volonté de développer la possibilité pour les ménages d'utiliser internet pour répondre aux enquêtes de l'Insee. Cette somme de changements significatifs a nécessité un long travail de préparation et d'expérimentation, ainsi qu'une vaste opération préalable pour estimer les ruptures de séries.

Il fallait bien un jour ouvrir les colonnes de la revue aux promoteurs de Fidéli, souvent cité dans les numéros précédents : le fichier démographique sur les logements et les individus n'est ni le résultat d'une enquête, ni une source administrative *stricto sensu*. Comme l'expliquent **Pierre Lamarche et Stéfan Lollivier**, Fidéli est une pure construction des statisticiens pour leurs propres besoins, un travail de mise en cohérence, et d'enrichissement de sources administratives, notamment fiscales. La cohérence, l'exhaustivité et la variété d'informations disponibles sont essentielles pour son insertion dans le système d'information du Service statistique public. À partir de plusieurs sources « brutes », le dispositif élabore une liste unique de logements d'habitation et une liste unique d'individus, localisés dans leur logement principal, tout en regroupant les informations socio-démographiques les concernant. Fidéli permet de réaliser des études spécifiques et d'échantillonner des enquêtes, mais aussi de compléter, par appariement, des données d'enquêtes avec des données socio-démographiques finement localisées, ce qui démultiplie son potentiel pour l'analyse sociale.

1. Voir le *numéro N3*.
2. Déclaration sociale nominative, voir le *numéro N1*.
3. Dont nous avions annoncé le cadre dans le *numéro N3* et qui fait écho à d'autres refontes (voir *numéro N2*).

L'échantillon démographique permanent (EDP) apporte une corde supplémentaire à notre arc : la profondeur temporelle, la possibilité de travailler sur des cohortes, de mieux comprendre des évolutions dans le temps et au fil des générations. C'est une source déjà ancienne, qui retrace pas moins de 3,7 millions de trajectoires individuelles, dont 200 000 depuis plus de 50 ans. **Isabelle Robert-Bobée et Natacha Gualbert** en décrivent les fonctionnalités actuelles − car son contenu et son périmètre n'ont cessé de s'étoffer au fil des années − et les principales innovations − car il a fallu également que l'EDP s'adapte aux évolutions de son environnement. Il s'est ainsi enrichi récemment de données socio-fiscales, comme Fidéli. La compilation de différentes sources fait l'originalité de l'EDP : si elle rend plus complexe son exploitation pour les études, elle offre en retour des possibilités d'analyse des trajectoires de plus en plus diversifiées.

Continuant l'exploration du vaste univers des sources administratives, **Christian Sureau et Richard Merlen** abordent le répertoire général des carrières unique (RGCU) mis au point par la Cnav (Caisse nationale d'assurance vieillesse). Cette gigantesque base de données doit, à terme, permettre aux organismes de retraite, aux entreprises, à leurs salariés et aux retraités, de partager une information respectant les mêmes concepts, sur les différentes dimensions des périodes qui composent une carrière professionnelle (activité salariée, période de chômage, etc.). La qualité de ce répertoire est assurée, entre autres, par un mécanisme sophistiqué de contrôle des données, à plusieurs niveaux. Au départ outil au service des usagers et de l'efficience administrative, voici une base de données qui est promise à un bel avenir pour des usages statistiques, de par sa richesse sans équivalent. Elle incorpore, comme l'EDP, une dimension temporelle : on remonte même bien plus loin, jusqu'aux années trente. Contrairement aux trois autres sources, le RGCU n'est pas encore disponible, et il faudra patienter encore quelques mois jusqu'en 2022 pour qu'une première version soit mise à disposition des chercheurs *via* le CASD.

Pour répondre aux besoins d'évaluation de la puissance publique, les services statistiques ministériels ont à leur disposition des sources de données, mais les relier entre elles n'a rien d'immédiat dès lors qu'on ne dispose pas d'identifiant commun. Cette activité d'« appariement » de fichiers (*record linkage* en anglais) a suscité depuis les années 80-90 une vaste littérature académique, notamment au Canada, aux Pays-Bas, en Australie, aux États-Unis, ou en Italie. Avec la multiplication des sources administratives disponibles, l'intérêt pour ces méthodes et leur mise en application s'accroît au sein de la statistique publique française.

Les travaux menés par la Depp s'inscrivent dans cette mouvance. InserJeunes, dispositif sur l'insertion professionnelle des jeunes, est ainsi basé sur l'appariement de sources administratives exhaustives, relatives à la scolarité des élèves et des apprentis, à la réussite aux examens, aux contrats d'apprentissage et aux contrats salariés, et de la déclaration sociale nominative (DSN). L'article de **Loïc Midy** explicite les étapes nécessaires pour mener à bien l'appariement de ces fichiers sur identifiant indirect : normalisation, indexation, calcul de similarités, classification des paires et évaluation/validation. Il met en évidence au passage la complexité de l'opération et les limites des approches « naïves » de ce sujet. L'auteur s'interroge également sur la manière d'outiller le processus, en effectuant un tour d'horizon des outils d'appariement *open source* existants.

S'intéresser aux données administratives, c'est aussi s'interroger sur la qualité de ces données, ne pas considérer celle-ci comme un acquis. De ce point de vue, la Belgique a clairement un temps d'avance, avec un travail de fond mené sur les anomalies dans les données et la manière d'y remédier en remontant à la source même de l'information : l'intérêt

de la démarche a même été acté par un « arrêté royal » qui l'impose aux administrations. **Isabelle Boydens, Gani Hamiti et Rudy Van Eeckhout** nous présentent ainsi un prototype original, appelé ATMS (*Anomalies & Transactions Management System*). Il permet un suivi des anomalies et des traitements, en support à la méthode dite du *back tracking* : dans une approche préventive de la qualité des données, la méthode est destinée à améliorer structurellement la qualité à la source, et fait également le lien avec des approches curatives, plus classiques (*data quality tools*).

Enfin, le dernier article de ce numéro nous emmène dans une direction beaucoup plus institutionnelle, avec le Conseil national de l'information statistique (Cnis). Il fait écho aux articles précédemment publiés, qui portaient sur deux autres acteurs de la gouvernance statistique en France, que sont l'Autorité de la Statistique Publique d'une part, le Comité du Label d'autre part[4]. **Isabelle Anxionnaz et Françoise Maurel** nous décrivent ici les principes de fonctionnement du Cnis et nous en dévoilent les arcanes. Elles nous rappellent notamment son rôle crucial dans l'organisation de la concertation entre producteurs et utilisateurs de statistiques publiques. Elles démontrent qu'institution n'est pas synonyme d'organisation figée : les rencontres et les groupes de travail du Cnis produisent, en toute transparence, une vision partagée sur les besoins de statistiques et sur la pertinence des productions. Et au-delà, les recommandations sont mises en application dans les programmes de la Statistique publique. Le Cnis contribue enfin, de par sa vigilance aux évolutions des usages, à l'adaptation en continu du Service statistique public dans ses méthodes de production, à leur rationalisation et à leur articulation avec les sources existantes.

Odile Rascol
Rédactrice en chef, Insee

4. Voir le *numéro N5*.

UNE NOUVELLE ENQUÊTE EMPLOI EN 2021

ENTRE IMPÉRATIF EUROPÉEN ET VOLONTÉ DE MODERNISATION

François Guillaumat-Tailliet et Chloé Tavan***

Le nouveau règlement européen sur les statistiques sociales (IESS), entré en vigueur en janvier 2021, exige une plus grande harmonisation des questionnaires de l'enquête européenne sur les forces de travail, dont l'enquête Emploi est la déclinaison française. La nécessaire refonte du questionnaire de l'enquête Emploi a été l'occasion de l'adapter aux nouvelles formes d'emploi et aux pratiques professionnelles émergentes. En donnant de plus aux enquêtés la possibilité de répondre sur internet en réinterrogation, l'enquête Emploi a modernisé son protocole de collecte. Elle fait ainsi figure de pionnière dans l'orientation prise par l'Insee de développer les enquêtes multimodes, avec pour ambition de maintenir dans la durée des taux de collecte élevés et de mieux cibler le recours aux enquêteurs.

Pour se préparer au passage à la nouvelle enquête et être en mesure de respecter le calendrier de publication, une opération de mise en pré-production de grande envergure, appelée Pilote, a été conduite sur l'ensemble de l'année 2020. Les données collectées, comparées pendant un an à celles de l'enquête Emploi en production, permettent d'estimer les ruptures qu'une telle refonte est susceptible de provoquer sur les principaux indicateurs d'emploi, de chômage et de formation. Dès juin 2021, l'Insee aura ainsi été en mesure de diffuser les principaux indicateurs du marché du travail issus de la nouvelle enquête, et des séries longues corrigées de la rupture.

The new European regulation on social statistics (IESS), which has come into force in January 2021, requires greater harmonisation of the questionnaires of the European Labour Force Survey (LFS). The necessary overhaul of the questionnaire of the French LFS was an opportunity to adapt it to new forms of employment and emerging professional practices. In addition, by giving respondents the possibility of answering on the Internet during re-interviews, the French LFS has modernised its data mode collection. It is thus a pioneer in the direction taken by INSEE to develop multi-mode surveys, with the aim of maintaining high response rates over time and better targeting the work of interviewers.

In order to prepare for the transition to the new survey and to be able to meet the publication schedule, a large-scale pre-production operation, called Pilot, was conducted throughout 2020. The data collected, compared for one year with those of the LFS in production, make it possible to estimate the breaks that such a redesign is likely to induce on the main employment, unemployment and training indicators. By June 2021, INSEE will thus have been able to disseminate main labour market indicators from the new survey, and time-series corrected for the break.

* Responsable du programme d'Évolution de l'enquête Emploi, Insee,
francois.guillaumat-tailliet@insee.fr
** Cheffe de la division Emploi, Insee,
chloe.tavan@insee.fr

Née en 1950 pour permettre une mesure régulière de l'emploi et du chômage, l'enquête Emploi a connu au fil des années de nombreux changements, sur les informations recueillies, mais aussi sur des aspects de nature méthodologique ou technique. L'histoire de l'enquête Emploi illustre à elle seule les évolutions importantes connues par les enquêtes auprès des ménages sur les dernières décennies. L'entrée en vigueur d'un nouveau règlement européen sur les statistiques sociales[1], en 2021, invite à tourner une nouvelle page de cette histoire. L'obligation européenne est l'occasion d'un réaménagement plus profond de l'enquête, qui n'avait pas connu d'évolution majeure depuis 2013. En particulier, un nouveau protocole de collecte propose désormais la réponse sur internet en réinterrogation. Singularité de cette refonte : l'entrée en vigueur de la nouvelle enquête a été précédée d'une vaste opération de mise en pré-production, sur l'ensemble de l'année 2020 et le premier trimestre 2021, baptisée Pilote.

LA SOURCE POUR MESURER L'EMPLOI ET LE CHÔMAGE AU SENS DU BIT

Même si ses usages vont bien souvent au-delà, en raison de la taille de son échantillon et de la richesse de son questionnaire, l'objectif premier de l'enquête Emploi est de mesurer l'emploi et le chômage, selon les concepts définis par le Bureau international du travail (BIT) (OIT, 1982). Ces concepts s'appuient sur des définitions factuelles de l'emploi et du chômage, indépendantes des régimes sociaux associés aux emplois ou encore des dispositifs de suivi ou d'indemnisation du chômage. Ils permettent ainsi de disposer, autant que possible, d'une mesure stable dans le temps et harmonisée entre les différents pays. Ils sont mis en œuvre par la plupart des instituts de statistique du monde, en particulier ceux de l'Union européenne. Pour améliorer la comparabilité entre les pays, Eurostat propose des définitions opérationnelles de ces concepts.

> *« Ces concepts s'appuient sur des définitions factuelles de l'emploi et du chômage, indépendantes des régimes sociaux associés aux emplois ou encore des dispositifs de suivi ou d'indemnisation du chômage. »*

Les critères pour classer la population entre emploi, chômage ou inactivité au sens du BIT sont nombreux et précis. Ils concernent, par exemple, le fait d'avoir travaillé ou non pendant une semaine donnée, dite de référence, les raisons d'absence pour les personnes ayant un emploi mais n'ayant pas travaillé cette semaine-là, le fait d'avoir effectué des démarches précises de recherche d'emploi ou encore le fait d'être disponible pour travailler. Aucun répertoire administratif ne contient de telles informations ; elles ne peuvent être recueillies qu'à travers des questions dans le cadre d'une enquête auprès des personnes. En France, cette enquête, c'est l'enquête Emploi.

1. Il s'agit du règlement IESS *(Integrated European Social Statistics framework regulation)*, voir *infra* et également les références réglementaires en fin d'article.

UNE ENQUÊTE CENTRALE DU DISPOSITIF STATISTIQUE SUR LE MARCHÉ DU TRAVAIL

L'enquête Emploi présente de nombreux atouts qui lui offrent une place centrale dans le dispositif statistique sur le marché du travail. La richesse de son questionnaire ***(figure 1)*** lui permet d'éclairer de nombreuses dimensions de la relation à l'emploi, bien au-delà de la seule mesure de l'emploi ou du chômage ; les concepts de sous-emploi et de halo autour du chômage, permettent par exemple de révéler la porosité des frontières entre l'emploi, le chômage et l'inactivité ; la description détaillée des emplois occupés (contrat, temps de travail, position socioprofessionnelle, statut, etc.) permet d'objectiver la qualité des emplois.

Conduite depuis des décennies, l'enquête Emploi peut saisir des transformations au long cours à l'œuvre sur le marché du travail. Enfin, son inscription dans un cadre européen (l'enquête Emploi est la déclinaison française de l'enquête européenne sur les forces de travail (*Labour Force Survey*, LFS)) en fait un outil précieux pour les comparaisons internationales.

L'enquête Emploi donne lieu à des valorisations variées, en France (Insee, services statistiques ministériels ou d'autres institutions françaises comme le Trésor et France Stratégie) et à l'étranger (Eurostat, OCDE, etc.). En plus de la publication trimestrielle d'un jeu d'indicateurs éclairant la conjoncture du marché du travail dont son indicateur phare, le taux de chômage, l'enquête est exploitée pour des études plus structurelles ou évaluatives sur des sujets variés, par exemple sur le chômage de longue durée, le halo autour du chômage ou encore les contrats courts.

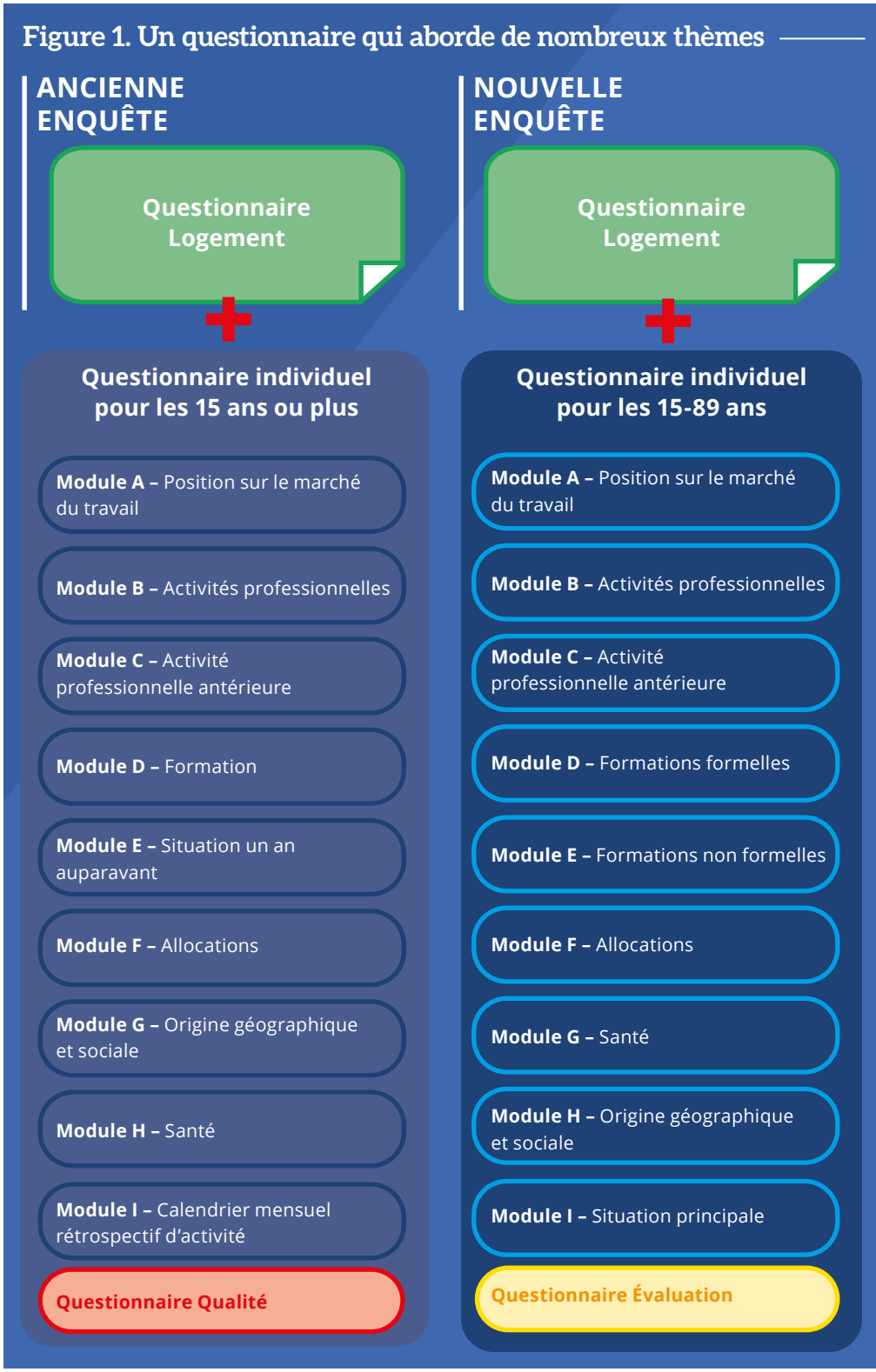
UNE HISTOIRE DÉJÀ LONGUE

Depuis sa naissance en 1950, l'enquête Emploi a connu de nombreux changements, de nature conceptuelle, pour se conformer aux orientations du BIT ou d'Eurostat ou mieux mesurer les transformations du marché du travail, mais aussi d'ingénierie statistique (méthodes d'échantillonnage ou de traitement de la non-réponse, modes de collecte, etc.) ou encore techniques avec l'informatisation croissante de la collecte et du traitement des données (Goux, 2003).

Sans revenir en détail sur l'intégralité de son histoire, les vingt dernières années illustrent l'importance et la variété des modifications que l'enquête a connues. L'année 2003 constitue une première date marquante : à la suite d'une décision européenne, d'annuelle l'enquête Emploi devient une enquête « en continu », c'est-à-dire qu'elle porte depuis sur l'ensemble des semaines de l'année. Elle constitue dès lors une source majeure pour l'analyse conjoncturelle. Autre date importante, 2009 : en réponse à une « polémique » sur les chiffres du chômage en 2006-2007[2] (Durieux *et alii*, 2007), son échantillon est augmenté de 50 %. En 2013, son questionnaire est rénové, pour en faciliter la passation, notamment au téléphone, améliorer la codification des variables de profession et de diplôme, enrichir la connaissance du marché du travail et se conformer aux orientations d'Eurostat sur certains

2. Le niveau et l'évolution du chômage étaient à l'époque suivis à partir de deux sources : les données administratives sur les demandeurs d'emploi et les données de l'enquête Emploi. Les résultats des deux sources étaient raccordés par un calage des données administratives sur le chômage mesuré par l'enquête Emploi. En 2006, pour la première fois en 20 ans, ce calage n'a pas pu être effectué en raison d'une forte divergence entre les deux sources. Par la suite, il a été mis fin à ce calage et des mesures ont été prises pour améliorer la précision de l'enquête Emploi et la compréhension des écarts entre les deux sources.

Figure 1. Un questionnaire qui aborde de nombreux thèmes



indicateurs (formation, halo autour du chômage). La mesure de l'emploi et du chômage en sera affectée. En 2013, l'application informatique de gestion de l'enquête, instrument capital pour une enquête d'une telle ampleur soumise à de fortes contraintes de production, est également refondue. En 2014, le périmètre géographique de l'enquête est étendu : les départements d'outre-mer (hors Mayotte[3]) intègrent alors le processus de l'enquête Emploi en continu.

> *« L'enquête Emploi n'évolue pas chaque année, mais tous les dix ans environ, dans le cadre de lourds exercices de refonte. »*

Chaque modification de l'enquête affectant la mesure des indicateurs induit un travail important de rétropolation, c'est-à-dire de construction de séries sur longue période qui offrent une vision cohérente du marché du travail. Pour cette raison, l'enquête Emploi n'évolue pas chaque année, mais tous les dix ans environ, dans le cadre de lourds exercices de refonte.

UNE ENQUÊTE QUI S'INSCRIT DANS UN CADRE EUROPÉEN

Depuis 1973, les enquêtes sur les forces de travail s'inscrivent dans un cadre réglementaire européen, qui a évolué, allant dans le sens d'une harmonisation croissante. Ce cadre fixe un certain nombre de contraintes, mais laisse aussi aux pays une marge de liberté pour la mise en œuvre de l'enquête. Ainsi, qu'il s'agisse du règlement qui prévalait jusqu'en 2020 (règlement 577/1998 du 9 mars 1998) ou de celui qui s'impose désormais (IESS FR – *Integrated European Social Statistics framework regulation*), les informations que l'enquête doit collecter sont définies au niveau européen, ce qui contraint, dans une certaine mesure, le questionnaire.

De même, ce règlement impose certains principes méthodologiques, qu'ils soient relatifs à l'échantillon (comme le fait d'avoir un échantillon rotatif, uniformément réparti sur l'ensemble de l'année et de taille suffisante, *via* des exigences en termes de précision des indicateurs), ou aux traitements aval (en imposant par exemple que l'enquête soit calée sur des marges de population au niveau régional).

Enfin, en limitant par exemple à cinq semaines la durée maximale de collecte, le règlement encadre certains aspects du protocole.

Les contraintes européennes sont des exigences minimales. Dans le cas de la France, on va bien souvent au-delà, par exemple sur les délais de collecte où on s'impose une fenêtre plus réduite, ou sur le questionnaire dont l'édition française comporte de nombreuses questions non requises.

D'autres aspects de l'enquête sont en revanche laissés à la discrétion des pays. Il en va ainsi du fait d'avoir un échantillon d'individus ou de logements ou du mode de collecte ; les pays sont ainsi libres de recourir à des modes de collecte intermédiés par un enquêteur (par téléphone ou en face-à-face) ou non (par internet ou sur papier) (Eurostat, 2019).

3. Une enquête annuelle est menée depuis 2013 à Mayotte.

UNE ENQUÊTE MENÉE EN CONTINU, AUPRÈS D'UN ÉCHANTILLON DE GRANDE TAILLE

Une caractéristique centrale de l'enquête Emploi, qui détermine son échantillonnage et conditionne fortement l'organisation de sa collecte, est le fait qu'il s'agit d'une enquête en continu : l'échantillon est divisé sur l'ensemble des semaines de l'année ; c'est comme si l'enquête était réalisée auprès de 52 échantillons. Chaque échantillon hebdomadaire est associé à une semaine, dite de référence, relativement à laquelle les personnes interrogées doivent décrire leur situation par rapport à l'emploi.

Pour éviter les biais de mémoire et garantir des données de qualité, la période de collecte est très courte : jusqu'en 2020, les enquêtés devaient répondre dans les deux semaines et deux jours qui suivaient la semaine de référence.

Une autre particularité de l'enquête est qu'il s'agit d'un panel de logements. Les habitants des logements échantillonnés sont interrogés six trimestres de suite, afin d'estimer de façon plus robuste les évolutions et de produire des analyses longitudinales. Jusqu'en 2020, en première et dernière interrogations, les ménages étaient interviewés en face-à-face ; en ré-interrogation, où le questionnaire était plus court, ils étaient enquêtés par téléphone. Le questionnaire se compose d'une partie qui décrit succinctement l'ensemble des habitants du logement, puis d'un questionnaire individuel auquel doit répondre chaque habitant du logement âgé de 15 ans ou plus. Le questionnaire comporte un module complémentaire, ensemble d'une vingtaine de questions qui, chaque année, vient éclairer un thème particulier.

L'enquête Emploi est une opération de grande envergure. Chaque trimestre, ce sont près de 100 000 personnes qui répondent à l'enquête. En 2019, elle représentait à elle seule plus d'un tiers de la charge d'enquête auprès des ménages de l'Insee.

POURQUOI UNE NOUVELLE ENQUÊTE ?

Depuis 2014, l'enquête Emploi n'avait pas connu de changement majeur, mais, début 2021, une nouvelle enquête Emploi est mise sur le terrain, pour répondre avant tout à un impératif européen. En effet, le nouveau règlement cadre européen sur les statistiques sociales (IESS) s'impose aux instituts statistiques depuis début 2021 (Cases, 2019). Celui-ci ne modifie pas en profondeur les requis européens en termes d'ingénierie statistique ou de protocole ; en particulier, il laisse toujours toute latitude aux pays pour le choix de leur mode de collecte[4]. Sa principale ambition est d'aller plus loin sur la voie de l'harmonisation des informations collectées, aussi bien entre les pays qu'entre les différentes enquêtes qu'il couvre. L'innovation portée par IESS est en effet d'avoir un seul règlement-cadre qui chapeaute les différentes enquêtes sociales européennes.

Le nouveau règlement réaffirme aussi la nécessité de couvrir l'ensemble du territoire national, ce qui n'est pas encore le cas avec la nouvelle enquête Emploi en continu, qui ne couvre pas le département de Mayotte (voir *supra*). Mayotte rejoindra le dispositif en vigueur sur le reste du territoire en 2024, le temps d'instruire ce chantier et de mettre en place l'organisation adéquate.

4. Le nouveau règlement européen impose également des niveaux de précision à atteindre pour certains indicateurs-clés, comme la part de chômage pour l'enquête Emploi, induisant des contraintes dans la taille de l'échantillon et son allocation régionale. Pour l'enquête Emploi, ces contraintes ont été prises en compte dans le cadre du renouvellement de son échantillon, à compter du 3e trimestre 2019, dans le cadre du renouvellement de l'échantillon-maître de l'Insee (Sillard *et alii,* 2020).

> *« La refonte rendue nécessaire par cet impératif européen a été l'occasion d'embarquer d'autres évolutions souhaitées au niveau national. »*

Afin d'introduire une rupture de série unique, la refonte rendue nécessaire par cet impératif européen a été l'occasion d'embarquer d'autres évolutions souhaitées au niveau national, mais non requises au niveau européen. La première était de moderniser le protocole de collecte de l'enquête en proposant internet comme mode de réponse supplémentaire, conformément à l'orientation prise par l'Insee de développer les enquêtes multimodes, à l'instar de nombreux autres instituts nationaux de statistique[5]. La seconde consistait à rénover la méthode de pondération de l'enquête.

UN QUESTIONNAIRE QUI RÉPOND À DES IMPÉRATIFS EUROPÉENS…

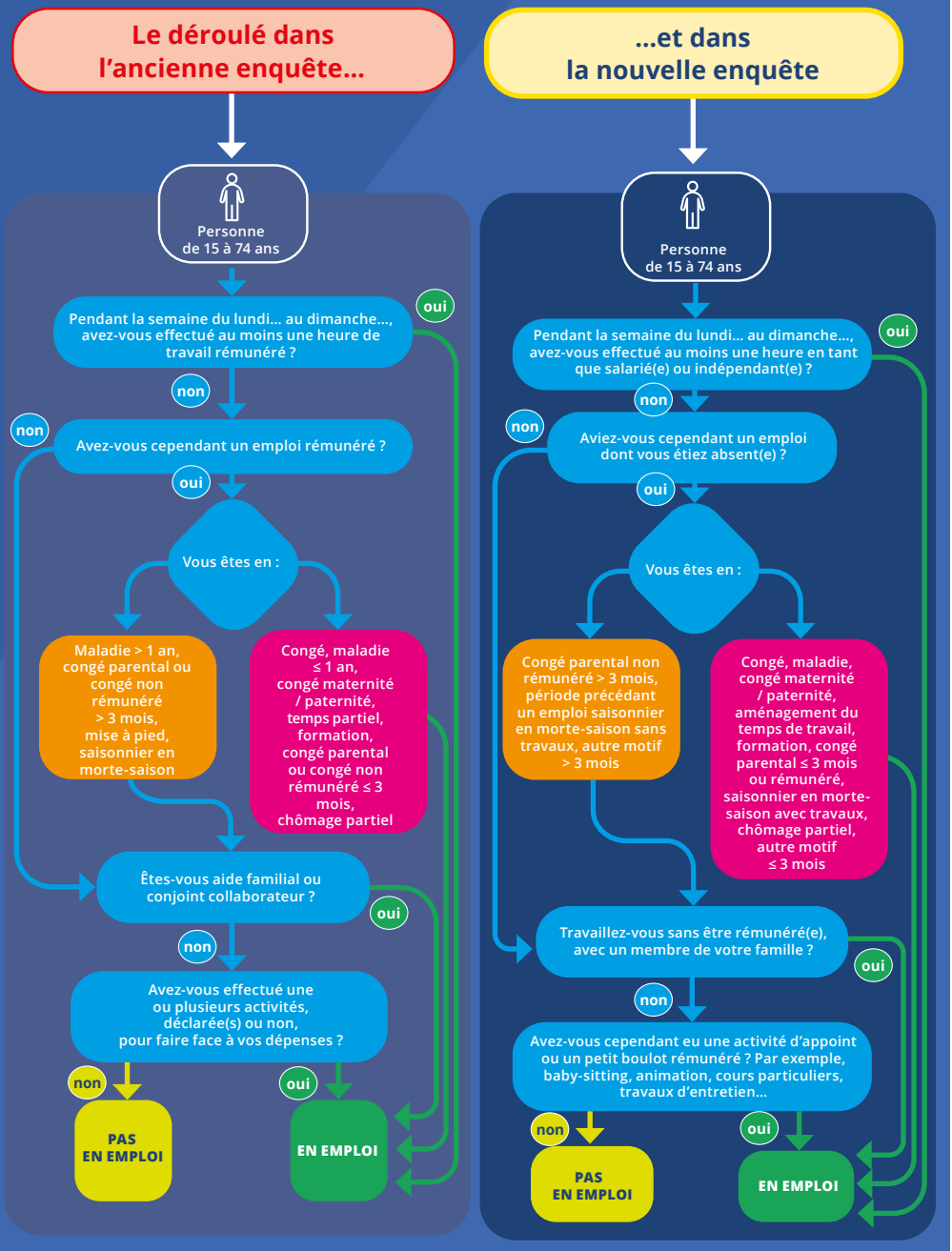
La rénovation du questionnaire répond à trois objectifs. En premier lieu, et c'était incontournable, il s'agissait de le mettre en conformité avec le règlement IESS, qui a notamment actualisé la liste des informations demandées. Par exemple, les raisons de migration, la durée contractuelle de travail ou encore le fait d'avoir travaillé pendant ses études font désormais partie des variables requises. Cependant, dans la mesure où le questionnaire français était déjà bien plus riche que ce qu'exigeait Eurostat, peu de questions nouvelles ont dû être introduites.

Mais l'Europe a poussé plus loin l'ambition d'harmonisation entre les pays : pour les variables les plus centrales, à savoir celles servant à déterminer le statut d'activité au sens du BIT, l'ordre des questions et le déroulé du questionnaire sont désormais imposés. C'est là la grande nouveauté introduite par IESS. Si cela ne s'est pas traduit par un changement en profondeur du questionnaire français, qui était finalement assez proche du nouveau canevas européen, à y regarder de plus près, les changements sont plus nombreux qu'il n'y paraît. Par exemple, la nouvelle interprétation par Eurostat des critères du BIT conduit à classer en emploi les personnes qui déclarent avoir un emploi mais en être absentes pour cause de maladie quelle que soit la durée de l'absence, contre un plafond de un an dans l'ancienne enquête. De même, les personnes absentes pour congé parental sont désormais classées en emploi si leur absence est inférieure à 3 mois mais aussi, et c'est nouveau, si elles perçoivent un revenu compensatoire, comme la prestation partagée d'éducation de l'enfant en France ***(figure 2)***. Autre exemple, la position des questions sur le souhait de travailler et sur la recherche d'emploi est inversée, pour revenir à la situation qui prévalait avant 2013. Enfin, la liste des démarches de recherche d'emploi est nettement raccourcie par rapport à l'ancienne enquête. Ce ne sont ici que des exemples. Tous ces changements sont susceptibles de modifier la teneur des informations recueillies et donc d'affecter la mesure de l'emploi et du chômage par l'enquête.

5. En particulier, les Pays-Bas et le Danemark ont été pionniers en matière de réponse à la LFS sur internet. Voir également (Signore, 2019).

Figure 2. Du concept à la mise en oeuvre dans le questionnaire, l'exemple de l'emploi

Le classement en emploi est le fruit d'une séquence de questions qui a dû être revue avec le nouveau règlement.



… AINSI QU'À DES DEMANDES NATIONALES…

Au-delà des requis européens, la refonte du questionnaire a été l'occasion de répondre à des attentes nationales sur des sujets jusqu'à présent peu ou pas couverts par l'enquête Emploi. Il s'agit essentiellement de questions sur :

- les conditions de travail, avec notamment des questions sur le télétravail (seul le travail à domicile, information européenne, était jusqu'à présent couvert par l'enquête) ;
- les nouvelles formes d'emploi (visant par exemple à identifier les CDI intérimaires, les apprentis en CDI ou les situations de dépendance des travailleurs indépendants), en lien avec les recommandations du Conseil national de l'information statistique (Cnis) (Gazier, Minni et Picart, 2016) ;
- et la formation non formelle à but professionnel.

> *« Mettre en œuvre la nouvelle nomenclature du BIT sur le statut d'emploi. »*

De même, le questionnaire intègre des questions nouvelles afin de pouvoir mettre en œuvre la nouvelle nomenclature du BIT sur le statut d'emploi, définie lors de la 20e conférence des statisticiens du travail d'octobre 2018 (OIT, 2018). Sur ce point, la France est en avance sur l'agenda européen : Eurostat vient seulement de lancer, début 2021, une *task force* pour réfléchir à la mise en œuvre de cette nouvelle nomenclature dans les enquêtes sur les forces de travail.

La nouvelle enquête Emploi est également la première enquête auprès des ménages à mettre en œuvre la nomenclature des professions et catégories socioprofessionnelles (PCS) rénovée (Amossé, 2020).

Enfin, certaines spécificités françaises ont été réaffirmées à l'occasion de cette refonte ***(encadré 1)*** ; il s'agit notamment du recueil de la catégorie socioprofessionnelle des parents ou des questions sur la perception d'allocations.

… ET ADAPTÉ À LA COLLECTE PAR INTERNET

La conception du nouveau questionnaire a enfin dû prendre en compte l'autre innovation majeure portée par la refonte : l'introduction de la réponse auto-administrée par internet en réinterrogation. Comme l'objectif était d'avoir un questionnaire unique quel que soit le mode de collecte, il a fallu trouver des formulations simples et adaptées à l'auto-administré.

L'évolution la plus emblématique à ce sujet concerne le recueil des libellés de profession et de diplôme. Le protocole qui avait cours jusque-là consistait en une saisie en clair de ces libellés. Il nécessitait un important travail de formation et d'accompagnement des enquêteurs et n'était clairement pas adapté à la collecte en auto-administré. Un recueil *via* la sélection dans une liste prédéfinie a donc été mis en place. Face au nombre important de professions (la liste ne comporte pas moins de 5 000 libellés), l'enjeu a été de mettre au point des outils efficaces de navigation dans la liste, suivant de près le protocole établi par le groupe de travail du Cnis dédié à la refonte de la PCS (Amossé, 2020).

Par ailleurs, la collecte par internet pour les réinterrogations supposait un questionnaire le plus fluide et le plus court possible, tout en veillant à bien repérer les changements de situation. En particulier, il a fallu fortement alléger le questionnaire de dernière interrogation (voir *infra* le protocole de collecte), qui auparavant, du fait d'une passation en face-à-face, était plus long qu'une ré-interrogation intermédiaire.

Pour ce faire, des questions ont été déplacées de la dernière à la première interrogation. C'est notamment le cas du module complémentaire. Pour compenser, le module sur la situation un an auparavant, posé en première interrogation, a été supprimé ; il s'est avéré qu'il pouvait être entaché de biais de mémoire et il perdait de la pertinence avec le développement des exploitations longitudinales de l'enquête.

Enfin, les modalités du passage à la nouvelle enquête avaient été arrêtées : une bascule à une date donnée de l'ensemble des unités enquêtées, quelle que soit leur ancienneté dans le panel. Il a donc fallu concevoir un questionnaire spécifique, dit de bascule, pour gérer cette transition. Le questionnaire de bascule est un questionnaire hybride entre une première interrogation et une réinterrogation.

Encadré 1. Quelques exemples de questions nouvelles ou répondant à des besoins nationaux

Questions franco-françaises qui existaient et sont maintenues

- Statut d'occupation du logement, perception d'aides au logement (questionnaire logement)
- Caractéristiques de l'emploi recherché ou souhaité (type de contrat, temps de travail) (modules A et B)
- Situation à l'entrée dans l'entreprise, situation avant le début de l'emploi, contrats aidés (module B)
- Date de fin des études initiales (module D)
- Retraites et autres allocations (module F)
- Profession du père et de la mère (module H)

Questions européennes nouvelles

- Heures contractuelles (module B)
- Expériences d'emploi pendant le cursus menant au plus haut diplôme (module D)
- Pays de résidence antérieur, motif de migration (module H)

Questions franco-françaises nouvelles

- Questions pour coder la PCS 2020, module sur les clients et autres relations des indépendants, télétravail, profession et contrat dans l'emploi secondaire (module B)
- Contrat du dernier emploi (module C)
- Description de la formation non formelle à but professionnel la plus récente (module E)

LA REFONTE DU QUESTIONNAIRE : UN CHANTIER DE LONGUE HALEINE

Les travaux de conception du nouveau questionnaire ont été menés dans le cadre européen et se sont étalés sur une dizaine d'années. Pour préparer les nouveaux textes régissant l'enquête Emploi, Eurostat a mis en place plusieurs groupes de travail (ou *task forces*) dédiés sur certaines parties du questionnaire. En particulier, au printemps 2011, une *task force* a été chargée d'élaborer une nouvelle définition opérationnelle de l'emploi et du chômage dérivée des principes du BIT, restés eux inchangés ; une autre a été consacrée à la mesure de la durée du travail (Eurostat, 2018) ; une autre avait enfin pour objectif de redéfinir le contenu de l'ensemble des modules complémentaires[6].

> *« Les travaux de conception du nouveau questionnaire ont été menés dans le cadre européen et se sont étalés sur une dizaine d'années. »*

Une fois le cadre européen suffisamment dessiné, les travaux de conception au niveau français ont pu être lancés. En 2017-2018, pas moins de dix groupes de travail reprenant les grands thèmes de l'enquête ont réuni des experts issus de l'ensemble de la statistique publique.

Le nouveau questionnaire[7] a été testé à diverses reprises : en juin 2018, le questionnaire de première interrogation a été testé en face-à-face auprès de 1 000 ménages ; en décembre 2018, un test 100 % internet a été réalisé auprès de personnes volontaires ; en 2019, un test complet, en deux vagues (première interrogation et réinterrogation), a été mené auprès de 1 400 ménages selon le protocole que l'on souhaitait dérouler à la cible.

UNE ORGANISATION GÉNÉRALE DE LA COLLECTE INCHANGÉE

Le protocole de la nouvelle enquête a été mis au point dans le cadre d'un groupe de travail interne à l'Insee, associant notamment des acteurs du « terrain » de l'enquête : des gestionnaires d'enquête et des enquêteurs.

L'organisation générale de la collecte ne change pas : la collecte reste organisée en continu sur toutes les semaines de l'année et structurée autour de la semaine de référence ; l'unité enquêtée reste le logement, qui est enquêté six trimestres consécutifs. La réponse par un tiers (*proxy*), en cas d'absence de la personne concernée, est toujours autorisée.

Le protocole de l'enquête Emploi est fortement différencié selon que le ménage est en première interrogation ou en réinterrogation. Avec la refonte, le protocole de première interrogation a été peu modifié ; celui de réinterrogation l'a été bien davantage.

6. Ces modules font désormais pleinement partie de l'enquête Emploi. Un cycle de huit ans a été défini, avec six thèmes récurrents : accidents au travail et problèmes de santé liés au travail, situation sur le marché du travail des immigrés et de leurs descendants, retraite et participation au marché du travail, jeunes sur le marché du travail, conciliation entre vie professionnelle et vie familiale, organisation du travail et aménagement du temps de travail.
7. Voir *https://www.insee.fr/fr/metadonnees/source/operation/s2022/processus-statistique*.

TOUJOURS UN ENQUÊTEUR EN FACE-À-FACE POUR LA PREMIÈRE INTERROGATION

Le protocole de première interrogation conserve les deux caractéristiques principales qui prévalaient jusque fin 2020 : un déplacement sur le terrain et un entretien en face-à-face[8].

L'échantillon de l'enquête Emploi comporte des résidences non principales au sens de la base de sondage (logements vacants et résidences secondaires), le statut du logement pouvant varier entre la base de sondage et la collecte. Il est donc important que l'enquêteur vérifie, en se déplaçant sur le terrain, que le logement existe toujours, qu'il est bien occupé et bien à usage d'habitation. Pendant la semaine de référence, l'enquêteur procède à cette phase de repérage et prend contact avec les personnes à enquêter. Pour l'aider, il bénéficie d'informations nouvelles : lorsqu'ils figurent dans la base de sondage, le numéro de téléphone et l'adresse mail des occupants du logement lui sont désormais mis à disposition.

La première interrogation d'un ménage reste menée par un enquêteur en face-à-face. C'est le cas des logements qui entrent dans l'échantillon, mais aussi des nouveaux ménages en cours de panel, soit que le logement était précédemment vacant, soit que le ménage n'avait pas répondu à l'enquête, soit qu'il y a eu un changement de ménage consécutif à un déménagement. Lors de la première interrogation, les concepts peuvent être précisés par l'enquêteur et le questionnaire est plus long. Ce contact avec l'enquêteur est également important pour établir la confiance avec le ménage et un engagement sur l'ensemble de la durée du panel.

Le maintien d'une première interrogation en face-à-face est un point clivant entre les pays : certains ont fait le même choix que la France, comme l'Allemagne, le Portugal ou la Belgique ; d'autres ont préféré introduire internet dès la première interrogation, contraints fortement par les coûts et après avoir réduit substantiellement la taille du questionnaire ; c'est le cas, par exemple, des Pays-Bas ou du Danemark.

En France, le principal changement dans le protocole de la première interrogation réside dans l'allongement à trois semaines de la durée de collecte. Les enquêteurs ont toujours pour consigne de réaliser le plus grand nombre d'entretiens possible en début de période, d'une part, pour limiter les biais de mémoire (la plupart des questions se rapportant à une période bien précise, la semaine de référence), d'autre part, parce que l'enquêteur se donne ainsi les moyens de parvenir à interroger les personnes les moins disponibles.

LE CHOIX ENTRE INTERNET OU LE TÉLÉPHONE POUR LES RÉINTERROGATIONS

En réinterrogation, qu'il s'agisse des réinterrogations intermédiaires ou de la dernière interrogation, le protocole évolue en revanche fortement, avec l'introduction d'internet comme mode de réponse alternatif au téléphone. L'enjeu a été de mettre au point une bonne articulation entre les deux modes pour que, dans la période de collecte limitée à trois semaines, l'ajout d'un mode se traduise au moins par un maintien du taux de collecte.

8. Du fait de la crise sanitaire, depuis mars 2020, le face-à-face a été à plusieurs reprises suspendu et remplacé temporairement par le téléphone. Dans certaines périodes où les restrictions de circulation étaient les plus strictes, les opérations de repérage n'ont pas pu être menées.

Si les grandes lignes étaient définies, plusieurs expérimentations ont été nécessaires pour affiner tous les paramètres du protocole : durée de l'exclusivité internet, supports (courrier ou mail) à utiliser pour communiquer avec les ménages, dates des relances, etc. ***(encadré 2)***.

Plus concrètement, tous les enquêtés reçoivent, le lundi de la première semaine de collecte, une lettre-avis et un mail-avis avec l'adresse du site internet de collecte[9] et leurs identifiants de connexion les invitant à répondre à l'enquête sur internet.

Dans le cas général, les trois premiers jours de la collecte sont réservés à la réponse par internet ***(figure 3)***, de façon à décharger les enquêteurs des réponses que certains enquêtés apportent sans difficulté et sans délai par eux-mêmes[10] (Garnero, 2019). C'est seulement à partir du jeudi de la première semaine de collecte que l'enquêteur peut appeler le ménage s'il n'a toujours pas répondu et lui proposer de réaliser l'enquête par téléphone. Si l'interview n'est pas réalisée lors de ce premier contact, et si l'enquêteur constate au bout de quelques jours que le questionnaire n'a toujours pas été renseigné sur internet, il rappelle l'enquêté, et ainsi de suite jusqu'à la fin des trois semaines de collecte. Le site de collecte reste accessible au ménage tant qu'il n'a pas répondu à l'enquête, quel que soit le mode, et au plus tard jusqu'au dimanche de la troisième semaine de collecte. Les ménages pour lesquels on dispose d'une adresse mail valide sont, dans le même temps, relancés par mail (jusqu'à trois relances) ; les autres le sont par courrier.

> *« Plusieurs expérimentations ont été nécessaires pour affiner tous les paramètres du protocole. »*

Dans certains cas particuliers, l'enquête par téléphone peut commencer dès le lundi de la première semaine de collecte. Il s'agit d'enquêtés identifiés comme peu susceptibles de répondre par internet[11] ou de cas où les enquêteurs savent par avance qu'ils seront absents, ce qui permet de lisser leur charge de travail. Comme auparavant, dans de rares cas laissés à l'appréciation de l'enquêteur, l'enquête pourra se réaliser en face-à-face, par exemple pour des ménages de grande taille ou comprenant mal le français.

POUVOIR GÉRER LE MULTIMODE…

Pour que la gestion du multimode se passe au mieux, l'enquêteur doit être parfaitement informé de ce qu'il doit faire et de l'état d'avancement de l'enquête. Pour cela, il dispose sur son poste informatique d'un « carnet de tournée », avec l'ensemble des logements qui lui sont confiés, et sur lequel figure la date de début de son intervention (lundi ou jeudi) ainsi que l'état d'avancement du questionnaire sur internet.

9. Le portail des enquêtes auprès des ménages par internet est accessible *via* le site *web* de l'Insee. Pour l'enquête Emploi, il est disponible à l'adresse : *https://particuliers.stat-publique.fr/eec*.
10. Les tests menés dans le cadre du projet Muse ont montré que plus de 10 % des enquêtés répondent dès les premiers jours, et que la première relance par mail est la plus efficace lorsqu'elle est envoyée en fin de semaine. Le cadencement des relances soutient la réponse sur internet tout au long des trois semaines de collecte.
11. Ces ménages ont répondu par téléphone à l'interrogation précédente, n'ont pas communiqué d'adresse mail, et ont répondu par la négative à la question *« Auriez-vous eu l'équipement et les connaissances nécessaires pour répondre sur internet ? »*. Ils reçoivent tout de même des identifiants de connexion dans la lettre-avis.

Figure 3. Trois semaines en réinterrogation pour répondre par internet ou par téléphone

Lorsqu'il entre en contact avec le ménage, l'enquêteur est ainsi en mesure de savoir s'il doit l'inciter à terminer le questionnaire s'il est bien avancé sur internet, ou s'il est préférable de faire ou de reprendre avec lui le questionnaire par téléphone[12].

Les enquêtés qui rencontrent des difficultés liées à l'enquête (problèmes de connexion, difficultés pour répondre à une question, etc.) ou s'interrogent sur ses objectifs, sa durée, les réinterrogations futures, etc. peuvent contacter un service d'assistance, par mail ou par téléphone. Une difficulté couramment rencontrée par les ménages est l'oubli ou la perte de leur mot de passe. Dans ce cas, un nouveau mot de passe peut être obtenu sans même passer par l'assistance en temps réel depuis le portail de l'enquête pour les enquêtés qui ont communiqué une adresse mail ; pour les autres, une procédure d'authentification est mise en œuvre.

… ET LIMITER LA CHARGE DES ENQUÊTÉS COMME DES ENQUÊTEURS

Ce protocole multimode a pour objectif de favoriser la réponse des ménages, car ceux-ci utilisent de plus en plus volontiers internet dans leurs démarches de la vie quotidienne ou pour répondre aux enquêtes. Il a également pour but d'alléger la charge des enquêteurs pour qui les enquêtes téléphoniques sont vécues comme répétitives. Cependant, même en réinterrogation, le rôle des enquêteurs reste crucial pour obtenir la réponse de ménages moins familiers d'internet ou plus sensibles à la relance téléphonique.

> *« Même en réinterrogation, le rôle des enquêteurs reste crucial pour obtenir la réponse de ménages moins familiers d'internet ou plus sensibles à la relance téléphonique. »*

L'enquête étant un panel relativement lourd (six interrogations séparées seulement d'un trimestre), un enjeu important est de limiter au maximum la charge d'enquête pour les ménages. Cela passe par un travail sur le questionnaire, comme évoqué *supra*, mais aussi par la variété des modes de réponse possibles, adaptés aux aspirations des ménages.

Certaines populations se sentent moins concernées par l'enquête et seraient difficiles à fidéliser sur toute la durée du panel alors même que leur situation ne change pas. Il s'agit des seniors et des personnes en situation de handicap. Eurostat a proposé des règles simplifiées dans leur cas : dès la première interrogation, les personnes de 90 ans ou plus sont sorties du champ de l'enquête ; les personnes inactives âgées de 70 ans ou plus ainsi que les personnes inactives en situation de handicap[13] ne sont quant à elles pas réinterrogées et leurs réponses de la première interrogation sont « recopiées ».

Enfin, pour alléger la charge des enquêteurs, les résidences secondaires et les logements non ordinaires, qui ont peu de chances de changer d'affectation sur la durée du panel, sont sortis de l'échantillon une fois que leur statut a été confirmé sur le terrain. En revanche, les logements identifiés comme vacants, dont une part non négligeable sont susceptibles de devenir des résidences principales sur la durée du panel, sont remis en collecte chaque trimestre.

12. Il n'est pour l'instant pas possible de finir par téléphone un questionnaire débuté par internet.
13. En France, des conditions d'âge ou d'exercice passé d'une activité ont été ajoutées.

Encadré 2. Les principaux enseignements des expérimentations de Muse

Le projet *Multimode sur l'enquête Emploi* (Muse) visait à expérimenter le recours à internet lors de la collecte de l'enquête Emploi (Garnero, 2019). En tout, sept expérimentations ont été menées entre 2014 et 2018 :

- pour mettre au point un questionnaire « fluide » pour la version internet (2014-2015) ;
- pour tester le fonctionnement à grande échelle sur les serveurs de l'Insee, avec un échantillon de 30 000 ménages interrogé à trois reprises exclusivement sur internet (2016) ;
- et pour tester plusieurs variantes de protocole multimode internet / téléphone (2017-2018).

Les tests ont confirmé que la première interrogation doit être maintenue en face-à-face :

- parce qu'il est préférable que les différents concepts soient introduits par un enquêteur ;
- parce que le temps de passation du questionnaire n'est pas adapté à une réponse sur internet.

En revanche, **en réinterrogation, le temps de réponse pour un questionnaire individuel est bien plus court.** Il est deux fois plus long par internet que par téléphone, mais ce n'est pas forcément perçu comme tel par les enquêtés.

Comme pour le recensement, les ménages les plus susceptibles de répondre sur internet sont plus jeunes, plus aisés et plus « connectés », au sens où ils ont communiqué plus souvent une adresse mail à l'administration fiscale.

Sur internet, les enquêtés répondent plus souvent à leur propre questionnaire individuel que par téléphone ou en face-à-face. Tout se passe comme si les membres du ménage se passaient le lien sur le questionnaire, alors qu'au téléphone il est fréquent qu'une personne réponde pour tout ou partie des membres du ménage.

Lors de l'expérimentation de grande ampleur de 2016 (avec obligation de réponse), le taux de réponse en première interrogation a été de 31 % ; 57 % ont ensuite répondu en 2e interrogation, et parmi ces derniers 78 % en 3e interrogation. Lors du test (non obligatoire) du multimode internet / téléphone, le taux de réponse des grappes pour lesquelles le test avait pu se dérouler sans incident informatique majeur a été de 35 % sur internet et 26 % par téléphone.

Les réponses sur internet sont plutôt concentrées sur le début de la période de collecte.

Les lettres-avis sont plus efficaces que les mails-avis. L'impact d'une relance papier diffère en revanche peu de celui d'une relance par mail.

Les enquêteurs ont bien accueilli le multimode internet / téléphone lors de ces tests, appréciant de voir diminuer la charge jugée trop répétitive des réinterrogations téléphoniques. La période initiale réservée à la réponse sur internet décharge les enquêteurs et permet de mieux cibler leur intervention.

La période de trois semaines est jugée suffisante pour basculer d'un mode de collecte à l'autre. Les relances téléphoniques pour inciter les ménages à répondre sur internet ne semblent pas opportunes : lorsque l'enquêteur contacte une personne par téléphone, le plus efficace est de lui proposer de réaliser l'enquête immédiatement par téléphone.

POURQUOI UNE ENQUÊTE PILOTE D'ENVERGURE DÈS 2020 ?

L'expérience de la précédente refonte de l'enquête Emploi, en 2013, lors de laquelle un simple « toilettage » du questionnaire ne devait pas avoir d'impact notable sur les indicateurs, a marqué les esprits. La surprise a en effet été totale lorsque les premières exploitations ont mis en évidence des ruptures sur les indicateurs. Après moult vérifications, une fois le bon fonctionnement de l'ensemble de la chaîne de production confirmé, il a fallu se rendre à l'évidence : les modifications introduites, notamment dans l'ordre des questions sur le souhait de travailler et la recherche d'emploi, avaient bien provoqué une rupture dans les séries de taux de chômage. La refonte de 2021, qui devait remodeler en profondeur le questionnaire, modifier le protocole et rénover la méthode de pondération, avait toutes les chances de provoquer elle aussi une rupture sur les principaux indicateurs du marché du travail. Il s'imposait donc de se donner les moyens de mener une opération préparatoire, visant à s'assurer du bon fonctionnement des outils (application de gestion, architecture informatique adaptée au multimode), à préparer la chaîne de traitements aval et à fournir une information suffisamment précise pour estimer au mieux les ruptures de série et pouvoir rétropoler les séries.

Conscient de la nécessité d'éclairer les utilisateurs sur les changements affectant ces indicateurs, au premier rang desquels le taux de chômage, Eurostat a fortement incité les États membres à mettre en place des dispositifs ambitieux, leur permettant d'estimer la rupture occasionnée par la mise en œuvre du nouveau règlement pour proposer des séries rétropolées. Les orientations européennes rejoignaient ainsi les objectifs que se donnait l'Insee pour cette nouvelle refonte.

UN PILOTE NÉ D'UN JEU DE CONTRAINTES

Dès 2017-2018, l'Insee s'est engagé dans la conception de cette opération préparatoire qui devait prendre en compte tout un ensemble de contraintes.

La première était l'ampleur de la refonte envisagée, qui affectait plusieurs dimensions de l'enquête : il n'était pas possible, comme en 2013, de procéder à une déconstruction analytique du questionnaire ; il fallait mener une enquête complète, sur le terrain.

La deuxième contrainte concernait le calendrier : elle portait sur la date d'entrée en vigueur du nouveau règlement, 2021, sur le fait que l'Insee envisageait de garder inchangé le rythme trimestriel de ses publications et sur le fait que l'expérience de 2013 avait montré que les impacts pouvaient varier au fil de l'année, sous l'effet de comportements saisonniers, en particulier pendant l'été. Il a donc été retenu de mener l'opération préparatoire en 2020 et de couvrir l'ensemble des trimestres. Cela imposait un calendrier de préparation très serré, d'autant plus que des paramètres importants dépendaient de décisions européennes qui étaient alors encore en cours.

La troisième contrainte, qui devait déterminer la taille de l'opération, devait concilier deux exigences contradictoires : une exigence de précision, pour être en mesure de détecter des impacts significatifs, et une exigence de limitation du coût. Il a ainsi été décidé de ponctionner un quart de l'échantillon de l'enquête en production pour réaliser le Pilote.

Enfin, le Pilote devait servir à s'assurer que l'ensemble de la mécanique de la nouvelle enquête fonctionnait et à estimer les ruptures de série introduites par la nouvelle enquête. Il devait donc jouer à l'avance exactement ce qui allait se passer lors de l'entrée en vigueur de la nouvelle enquête et être parfaitement conforme à ce que serait la nouvelle enquête, en termes de questionnaire, de protocole ou encore de méthode de traitement aval.

Au final, le Pilote comporte deux composantes ***(figure 4)*** :

- une **bascule anticipée en nouvelle enquête** : un quart de l'échantillon, sur l'ensemble des rangs d'interrogation, a basculé en nouvelle enquête dès le 1er trimestre 2020. Cette bascule préfigurait celle opérée au 1er trimestre 2021 sur le reste de l'échantillon[14] ;
- un **sur-échantillon en ancienne enquête** : un échantillon supplémentaire, introduit progressivement à partir du 4e trimestre 2019, a été réinterrogé dans l'ancienne enquête jusqu'au 1er trimestre 2021 inclus. On disposait ainsi d'un échantillon témoin complet (rangs 1 à 6) en ancienne enquête au 1er trimestre 2021, quand la nouvelle enquête entrait en production. Ce sur-échantillon a également été utilisé pour la diffusion en 2020, compensant en partie la ponction effectuée pour la composante « bascule ».

Au 1er trimestre 2021, la partie qui avait basculé un an plus tôt a été complétée par les trois quarts restants qui étaient restés en ancienne enquête et ont basculé à leur tour. L'échantillon de l'enquête retrouvait ainsi sa « pleine » taille.

ESTIMER LA RUPTURE DE SÉRIE ET PRÉPARER LA RÉTROPOLATION

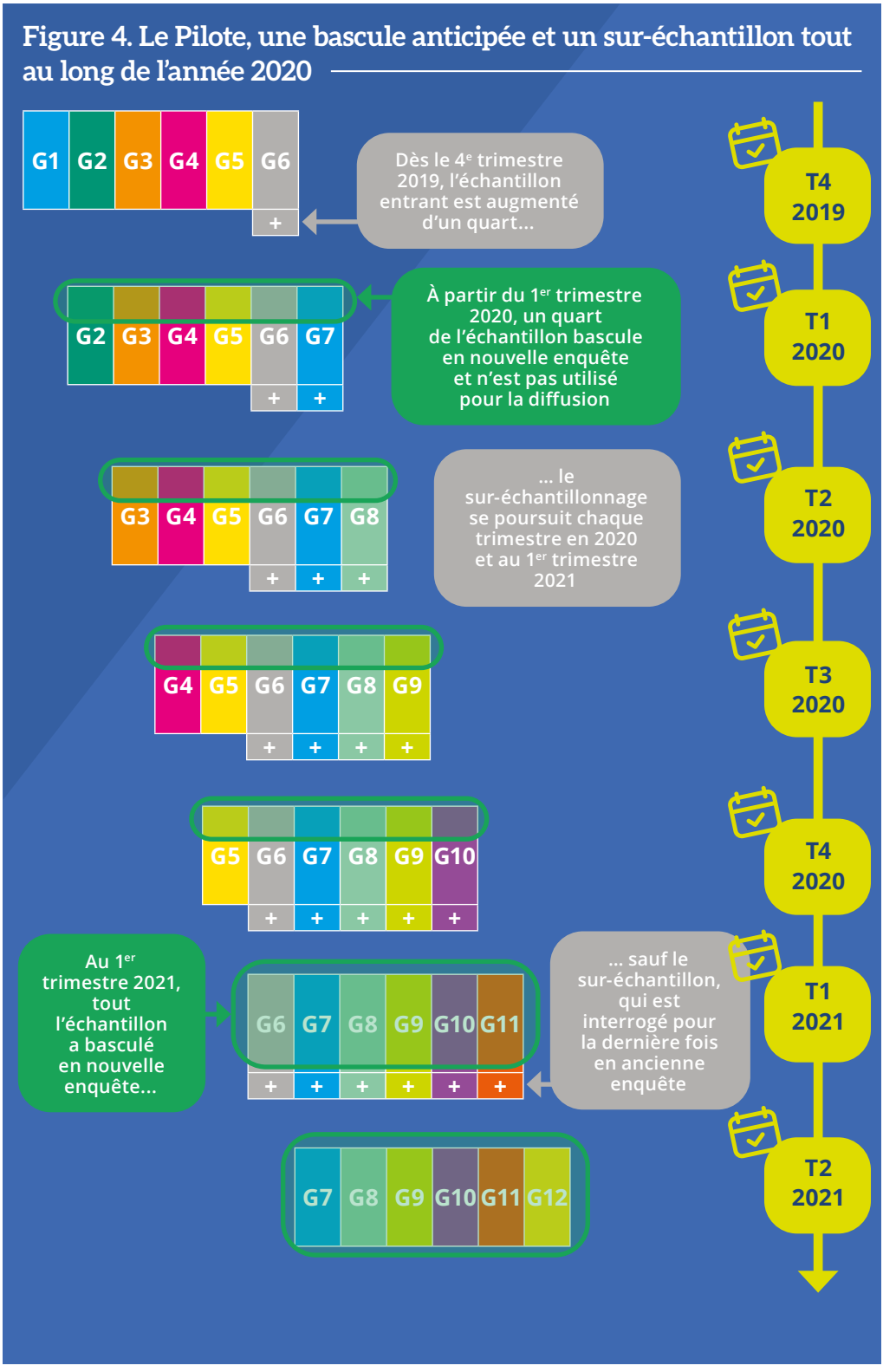
L'objectif principal du Pilote était de fournir un ensemble d'informations permettant d'estimer le plus précisément possible les ruptures de série induites par la nouvelle enquête. La difficulté majeure pour estimer les ruptures de série était la marge d'incertitude importante qui entoure les estimations : sur l'ancienne enquête, malgré la taille importante de l'échantillon, le taux de chômage trimestriel est estimé avec une précision de ± 0,3 point. Avec le dispositif du Pilote, en comparant les indicateurs mesurés *via* l'ancienne enquête et *via* la nouvelle enquête, on disposait de cinq mesures trimestrielles des ruptures de série, du 1er trimestre 2020 au 1er trimestre 2021. Disposer de plusieurs trimestres permettait de mesurer la rupture de série en moyenne annuelle, avec une précision accrue, mais également de détecter une éventuelle saisonnalité dans les impacts. Au vu de la marge d'incertitude importante attendue pour les estimations, seuls des changements de saisonnalité importants pouvaient cependant être identifiés. Enfin, la crise sanitaire de 2020 a rendu plus compliquée la mesure des ruptures de série, cette dernière pouvant avoir des impacts différents sur les deux enquêtes, l'ancienne et la nouvelle.

En raison de la multiplicité des sources de rupture de série et des possibles effets croisés, il a été privilégié d'estimer de façon globale la rupture de série par comparaison des deux versions de l'enquête, sans chercher à quantifier systématiquement la contribution propre à chaque évolution. Malgré tout, il a été possible d'isoler certains effets propres, en ce qui concerne par exemple des changements conceptuels ou dans la méthode de pondération (Insee, 2021). Isoler les effets liés au changement de questionnaire ou les effets de mode liés à la réponse par internet (Vinceneux, 2018), qui combinent effet « de mesure » (le fait qu'une personne réponde différemment sur internet ou à un enquêteur) et effet « de sélection » (le fait que la réponse par internet permette de capter de nouveaux répondants), est plus complexe et pourra faire l'objet de travaux ultérieurs.

L'objectif était d'obtenir une estimation des ruptures de série pour rétropoler les principales séries dès la publication des résultats du 1er trimestre 2021, afin de pouvoir communiquer sur des évolutions entre le 4e trimestre 2020 et le 1er trimestre 2021 qui ont du sens.

14. En raison du caractère rotatif de l'échantillon et de sa taille, le passage de l'ancienne enquête à la nouvelle a été fait sous la forme d'une bascule (et non d'une montée en charge progressive), à l'image de ce qui avait été fait en 2013.

Figure 4. Le Pilote, une bascule anticipée et un sur-échantillon tout au long de l'année 2020



BIBLIOGRAPHIE

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 11 juin 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497076/courstat-4-5.pdf*.

CASES, Chantal, 2019. IESS : l'Europe harmonise ses statistiques sociales pour mieux éclairer les politiques. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 125-139. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254233/courstat-3-10.pdf*.

DURIEUX, Bruno, DE NANTEUIL, Yann, RÉMOND, Sébastien, DU MESNIL DU BUISSON, Marie-Ange, GRIVEL, Nicolas et WANECQ, Thomas, 2007. *Rapport sur les méthodes statistiques d'estimation du chômage.* [en ligne]. 1er septembre 2007. Inspection générale des Finances, Inspection générale des Affaires sociales. N° 2007-M-066-01 et RM 2007-141. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.vie-publique.fr/rapport/29333-les-methodes-statistiques-destimation-du-chomage*.

EUROSTAT, 2018. *Quality issues regarding the measurement of working time with the Labour Force Survey (LFS).* [en ligne]. 13 mars 2018. Statistical reports. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/fr/web/products-statistical-reports/-/KS-FT-17-005*.

EUROSTAT, 2019. *Labour Force Survey in the EU, candidate and EFTA countries — Main characteristics of national surveys 2018.* [en ligne]. 15 novembre 2019. Statistical reports. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/fr/web/products-statistical-reports/-/KS-FT-19-008*.

GARNERO, Marguerite, 2019. *Le projet Muse : 5 ans d'expérimentations pour préparer l'introduction d'Internet dans l'enquête Emploi.* [en ligne]. 11 décembre 2019. Insee. Documents de travail, n° F1907. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4263350*.

GAZIER, Bernard, MINNI, Claude et PICART, Claude, 2016. *La diversité des formes d'emploi.* [en ligne]. Juillet 2016. Cnis, rapport de groupe de travail, n°142. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2016_142_diversite_forme-demploi.pdf*.

GOUX, Dominique, 2003. Une histoire de l'Enquête Emploi. In : *Économie et Statistique*. [en ligne]. 1er juillet 2003. Insee. N°362, pp 41-57. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/1376194/es362c.pdf*.

INSEE, 2021. *L'enquête Emploi se rénove en 2021 : des raisons de sa refonte aux impacts sur la mesure de l'emploi et du chômage.* [en ligne]. 29 juin 2021. Insee Analyses, n° 65. [Consulté le 29 juin 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5402123*.

ORGANISATION INTERNATIONALE DU TRAVAIL (OIT), 1982. *Résolution concernant les statistiques de la population active, de l'emploi, du chômage et du sous-emploi*. [en ligne]. 1er octobre 1982. Treizième Conférence internationale des statisticiens du travail. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.ilo.org/global/statistics-and-databases/standards-and-guidelines/resolutions-adopted-by-international-conferences-of-labour-statisticians/WCMS_087482/lang--fr/index.htm*.

ORGANISATION INTERNATIONALE DU TRAVAIL (OIT), 2018. Résolution concernant les statistiques sur les relations de travail. In : *Vingtième Conférence internationale des statisticiens du travail.* [en ligne]. 10-19 octobre 2018. Genève. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://ilostat.ilo.org/fr/about/standards/icls/icls-documents/*.

RAZAFINDRANOVONA, Tiaray, 2015. *La collecte multimode et le paradigme de l'erreur d'enquête totale*. [en ligne]. 27 mars 2015. Insee. Documents de travail, Méthodologie statistique, n°M2015/01. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1381054*.

SIGNORE, Marina, 2019. *Mixed-Mode Designs for Social Surveys – MIMOD. Final methodological report summarizing the results of WP 1-5.* [en ligne]. 26 mars 2019. Eurostat et Istat. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.istat.it/it/files//2011/07/MIMOD-project-Final-report-WP1-WP5.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 11 juin 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497081/courstat-4-6.pdf*.

VINCENEUX, Klara, 2018. *Mode de collecte et questionnaire, quels impacts sur les indicateurs européens de l'enquête Emploi ?* [en ligne]. 4 octobre 2018. Insee. Documents de travail, Direction des Statistiques Démographiques et Sociales, n°F1804. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/version-html/3625351/F1804.pdf*.

FONDEMENTS JURIDIQUES

Règlement 577/98 du Conseil du 9 mars 1998 relatif à l'organisation d'une enquête par sondage sur les forces de travail dans la Communauté. In : *Office des publications de l'Union européenne.* [en ligne]. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://op.europa.eu/fr/publication-detail/-/publication/a5ec350e-ad3c-4936-9fbf-5f63560cce0d*.

Règlement 2019/1700 du Parlement européen et du Conseil du 10 octobre 2019 établissant un cadre commun pour des statistiques européennes relatives aux personnes et aux ménages fondées sur des données au niveau individuel collectées à partir d'échantillons. In : *site EUR-Lex.* [en ligne]. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=OJ:L:2019:261I:FULL&from=FR*.

Règlement d'exécution 2019/2240 de la Commission du 16 décembre 2019 spécifiant les éléments techniques de l'ensemble de données, établissant les formats techniques de transmission des informations et spécifiant les modalités et le contenu détaillés des rapports de qualité concernant l'organisation d'une enquête par sondage dans le domaine de la main-d'œuvre conformément au règlement (UE) 2019/1700 du Parlement européen et du Conseil. In : *site EUR-Lex.* [en ligne]. [Consulté le 11 juin 2021]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32019R2240&from=EN*.

FIDÉLI, L'INTÉGRATION DES SOURCES FISCALES DANS LES DONNÉES SOCIALES

Pierre Lamarche et Stéfan Lollivier***

Les fichiers fiscaux se sont imposés dans la statistique publique comme des sources précieuses dont l'usage permet de satisfaire différents besoins du système d'information, à condition de faire l'objet d'un travail ambitieux de mise en cohérence. Le Fichier démographique sur les logements et les individus (Fidéli) produit par l'Insee à partir des données fiscales reflète bien ce potentiel associé à cette exigence ; en venant s'insérer en complément de sources telles que le recensement de la population, il permet d'éclairer des aspects des conditions de vie des individus jusqu'alors peu documentés, tant au niveau national qu'au niveau local. Ainsi, l'information finement localisée des individus permet de dresser un état des lieux plus complet, y compris au niveau infra-communal. Cette insertion se fait au prix d'un processus de production lourd et exigeant, et s'accompagne naturellement de contraintes d'utilisation assez élevées afin de préserver le secret statistique.

Pour rendre durable la présence de cette source dans le système d'information de la statistique publique, il faut dans un premier temps relever le défi posé par les réformes fiscales en cours, en particulier la suppression de la taxe d'habitation ; pour accroître son utilité, il faudra également poursuivre l'effort d'enrichissement, afin de tirer toujours plus parti de la profondeur de l'information fiscale. Enfin, Fidéli doit participer à terme à la mise en cohérence d'un ensemble encore plus vaste de sources administratives, et pas seulement fiscales.

Tax data have turned out as valuable sources for official statistics, as they fulfil a lot of needs of the information system, as long as they are produced through a demanding effort of data compilation. The Fichier Démographique des Logements et des Individus (Fidéli) ***produced by INSEE from fiscal data has proven the point. Usefully complementing traditional sources such as Census, it makes it possible to enlighten aspects of individuals' living conditions so far poorly documented, at national and at local level as well. Hence, the precisely geolocalized data makes it possible to renew the analysis also below municipality level. This completion is performed at the cost of a highly demanding and comprehensive production process while it also comes with strong dissemination conditions, so as to avoid unforeseen information disclosures.***

In order to make this statistical source sustainable, one should first address data issues posed by the ongoing fiscal reforms in France, especially the abolition of the Housing Tax. Second, one should enlarge its scope and encompass a always wider range of information coming from fiscal sources. Third and last, in the long run the source should fit in a more ambitious and comprehensive process involving larger parts of administrative – and not only fiscal – data.

* Chef de la division Logement, Insee,
pierre.lamarche@insee.fr
** Expert auprès du directeur général, Programme Répertoire des logements, Insee,
stefan.lollivier@insee.fr

L'administration fiscale, dans son activité de recouvrement de l'impôt, collecte à cette fin une grande quantité d'informations sur les individus et les logements. Ces données présentent un intérêt naturel pour le statisticien. Ainsi, l'usage des sources fiscales dans le système d'information des statistiques démographiques et sociales s'est imposé au cours de ces dernières décennies, avec en point d'orgue la production annuelle d'un ensemble de données permettant de répondre à des besoins à la fois très différents et très précis : le **Fichier démographique sur les logements et les individus (Fidéli)**.

La mise en cohérence des sources à l'origine de ces données illustre bien le processus qui permet de passer de données purement administratives à une information statistique de grande valeur pour le Service statistique public. Fidéli réalise en effet un traitement mutualisé des sources fiscales sur les logements et les personnes, et permet ainsi de mettre ces données à la disposition des chargés d'étude, des services en charge de l'échantillonnage des enquêtes, etc. Ces utilisateurs disposent chacun de « livrables » spécifiques ***(figure 1 et encadré 1)*** qui bénéficient de la même mise en cohérence des données administratives pour des finalités statistiques.

L'USAGE DES SOURCES ADMINISTRATIVES, DES PRATIQUES HÉTÉROGÈNES EN EUROPE

Si un observateur dressait un inventaire des différentes sources d'information utilisées par les instituts nationaux statistiques à travers le monde pour décrire les populations, il s'étonnerait de leur grande diversité. En Europe en particulier, il constaterait de très fortes différences d'un pays à l'autre.

> *« Les pays nordiques (Danemark, Finlande, Suède, Norvège et Islande) construisent l'essentiel de leurs statistiques démographiques et sociales sur le recours intensif à des registres de population. »*

Les pays nordiques (Danemark, Finlande, Suède, Norvège et Islande) construisent l'essentiel de leurs statistiques démographiques et sociales sur le recours intensif à des registres de population, à tel point que le recensement de population se fait généralement sur cette base (Unece, 2007 ou Statistics Finland, 2004). Mais au-delà du dénombrement de population, les administrations de ces pays collectent et mettent à jour un spectre très large d'informations sur l'ensemble de la population, ce qui leur permet de produire des analyses extrêmement variées. Une des clés de la réussite de ce modèle repose bien souvent dans l'adoption d'identifiants uniques pour repérer les unités d'intérêt – ici les individus et les logements – afin de faciliter les appariements entre les différentes sources. Très généralement, un tel système repose également sur deux piliers essentiels qui vont d'ailleurs assez naturellement de pair : un contexte législatif favorable, voire incitatif, ainsi qu'un grand degré d'acceptation et de confiance de la part de la population vis-à-vis d'un tel processus et des institutions qui sont les garantes de son bon usage.

De l'autre côté du spectre, de nombreux pays ont un recours encore assez limité aux registres et autres sources administratives, pour des raisons qui peuvent être tout autant d'ordre historique que d'ordre pratique. Ainsi, la constitution d'un registre, régulièrement actualisé, suppose la mise en place d'une infrastructure administrative pérenne et uniforme sur l'ensemble du territoire national ; un exemple très connu à ce sujet est l'absence de cadastre en Grèce.

De manière générale, les pays européens, en dehors de l'exemple nordique, ont pendant longtemps privilégié les techniques de sondage et les données d'enquêtes pour pouvoir éclairer les conditions de vie de la population ; les recensements ont été longtemps exclusivement basés sur de grandes opérations décennales de dénombrement de la population. De fait, la longue tradition de social-démocratie qui prédomine depuis longtemps dans les pays nordiques apporte un élément explicatif de l'usage intensif que font ces pays des registres pour éclairer les conditions sociales de leur population.

EN FRANCE, UNE LENTE ÉVOLUTION MARQUÉE PAR LES IMPÉRATIFS DU RECENSEMENT

La France ne dispose pas de registre de population, et en matière de croisement de données d'origine administrative, l'administration et la statistique publique françaises ont longtemps été marquées par les polémiques dans les années 1970 autour du fichier Safari (Boucher, 1974)[1] ; cette affaire illustre bien la nécessité des deux piliers mentionnés plus haut comme pré-requis à un usage statistique des sources administratives, usage qui passe naturellement par l'appariement des différents registres. Le rejet par l'opinion publique d'un projet d'interconnexion de différents fichiers relatifs à la population a constitué par la suite un motif pour un recours assez exclusif à la collecte par enquête(s) : au fil du temps, l'Insee s'est constitué une expertise reconnue en la matière, permettant la description des grands phénomènes sociaux qui caractérisent la population française.

Les grandes enquêtes statistiques de l'Insee se sont historiquement adossées au recensement de population, qui permettait de constituer tous les dix ans environ une base de sondage exhaustive des logements, ménages et individus vivant en France. Les informations collectées à l'occasion de ces grandes opérations décennales étaient naturellement limitées, et la base de sondage était généralement complétée par des typologies socio-économiques des quartiers et communes de manière à accroître les possibilités de stratification. Ce modèle présentait alors un fort défaut d'actualisation, les échantillons sélectionnés alors en fin de cycle dans une base de données remontant à une dizaine d'années souffrant de problèmes liés entre autres aux mobilités résidentielles depuis la date du dernier recensement. La construction neuve était prise en compte en sélectionnant un échantillon complémentaire dans la base de logements neufs (Sitadel), mais sans connaissance des occupants de ces logements.

Le passage en 2004 à un recensement rotatif annuel a constitué de ce point de vue une amélioration substantielle puisqu'il a permis de constituer des bases de sondage renouvelées annuellement. Pour autant, l'usage du nouveau recensement de population comme base de sondage a également posé de nouveaux défis qu'il a fallu résoudre, comme le fait que le processus de recensement résultait lui-même d'un processus d'échantillonnage. Au-delà de ces questions méthodologiques, les unités primaires – c'est-à-dire les zones géographiques dans lesquelles sont sélectionnés aléatoirement les unités à enquêter – étaient menacées

1. Voir (Sénat, 2021) : « *En 1974, le ministère de l'Intérieur avait bâti un fichier informatisé au nom évocateur : SAFARI, acronyme de système automatisé pour les fichiers administratifs et le répertoire des individus. Ce système prévoyait de créer une base de données centralisée de la population, en utilisant le fichier de sécurité sociale comme identifiant commun à tous les fichiers administratifs. Devant le tollé généralisé provoqué par ce projet – le journal Le Monde allant jusqu'à titrer « SAFARI, ou la chasse aux Français » –, le Premier ministre de l'époque […] ne put que le retirer et créer dans la foulée une commission dite Informatique et liberté, chargée de proposer une réglementation sur l'utilisation des moyens informatiques* ».

peu à peu d'épuisement ; la représentativité des échantillons sélectionnées s'en est trouvée petit à petit affectée[2]. En outre, la rotation annuelle des échantillons en petites communes entraînait pour les enquêteurs des déplacements plus nombreux (Sillard *et alii*, 2020).

Parallèlement, la numérisation des données administratives, et en particulier fiscales, allait offrir de nouvelles opportunités. Dès 1996, l'Insee a pu disposer de la totalité des données fiscales relatives à la taxe d'habitation et à l'impôt sur le revenu. Une fois celles-ci

Encadré 1. Les livrables de Fidéli

Les produits issus de Fidéli sont multiples, et peuvent même faire l'objet d'une production ponctuelle lorsque le besoin s'en fait sentir. Les données produites de manière récurrente sont les suivantes :

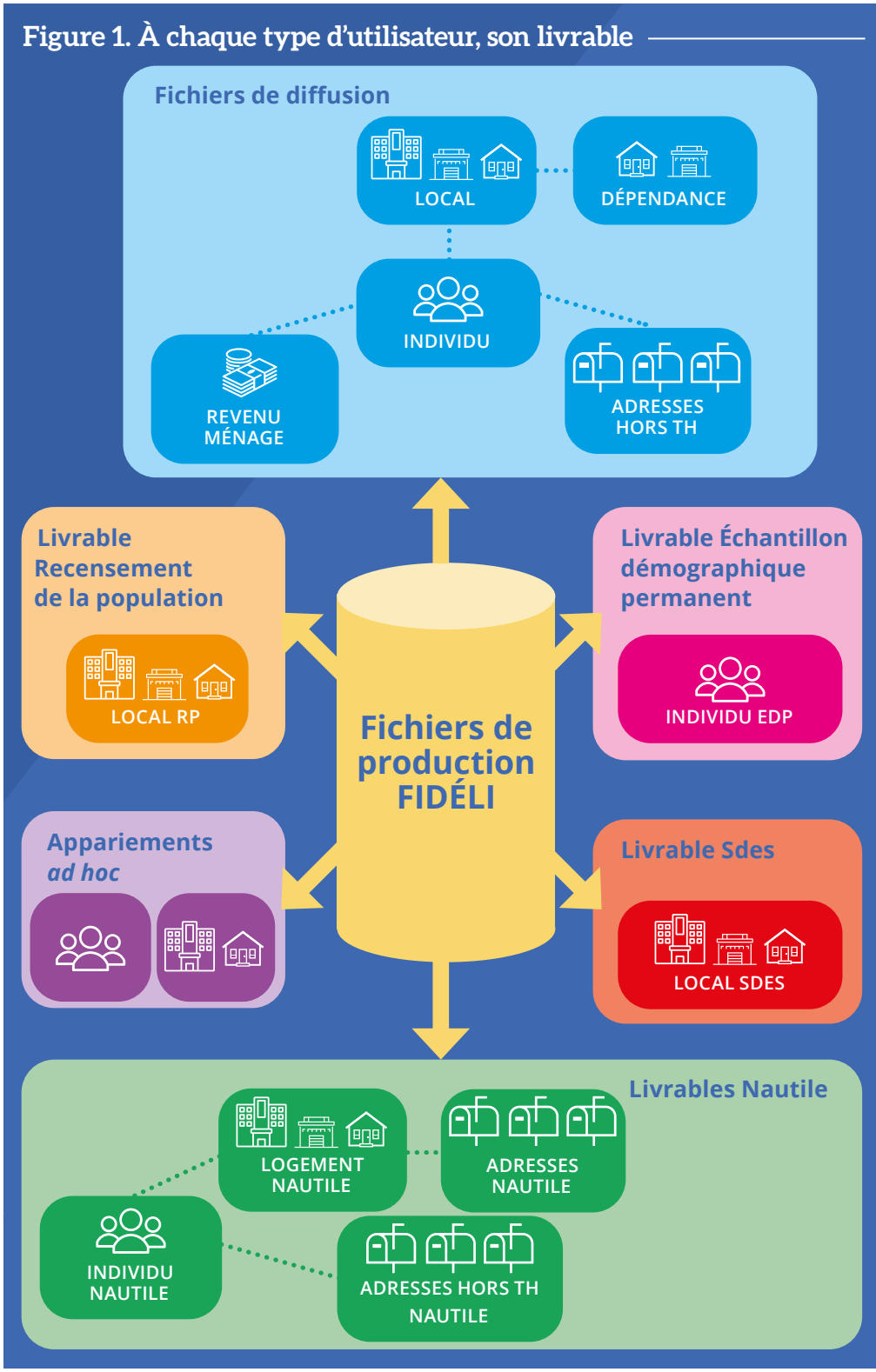
- **les fichiers de diffusion livrés au CASD***, qui sont les plus utilisés, sont constitués de cinq grandes tables, une table sur les locaux, une table sur les dépendances, une table sur les individus, une table sur les revenus des ménages localisés (par Filosofi) dans chaque logement recensé dans la table des locaux, ainsi qu'un complément, les adresses hors taxe d'habitation, qui correspond à l'ensemble des logements pour lesquels on retrouve des individus dans la source fiscale sans qu'ils ne soient connus à la taxe d'habitation. L'information contenue dans ces fichiers, très dense et précise, permet un grand nombre d'analyses, y compris au niveau local et constitue l'essentiel du socle sur lequel peuvent s'appuyer les utilisateurs de Fidéli. Comme il s'agit de fichiers de diffusion, l'information est anonymisée (au sens où l'ensemble des variables nominatives contenues dans les sources originelles sont supprimées), mais celle-ci demeure très largement identifiante, compte tenu de l'exhaustivité et de la précision de la source. Leur accès se fait dans un cadre très contraint, avec un examen des projets par le Comité du Secret, et une mise à disposition *via* le CASD pour les utilisateurs en dehors du Service statistique public, pour laquelle les analyses des données réalisées font l'objet d'un contrôle de confidentialité. Il s'agit des seuls fichiers mis à disposition en dehors du SSP ;
- **le livrable Nautile**, qui constitue la base de sondage des enquêtes auprès des ménages, contient l'information nécessaire pour la constitution des plans de sondage et l'établissement de fiches-adresses des individus ou logements échantillonnés. La seule finalité de ces fichiers est la constitution d'échantillons ; ils ne font pas l'objet de diffusion en dehors de la collecte de certaines enquêtes auprès des ménages, très encadrées par des conventions avec les services fiscaux (DGFiP) ;
- **un livrable spécifique pour le Sdes****, qui autorise un chaînage sur cinq ans (au lieu de deux dans le cas des fichiers de diffusion) des données sur les logements, de manière à reproduire une information similaire à celle produite par le fichier Filocom ;
- un **livrable permettant d'alimenter l'Échantillon démographique permanent** ;
- un **livrable utilisé par le recensement de population** pour anticiper la charge de la collecte dans les parties du territoire concernées par la prochaine Enquête annuelle de recensement (EAR).

** Centre d'accès sécurisé aux données (Gadouche, 2019).*

*** Services des données et études statistiques, service statistique des ministères chargés de l'environnement, de l'énergie, de la construction, du logement et des transports.*

2. Ce d'autant que l'Insee a acquis une mission de service public en matière d'échantillonnage pour les enquêtes auprès des ménages : pour toute enquête ayant reçu le label d'intérêt statistique, l'Insee doit assurer le tirage de l'échantillon, généralement dans Nautile.

Figure 1. À chaque type d'utilisateur, son livrable



appariées avec l'enquête sur l'Emploi, l'Insee a pu annualiser la production de l'enquête sur les revenus fiscaux et répondre ainsi à une demande accrue d'informations sur les revenus des personnes et les inégalités.

L'utilisation à partir de 2009 des fichiers de la taxe d'habitation comme base de sondage de cette même enquête Emploi constituait alors la preuve que les sources fiscales peuvent représenter une alternative au recensement pour l'échantillonnage des enquêtes ménages. Le recours dans le même temps aux sources fiscales pour la description exhaustive, y compris au niveau local, de la distribution des revenus dans la population française[3], et les outils de diffusion des résultats sous forme de données carroyées, ont renforcé la prise de conscience du potentiel des sources fiscales.

UN CONTEXTE DE PLUS EN PLUS PORTEUR

Dans le même temps, le contexte européen a vu l'émergence des sources administratives comme élément structurant du Système statistique européen (SSE), et particulièrement en France où les propriétés du recensement de population sous sa forme rotative pouvaient constituer un handicap (voir *supra* au sujet de la nature non-exhaustive du recensement rotatif). L'initiative GEOSTAT1 (Eurostat, 2018) a abouti ainsi à la diffusion de données de population sur l'ensemble du territoire européen à l'échelle de carreaux de 1 kilomètre de côté. La volonté d'enrichir la diffusion de ces données carroyées avec de nouvelles variables, par exemple sur les revenus, consacre l'apport des sources administratives comme complément précieux aux sources traditionnelles de la statistique publique. De ce point de vue, le mouvement est global, tant en Europe, où des pays comme l'Espagne ou l'Allemagne (Bens et Schukraft, 2019) font également de plus en plus appel aux données administratives, mais également au Canada avec l'Environnement de couplage de données sociales (Trainor et Trudeau, 2015), les Pays-Bas et le *System of social Statistical Datasets* (Bakker *et alii*, 2014), la Nouvelle-Zélande (Statistics New Zealand, 2014) ou encore l'Australie avec le *Multi-Agency Data Integration Project* (ABS, 2021).

> *« La loi pour une République numérique adoptée en 2016 a renforcé la possibilité de recours aux données sous forme numérique pour le besoin des enquêtes statistiques. »*

Comme en Allemagne, le contexte législatif français accompagne le mouvement d'un recours plus massif aux sources administratives. La loi de 1951 sur l'Obligation, la coordination et le secret en matière de statistiques[4] garantissait déjà de longue date l'accès pour le SSP aux données administratives. Mais la loi pour une République numérique adoptée en 2016 a renforcé la possibilité de recours aux données sous forme numérique pour le besoin des enquêtes statistiques. En outre, elle désigne l'Insee comme tiers de confiance[5] dans le processus d'appariement entre différentes sources administratives. Le cadre législatif reste néanmoins assez contraignant, en particulier sur la question du principe de minimisation[6], qui impose une durée de conservation des identifiants très courte, relativement aux besoins potentiels.

3. Dans le cadre de Revenus disponibles localisés (RDL) puis Filosofi (voir *infra*).
4. Voir les références juridiques en fin d'article.
5. Pour plus de précisions sur la notion de tiers de confiance, voir (Gadouche, 2019).
6. Il s'agit de l'obligation de ne disposer dans l'appariement que des données utiles à cet appariement ou au traitement pour lequel l'appariement doit donner lieu.

Dans la droite ligne de ces avancées et de cette prise de conscience du potentiel des données fiscales, le Service statistique public (SSP) s'est armé pour gagner en expertise et construire une information statistique de manière pérenne à partir de ces sources. Preuve de concept de l'intégration des sources fiscales dans le système d'information des statistiques démographiques et sociales, le **Fichier démographique sur les logements et les individus (Fidéli**) est ainsi apparu depuis 2016 dans ce système d'information : son objectif est de tirer parti des informations issues de l'administration fiscale sur l'impôt et les propriétés bâties pour compléter l'information déjà disponible sur le parc de logements et la démographie résidente. Au final, Fidéli constitue un bon exemple du travail de mise en cohérence, d'appariement et d'enrichissement de sources administratives permettant d'aboutir à une information statistique avec la constitution d'un objet dont la cohérence, l'exhaustivité et la variété d'informations disponibles sont essentielles pour son insertion dans le système d'information du SSP.

Le but de la construction du fichier est de disposer, à partir de plusieurs sources « brutes », d'une liste unique de logements d'habitation et d'une liste unique de personnes, puis de localiser ces personnes prioritairement dans leur logement principal, tout en regroupant les informations socio-démographiques les concernant.

DISPOSER D'UNE LISTE EXHAUSTIVE DES LOGEMENTS

Fidéli peut se définir comme **une base annuelle exhaustive de données statistiques sur les logements et de leurs occupants**, permettant d'éclairer le parc de logements ainsi que les mobilités résidentielles. Cette base de données repose tout d'abord sur les fichiers fiscaux sur le bâti, qu'il soit résidentiel ou non ***(encadré 2)***. Les données du bâti fournissent différents éléments :

- des éléments de repérage, comme l'adresse au cadastre (code Rivoli[7], numéro de rue) ou encore l'adresse postale du propriétaire utilisée pour les correspondances par l'administration fiscale ;
- des informations sur la nature du propriétaire (personne physique ou morale) ;
- des informations sur la nature du bâti (maison, appartement, etc.), la superficie et le nombre de pièces, le nombre d'étages quand c'est pertinent, et d'autres éléments caractéristiques (présence d'un ascenseur, date de construction, etc.) ;
- des éléments de géolocalisation tels que la référence cadastrale[8] ;
- et d'autres informations, telles que la présence de dépendances par exemple.

La fusion entre le fichier du bâti et les données de la taxe d'habitation permet d'enrichir la description des locaux, notamment sur le fait qu'ils sont occupés par le propriétaire ou un locataire, à titre de résidence principale, secondaire ou qu'il s'agit d'un logement vacant. La taxe d'habitation permet en outre de connaître les foyers fiscaux assujettis à la taxe pour un local donné, ce qui permet ultérieurement dans Fidéli d'en déduire les occupants. Il faut enfin être en mesure de caractériser, parmi les locaux, quels sont ceux qui peuvent être qualifiés de logements.

7. Avec Fantoir (fichier annuaire topographique initialisé réduit), anciennement fichier Rivoli (Répertoire informatisé des voies et lieux-dits), la Direction générale des finances publiques (DGFiP) recense, pour chaque commune, différents types de « voies » et leur attribue un identifiant appelé code Rivoli.
8. La référence cadastrale désigne une parcelle cadastrale de manière unique.

Fidéli introduit alors deux notions distinctes de logement au sens de la source fiscale :

- une première définition se base sur l'unique variable de nature du local, en considérant que le local d'habitation est nécessairement soit une maison, soit un appartement. Il s'agit de la définition centrale du logement dans Fidéli, et la plupart des traitements se rapportant au champ des logements réalisés dans Fidéli se font en cohérence avec cette définition ;
- une seconde définition plus proche du recensement, considère un univers plus large de natures de locaux (telles que des chambres de domestique, des pièces indépendantes, etc.), mais ne prend en compte que les informations issues de la taxe d'habitation pour définir le type d'occupation du logement.

Ces deux caractérisations des logements dans Fidéli coexistent, pour des usages distincts, l'un visant la cohérence interne des informations utilisées, l'autre cherchant à se rapprocher d'un concept similaire au recensement : lorsqu'on compare le volume de logements dans Fidéli selon la seconde définition avec celui mesuré dans le recensement, on obtient

Encadré 2. Les principales sources fiscales utilisées pour la constitution de Fidéli

Les sources fiscales mises en cohérence et intégrées dans le cadre de la production de Fidéli sont de différentes natures. Tout d'abord, il y a les sources sur le bâti :

- les données **Majic** (Mise à jour des informations cadastrales) fournissent toute l'information connue au cadastre, sur la nature du bâti, et sur la présence de dépendances. L'ensemble des locaux, y compris les logements, y sont recensés. Les données contiennent de l'information sur les caractéristiques des logements (nombre de pièces, superficie), ainsi que des immeubles dans le cadre de l'habitat collectif (nombre d'étages, nom d'usage éventuel). La source apporte également des informations sur la nature du propriétaire (particulier ou société), sur la localisation, telles que l'adresse au cadastre et l'adresse postale utilisée par l'administration fiscale pour les correspondances avec le propriétaire ;
- les données de la **taxe d'habitation** (fichier PLFC) sont partiellement redondantes avec celles de Majic, puisqu'elles décrivent les logements assujettis à la taxe d'habitation. Par ailleurs, elles contiennent une information sur le type d'occupation du logement (résidence principale, résidence secondaire ou occasionnelle, logement vacant), ainsi que la liste de l'ensemble des foyers fiscaux des occupants du logement (fournissant ainsi le lien logement occupant) ;
- le **FIP (fichier d'imposition des personnes)** quant à lui est une liste d'occurrences fiscales correspondant pour chacune d'entre elles à un ensemble d'individus connus par les services fiscaux, non nécessairement unique. Ce fichier permet d'associer un individu à un foyer fiscal ; il contient également de nombreuses informations identifiantes au sujet des individus (en particulier les traits d'identité, c'est-à-dire les noms, prénoms, date de naissance ou encore lieu de naissance). Ce fichier recense essentiellement les déclarants et leur éventuel conjoint, ainsi que les personnes adultes à charge ; une partie des individus mineurs manque, et est fournie par un autre fichier ;
- le **POTE (fichier permanent des occurrences de traitement des émissions)** est issu de la déclaration d'impôt sur le revenu (en particulier le formulaire 2042). Ce fichier complète les informations du FIP avec les statuts matrimoniaux des individus, et recense en théorie le nombre des personnes à charge des ménages. Ces données permettent également de fournir une information sur les revenus individualisables (salaires, pensions, indemnités chômage, etc.) pour chaque déclarant.

des différences de l'ordre de 1 % ; celles-ci s'expliquent en partie par l'absence dans les fichiers fiscaux des logements ordinaires qui ne sont pas soumis à la taxe d'habitation. Les écarts s'amplifient lorsqu'on s'intéresse aux résidences principales uniquement : là encore, cela est imputable au fait qu'un certain nombre de logements ordinaires ne sont pas assujettis à la taxe d'habitation, mais aussi à un certain retard dans l'enregistrement des logements récents dans les bases fiscales. Enfin, Fidéli comporte un nombre sensiblement plus élevé de logements vacants, car les sources fiscales tardent à prendre en compte les destructions de logements.

Une fois ces définitions établies, on poursuit la caractérisation des logements par **l'identification du parc des bailleurs sociaux, au moyen d'un appariement avec les données du Répertoire du parc locatif social (RPLS)**[9]. Cet appariement se déroule en plusieurs étapes, et implique un appariement sur le numéro Siren et sur la dénomination du bailleur. Une partie du parc est également repérée grâce aux informations d'exonérations provenant des fichiers du bâti.

Le parc de logements en France étant ainsi caractérisé, il faut ensuite lister l'ensemble des individus susceptibles de l'occuper et parvenir à relier chaque individu à un ou plusieurs logements.

DISPOSER D'UNE LISTE DES PERSONNES SANS DOUBLONS…

Lorsque l'administration fiscale compile l'information dont elle dispose sur les individus, son objectif est le recouvrement de l'impôt, principalement de l'impôt sur le revenu et de la taxe d'habitation. En conséquence, si elle poursuit bien un objectif de complétude, elle ne se préoccupe pas de la question de la redondance ; de ce fait, un même individu peut être identifié par l'administration fiscale de manière non unique : il apparaît concrètement sur plusieurs lignes dans les fichiers fiscaux, sous des identifiants différents.

Le statisticien public poursuit, pour sa part, un double objectif d'exhaustivité et d'unicité dans le dénombrement : en clair, il s'agit de recenser précisément et en totalité les individus résidant sur le territoire français. Pour atteindre cet objectif, il faut repérer les doublons et les éliminer.

Les individus de 15 ans et plus vivant dans des foyers ayant déclaré des revenus ou payé la taxe d'habitation sont connus à partir du **fichier d'imposition des personnes (FIP)**. Celui-ci contient un grand nombre de variables décrivant les individus très précisément : date et lieu de naissance, sexe, nom (et éventuellement nom marital), prénoms, situation matrimoniale, et pour les individus décédés, année de décès. Comme il s'agit d'un fichier de gestion, plusieurs localisations peuvent figurer pour un même foyer fiscal (par exemple, ancienne et nouvelle adresse en cas de déménagement).

Les individus de moins de 15 ans ne sont connus que dans le **fichier relatif à l'impôt sur le revenu (POTE**[10]**)**, mais caractérisés seulement par leur année de naissance. L'information sur la situation matrimoniale des individus peut s'avérer incohérente entre les deux fichiers FIP et POTE. Fidéli applique alors des règles de mise en cohérence afin d'obtenir une donnée unique pour chaque individu connu de l'administration fiscale.

9. Pour plus d'information sur les données concernant le système d'information sur le logement, voir (Harmois et Lamarche, 2020).
10. « Fichier permanent des occurrences de traitement des émissions », élaboré par les services de la DGFiP à partir des émissions des avis d'imposition sur les revenus.

… BÂTIR UN IDENTIFIANT NON SIGNIFIANT…

Pour les personnes de 15 ans et plus, les données fiscales contiennent un identifiant individuel. Mais celui-ci ne convient pas pour les finalités statistiques, et notamment du fait qu'environ un million de personnes sont présentes dans les données fiscales avec plusieurs identifiants fiscaux distincts. C'est la raison pour laquelle **Fidéli reconstruit un identifiant individuel non signifiant et spécifique à chaque année fiscale**, à partir de l'identifiant fiscal, mais aussi des éléments d'état-civil contenus dans FIP.

Cette opération nécessite de retravailler certaines informations : c'est notamment le cas du lieu de naissance, dont la recodification est l'un des éléments de standardisation de l'information fiscale les plus délicats. La variable est critique pour la constitution de l'information statistique, tant pour le repérage des doublons sur la base des traits d'identité que pour l'analyse *in fine*. Mais elle ne présente que peu intérêt pour l'administration fiscale, qui par conséquent n'apporte pas beaucoup d'attention à la qualité de sa collecte. Par ailleurs, l'information reste assez parcellaire selon le statut des individus vis-à-vis de l'administration fiscale. Ainsi, on dispose de peu de renseignements sur les personnes de moins de 15 ans à charge des foyers fiscaux. Cela pose, entre autres, un problème pour repérer les enfants en garde alternée. Autre difficulté : identifier puis localiser les étudiants, qui peuvent figurer dans le foyer fiscal de leurs parents en tant que personne à charge, et par ailleurs être connus de l'administration fiscale en tant que contribuables au titre de la taxe d'habitation.

… CHAÎNER LES MILLÉSIMES…

Les mêmes informations relatives au repérage unique des individus dans les sources fiscales permettent également de retrouver la personne dans les fichiers fiscaux de l'année précédente, afin notamment d'appréhender les mobilités résidentielles et les changements d'état matrimoniaux. Le chaînage de l'information dans Fidéli entre deux années consécutives est fondamental, car il ouvre la porte à un ensemble d'analyses en matière de transition, que les données collectées dans le cadre du recensement de population ne permettaient jusqu'alors de traiter que de manière assez partielle. Ainsi, Fidéli permet d'observer les évolutions concomitantes en matière de revenu, de composition du ménage, d'éventuelles mobilités résidentielles (allant jusqu'à observer ces mobilités à l'aide de coordonnées géographiques).

… ET RATTACHER LES INDIVIDUS À UN LOGEMENT (OU PLUSIEURS)

« La résidence principale définit la localisation des individus. »

Une fois qu'on dispose d'une liste cohérente d'individus et de logements, il convient de relier les individus à un ou plusieurs logements, de manière à les localiser, puis, grâce au suivi rétrospectif des individus, d'apprécier les mobilités résidentielles.

Pour faire un usage statistique de Fidéli, il est primordial de savoir où un individu habite, quelle est sa résidence principale : la résidence principale définit la localisation des individus. C'est à cette étape qu'intervient la connaissance des foyers fiscaux qui résident dans un même logement selon la taxe d'habitation.

Mais parce qu'ils peuvent appartenir à plusieurs foyers fiscaux liés à des adresses différentes, et parce que chaque foyer fiscal peut également être associé à plusieurs adresses, les individus ne sont pas naturellement liés à une seule adresse dans les sources fiscales. La procédure de localisation exige, ici encore, la mise en place de règles de décision permettant de lier un individu à une seule résidence principale. Fidéli utilise les informations fiscales pour déterminer la résidence principale, les règles liées aux contraintes fiscales étant parfois un peu différentes de celles adoptées par le recensement.

> *« La procédure de localisation exige la mise en place de règles de décision permettant de lier un individu à une seule résidence principale. »*

Dans un premier temps, une étape de « nettoyage » est appliquée, afin de restreindre l'univers des possibles aux individus encore en vie à la date de référence, aux foyers connus à l'impôt sur le revenu et déclarant des revenus à l'adresse du logement, et aux foyers connus à la taxe d'habitation. Ensuite, si un individu appartient à plusieurs foyers fiscaux, on applique des règles de priorisation entre les différents foyers connus par l'administration fiscale, afin de ne conserver qu'une seule localisation dans leur résidence principale , dès lors que celle-ci est connue[11]. Lors de cette deuxième étape, il peut apparaître des individus figurant dans plusieurs foyers, dans l'un comme déclarant principal ou son conjoint, et dans les autres comme personne à charge (le cas typique est celui des jeunes adultes). La priorité est alors donnée en s'appuyant sur le statut de l'individu dans le foyer : il sera localisé dans le foyer où il est déclarant principal plutôt que dans celui où il est à charge (ses parents pour un jeune adulte).

Un foyer peut être localisé au sens de l'impôt sur le revenu à une adresse, et au sens de la taxe d'habitation à une autre adresse ***(figure 2)***. Si l'adresse de la taxe d'habitation est celle d'une résidence principale, c'est à cette adresse que l'individu est localisé. Si on ne lui connaît pas de résidence principale dans les sources fiscales, la personne est localisée à l'adresse de son imposition sur le revenu.

Par ailleurs, l'information sur l'adresse relative aux foyers fiscaux est multiple : dernière adresse connue, adresse de la résidence principale au sens de la taxe d'habitation, ou encore adresse utilisée par l'administration fiscale pour échanger avec les contribuables. Ces adresses seront plus ou moins pertinentes selon le but recherché au travers de la localisation des individus ; et elles présentent naturellement des imperfections en étant le résultat d'une gestion administrative, laquelle est de moins en moins fondée sur l'échange de courriers postaux.

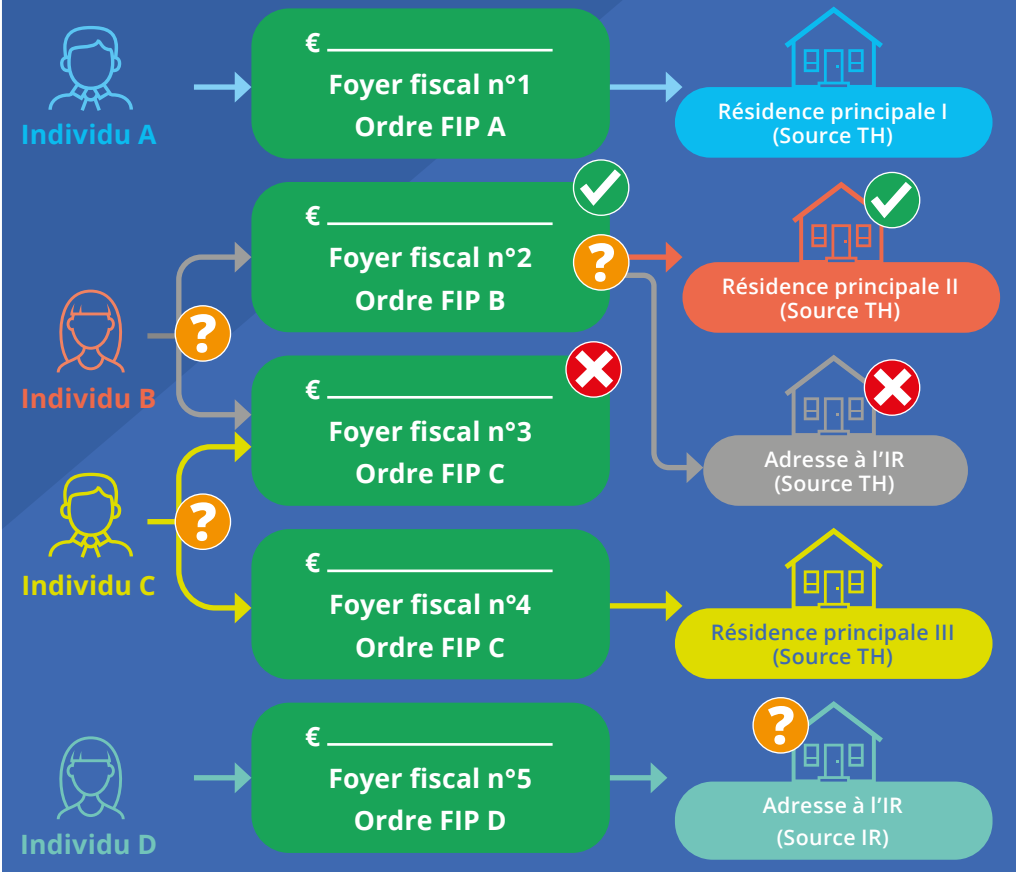
11. Si le foyer n'a pas de résidence principale au sens fiscal, il est localisé là où il déclare l'impôt sur le revenu.

REPÉRER LES MOBILITÉS RÉSIDENTIELLES…

Une fois la localisation des individus réalisée, la comparaison avec le millésime précédent permet de repérer les mobilités résidentielles au travers des changements d'adresse.

Les logements sont ensuite géolocalisés, sur la base d'abord des informations du cadastre, puis dans les cas problématiques, sur la base des éléments d'adressage (numéro, voie, etc.), avec une éventuelle procédure d'arbitrage dans les cas où les informations obtenues ne sont pas cohérentes. Fidéli construit ainsi une information nouvelle, sous forme de coordonnées géographiques, pour l'ensemble des logements du fichier. La géolocalisation permet également d'identifier les adresses appartenant à des IRIS[12] ou à des quartiers prioritaires de la politique de la ville. Cette information est extrêmement précieuse, car elle permet des analyses à un niveau local potentiellement très fin.

Figure 2. Le processus de localisation des individus dans un logement



Note : Ordre FIP est l'identifiant de l'adresse à laquelle le contribuable est connu pour une taxation donnée (impôt sur le revenu, taxe d'habitation principale, secondaire, etc.).
TH : fichiers de la taxe d'habitation ; IR : fichiers de l'impôt sur le revenu.

12. Îlots regroupés pour l'information statistique, découpage infra-communal des communes de 5 000 habitants et plus.

Au final, le processus aboutit à la production d'une information globale et la plus cohérente possible permettant de lister l'ensemble des individus résidant sur le territoire national, en regard de leurs caractéristiques socio-démographiques et de leur localisation.

UNE SOURCE COMPLÉMENTAIRE DANS LE SYSTÈME D'INFORMATION STATISTIQUE ACTUEL

Pour mieux cerner les logements ordinaires, les adresses dans Fidéli sont également enrichies avec les données du recensement sur les communautés ; de la même manière, les résidences hôtelières font également l'objet d'un traitement spécifique. Les données de Fidéli sont aussi rapprochées de Filosofi[13], ce qui permet de disposer d'un revenu disponible et d'un niveau de vie pour les ménages vivant dans les logements pour lesquels l'information est disponible dans Filosofi.

« Fidéli se positionne ainsi comme un complément extrêmement complet et riche du système d'information des statistiques démographiques et sociales. »

L'ensemble de ces traitements est pensé de manière à se rapprocher le plus possible conceptuellement et quantitativement des informations traditionnellement collectées dans le cadre du recensement de population ou des enquêtes auprès des ménages. Fidéli se positionne ainsi comme un complément extrêmement complet et riche du système d'information des statistiques démographiques et sociales, répondant au besoin d'exhaustivité qu'induit la demande grandissante d'analyses fines au niveau local.

Fidéli est un produit intermédiaire, dans le sens où malgré sa grande richesse, il ne donne pas lieu à une diffusion propre en dehors des fichiers détails qui alimentent le Service statistique public ainsi que le monde académique. Par exemple, sur le dénombrement de la population à l'échelle d'une commune, Fidéli ne peut prétendre remplacer le recensement de population, qui fait foi ; comme dit précédemment, les sources fiscales ne sont pas construites sur un besoin originel de dénombrement de la population, mais de recouvrement de l'impôt. De ce point de vue, elles tendent à surestimer le nombre d'individus résidant sur le territoire national, tendance que le traitement et la mise en cohérence statistique réalisés dans le cadre de la production de Fidéli ne pallient malheureusement pas totalement. En revanche, elles permettent d'apporter des éclairages sur la population à l'échelle infra-communale, que le recensement sous sa forme moderne ne permet pas toujours, comme les mobilités résidentielles.

Fidéli, bien que contenant de nombreuses variables sur les revenus, ne remplace pas Filosofi : ce dernier reste la source de référence pour l'analyse de la distribution des revenus au niveau local ; en revanche, les variables de revenu disponibles dans Fidéli permettent d'envisager cette dimension comme un descripteur bien souvent pertinent pour les analyses rendues possible par le fichier.

13. Dispositif sur les revenus localisés sociaux et fiscaux. Les choix de localisation n'étant pas les mêmes pour Fidéli et Filosofi, l'appariement n'est que partiel.

UNE RICHESSE EN MATIÈRE DE GÉORÉFÉRENCEMENT ET DE VARIABLES DESCRIPTIVES

Depuis sa création, Fidéli a permis d'éclairer certains angles morts laissés par le recensement de population et les enquêtes ménages, notamment en ce qui concerne la géolocalisation dans les communes de moins de 10 000 habitants.

L'apport de Fidéli est illustré par un exemple récent : il s'agit de la possibilité de dénombrer de manière très précise, à l'aide des données de géolocalisation, le nombre d'individus vivant en zone à risque autour de l'implantation des centrales nucléaires.

Il est également possible d'évaluer de manière assez globale l'impact d'une submersion des zones littorales sur l'économie locale et la population, en utilisant les autres informations disponibles désormais au niveau local sur l'activité économique et l'emploi, et en les combinant avec les informations sur les caractéristiques socio-démographiques et les revenus des ménages vivant sur place (Brendler *et alii*, 2020).

Fidéli a aussi permis de savoir que les personnes habitant des quartiers prioritaires ont des mobilités résidentielles assez semblables aux autres résidents, contrairement à certaines idées reçues (Vicaire *et alii,* 2018).

Plus récemment, les mobilités observées au tout début de la crise sanitaire de 2020 ont pu être éclairées à l'aide des données de Fidéli, qui permettent de lier les individus non seulement à une résidence principale, mais également à d'éventuelles résidences secondaires et à leur localisation. Il a été ainsi possible d'expliquer les mouvements observés au travers des données de téléphonie mobile, à l'aune des informations que fournit Fidéli sur les résidences secondaires possédées par les ménages vivant en région parisienne.

In fine, l'information contenue dans Fidéli, sur les caractéristiques des logements et sur celles des ménages, que ce soit en termes de composition, de revenu ou de niveau de vie, combinée aux données sur les mobilités résidentielles d'une année sur l'autre, permet d'éclairer de manière assez complète les mobilités et de mieux en comprendre les ressorts.

MAIS UN RISQUE RÉEL LIÉ À LA DIFFUSION DES DONNÉES LOCALISÉES

Toutefois, l'immense potentiel d'analyses que permet en théorie Fidéli masque de véritables contraintes dans l'usage qu'il est possible d'en faire. La richesse de ces données et le grand détail d'informations qu'elles contiennent révèlent en creux un risque accru de différenciation géographique du fait du grand nombre de zonages rendus possibles, et de la grande variété d'informations sur lesquels ces zonages peuvent porter.

Fidéli fait ainsi l'objet de conditions de diffusion très restrictives, dans la mesure où le spectre d'informations sur lequel il contient des variables est large. De ce fait, un usage mal contrôlé de ce fichier, par de multiples acteurs, souvent situés en dehors du Service statistique public et pour lesquels la coordination peut être plus difficile, pourrait amener au dévoilement involontaire d'informations individuelles sensibles. Son existence pose alors une exigence accrue pour le service producteur : exigence de vérification et exigence de centralisation des statistiques produites à partir des différents millésimes mis à disposition des utilisateurs. L'idée est donc de contraindre les utilisateurs *a priori* pour limiter le risque sur le secret statistique *a posteriori*.

Dans ce contexte, un autre facteur rend l'existence de Fidéli aisément valorisable par le monde de la recherche : la mise à disposition aux chercheurs des données détaillées dans le cadre d'une procédure et d'un accès sécurisés, grâce au Centre d'accès sécurisé aux données (Gadouche, 2019). En donnant l'assurance que les conditions de confidentialité et de diffusion sont strictement respectées, l'infrastructure du CASD permet un accès simplifié à des données telles que Fidéli, et démultiplie de ce fait les possibilités d'application dans le monde de la recherche.

AU-DELÀ DES ÉTUDES, DE NOUVELLES MISSIONS POUR LA PRODUCTION STATISTIQUE

« Fidéli constitue désormais la base de sondage de la plupart des enquêtes ménages menées dans le Service statistique public. »

La finalité de Fidéli n'est pas uniquement de mettre à disposition des données pour les études. À l'issue du processus de traitement mutualisé des données fiscales, des livrables sont utilisés pour des finalités de production, par d'autres applications. C'est ainsi le cas pour l'échantillon démographique permanent[14] ou quelques processus liés au recensement. Mais surtout, grâce à l'exhaustivité de la source, la qualité de son géoréférencement, et les très nombreuses variables descriptives des individus, ménages et logements qu'elle contient, Fidéli constitue désormais la base de sondage de la plupart des enquêtes ménages menées dans le Service statistique public.

Originellement, cette mission était confiée au recensement de population ; le passage en 2004 à un recensement annuel rotatif a ouvert la voie d'un changement de paradigme : dans cette mutation, il s'est recentré sur le dénombrement de population en continu, en abandonnant de ce fait son caractère traditionnellement exhaustif qui est une propriété désirable pour une base de sondage (Sillard *et alii*, 2020).

Les sources fiscales se sont alors imposées comme un candidat naturel pour constituer une base de sondage, et ce d'autant plus qu'elles contiennent, du fait du recouvrement de la taxe d'habitation, un élément très intéressant pour la collecte des enquêtes auprès des ménages : le lien logement-occupant. Ce lien est central, car il permet l'identification précise sur le terrain de l'unité de collecte qui a été échantillonnée pour être enquêtée. Il est en effet difficile pour un enquêteur de repérer soit un logement sans en connaître l'identité de ses occupants, soit un individu ou un ménage sans en connaître le lieu d'habitation principal. Fidéli fournit une information de grande valeur pour les enquêtes ménages, en constituant une base de sondage mise à jour chaque année, et détaillant pour l'ensemble des individus résidant sur le territoire national leur localisation ainsi que de nombreuses variables contextuelles utiles à l'élaboration d'un plan de sondage pertinent. Fidéli présente ainsi différents avantages attendus pour une base de sondage de qualité : quasi-exhaustivité de l'information (y compris contextuelle), faibles défauts de couverture, mises à jour de l'information régulière, et pluralité des unités possiblement échantillonnées. En revanche, de nouveaux obstacles se font jour, comme la difficulté de repérage des logements dans le collectif, puisque les informations de rang de logement usuelles dans le recensement n'existent pas dans les sources fiscales.

14. Voir l'article d'Isabelle Robert-Bobée et Natacha Gualbert sur l'échantillon démographique permanent, dans ce même numéro.

Par ailleurs, le développement des collectes multi-modes pour ces enquêtes auprès des ménages nécessite de disposer, en amont de la collecte, de données de contact telles que des adresses électroniques et des numéros de téléphone : les sources fiscales en disposent et vont constituer une grande opportunité d'amélioration substantielle des conditions de collecte dans ce contexte[15].

LES PERSPECTIVES : D'ABORD MAINTENIR LES PROGRÈS RÉALISÉS…

Il n'est pas possible de parler de Fidéli sans évoquer la disparition des fichiers de la taxe d'habitation suite à la réforme fiscale entreprise en 2017, même si ces données sont moins utiles pour Fidéli, contrairement à Filosofi. Avec la suppression de la taxe d'habitation pour les résidences principales (Bur et Richard, 2018), on perd le lien entre occupant et logement, qui permet de reconstituer des ménages, mais on conserve tout le reste, qui a permis de réaliser la plupart des études évoquées plus haut.

Fidéli n'utilise le lien entre occupant et logement que de façon assez marginale, principalement pour reconstituer des ménages. En revanche, le lien entre occupant et logement est indispensable pour l'échantillonnage des enquêtes réalisé à partir de Fidéli. Compte tenu de l'importance de cet objectif, l'Insee a investi dans différents projets visant à assurer la pérennité de cette information sur différentes échelles de temps.

Dans un premier temps, un projet de court terme vise à effectuer une forme de ré-ingénierie de la chaîne de production du fichier Fidéli, de manière à rendre plus modulaire l'incorporation du lien logement-occupant et ainsi autoriser l'usage de sources alternatives contenant de l'information sur ce lien. À horizon de 2023, l'administration fiscale doit par ailleurs collecter l'information sur les occupants des logements car si la taxe d'habitation sur les résidences principales aura disparu, les taxes d'habitation sur les résidences secondaires ou la taxe sur les logements vacants existeront toujours par la suite ; il faudra donc bien être en mesure de distinguer les résidences principales et leurs occupants des autres logements. Cette nécessité de complétude du système d'information fiscale doit permettre d'assurer la pérennité de Fidéli dans les années à venir.

… ET POURSUIVRE L'INVESTISSEMENT SUR LES SOURCES FISCALES POUR ENRICHIR LE FICHIER

La force des données administratives, et en particulier fiscales, est qu'elles sont par essence porteuses d'autres types d'information que l'information collectée primitivement à but statistique, pourvu que l'on mette en œuvre le traitement approprié. Ainsi, les fichiers du bâti contiennent naturellement des informations sur l'identité du ou des propriétaires des logements ; il est alors possible d'enrichir les données existantes au travers du lien logement-propriétaire, ouvrant alors un nouvel horizon en matière d'analyses, cette fois sur la dimension patrimoniale.

15. Voir l'application dans le cadre de l'enquête Emploi dans l'article de François Guillaumat-Tailliet et Chloé Tavan dans ce même numéro.

En appariant les données du bâti avec les fichiers de Demandes de valeurs foncières[16], il est même possible d'obtenir une valeur de marché pour l'ensemble des logements ayant fait l'objet d'une transaction dans les dernières années ; et à l'aide de modèles économétriques ou d'apprentissage correctement estimés, d'évaluer une valeur de marché pour l'ensemble des logements recensés dans Fidéli. Ces travaux ont déjà été entrepris, avec l'ambition d'incorporer dans les millésimes futurs les informations relatives au patrimoine immobilier des individus dans le champ de Fidéli. Les travaux relatifs au calcul du lien logement-propriétaire sont complexes, en particulier parce qu'ils nécessitent la mise en transparence des sociétés immobilières civiles, sans lesquelles l'information sur le patrimoine des ménages reste incomplète ; mais cette complexité est assez comparable à celle qui caractérise les traitements permettant la production de Fidéli sous sa forme actuelle, et peut donc être résolue avec des méthodes similaires. En revanche, ils renforcent l'utilité du fichier, et sa place de plus en plus centrale dans le système d'information des statistiques démographiques et sociales.

ET AU-DELÀ…

Au-delà, c'est une vision plus ambitieuse des sources administratives, et pas uniquement fiscales, qui prévaut afin d'assurer l'alimentation des besoins de la statistique démographique et sociale en matière de données localisées, variées et exhaustives. Il faut ainsi tendre vers une intégration beaucoup plus poussée des sources administratives pour assurer la résilience du système d'information vis-à-vis de la potentielle transformation, voire la disparition de certains fichiers. C'est à ce prix que les sources administratives occuperont toute la place qui peut être naturellement la leur dans le système statistique d'une économie administrée ; en contrepartie également, cette transformation suppose une vision beaucoup plus holistique, intégrée et cohérente du système d'information des statistiques démographiques et sociales.

16. DVF est un jeu de données sur les transactions immobilières en France produit par la Direction générale des finances publiques. Il est complémentaire des bases de données BIEN et PERVAL produites par les notaires. Voir (Harnois et Lamarche, 2020).

BIBLIOGRAPHIE

ABS, 2021. MADIP data and legislation. In : *site de l'Australian Bureau of Statistics*. [en ligne]. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Statistical+Data+Integration+-+MADIP+data+and+legislation*.

BAKKER, Bart F. M., VAN ROOIJEN, Johan, VAN TOOR, Leo, 2014. The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. In : *Statistical Journal of the IAOS*. 2 avril 2014. Vol. 30, n° 4, pp. 411-424.

BENS, Arno et SCHUKRAFT Stefan, 2019. Modernisation des registres administratifs en Allemagne – Développements actuels et enjeux pour la statistique publique. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 10-20. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168390/courstat-2-3.pdf*.

BOUCHER, Philippe, 1974. Safari ou la chasse aux Français. In : *Le Monde*. [en ligne]. 21 mars 1974. Page 9. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.lemonde.fr/archives/article/1974/03/21/une-division-de-l-informatique-est-creee-a-la-chancellerie-safari-ou-la-chasse-aux-francais_3086610_1819218.html*.

BRENDLER, J., COMTE, S., LOUZA, T., MOUNCHIT, N., DARDAILLON, B., ROSE et V., PAILLETTE, É. 2020. *Plus de 100 000 résidents, logements et emplois concernés par le risque de submersion marine en Normandie.* [en ligne]. 7 décembre 2020. Insee Analyses Normandie, N°87. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4989506*.

BUR, Dominique et RICHARD, Alain, 2018. *Mission Finances Locales. Rapport sur la refonte des finances locales*. [en ligne]. Mai 2018. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.vie-publique.fr/sites/default/files/rapport/pdf/184000278.pdf*.

EUROSTAT, 2018. Population grids. In : *Statistics explained.* [en ligne]. 16 juillet 2018. [Consulté le 25 mai 2021] . Disponible à l'adresse : *https://ec.europa.eu/eurostat/statistics-explained/index.php/Population_grids*.

GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la data science et la recherche scientifique. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254227/courstat-3-7.pdf*.

HARNOIS, Jérôme et LAMARCHE, Pierre, 2020. Le système statistique du logement – Étendue et perspectives. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee. N° N4, pp. 142-162. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497089/courstat-4-9.pdf*.

PADIEU, René, 2005. Grandes bases de données et protection des personnes. In : *Courrier des statistiques.* [en ligne]. Mars-juin 2005. Insee. N° 113-114, pp. 65-67. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/122548/1/cs113-114.pdf*.

SÉNAT, 2021. 1977 – 1978 : Le Sénat invente les autorités administratives indépendantes. In : site du Sénat. [en ligne]. Dossiers d'histoire. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.senat.fr/evenement/archives/D45/context.html*.

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 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497081/courstat-4-6.pdf*.

STATISTICS FINLAND, 2004. *Use of registers and administrative data sources for statistical purposes. Best practices of Statistics Finland*. [en ligne]. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.stat.fi/tup/julkaisut/kasikirjoja_45_en.pdf*.

STATISTICS NEW ZEALAND, 2014. *Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project.* [en ligne]. Juin 2014. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.stats.govt.nz/assets/Uploads/Retirement-of-archive-website-project-files/Methods/Linking-methodology-used-by-Statistics-New-Zealand-in-the-Integrated-Data-Infrastructure-project/linking-methodology-IDI-project.pdf*.

TRAINOR, Cathy et TRUDEAU, Richard, 2015. Environnement de couplage de données sociales (ECDS). In : *Conférence nationale du RCCDR des 5-6 novembre 2015*. [en ligne]. Statistique Canada et Réseau canadien des centres de données de recherche, atelier pré-conférence RCCDR. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://crdcn.org/sites/default/files/4._sdle_overview_-_crdcn_pre-conference_workshop_nov_4_2015_french_final.pdf*.

UNECE, 2007. *Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics.* [en ligne]. Nations Unies, New York et Genève. Séries Statistical standards and studies (Conference of European Statisticians). [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://digitallibrary.un.org/record/609979/files/Register_based_statistics_in_Nordic_countries.pdf*.

VICAIRE, Vincent, SÉMÉCURBE, François, FAIVRE, Cynthia et DARRIAU, Valérie, 2018. Mobilité résidentielle entre 2015 et 2016 : un mouvement de même ampleur dans les quartiers prioritaires que dans le reste de la ville. In : *ONPV, Rapport annuel 2017.* [en ligne]. [Consulté le 25 mai 2021]. Disponible à l'adresse : *http://www.onpv.fr/uploads/media_items/rapport-onpv-2017-synth%C3%A8se-stephanie-mas-1.original.pdf*.

FONDEMENTS JURIDIQUES

Loi n° 51-711 du 7 juin 1951 sur l'Obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance*. [en ligne]. Mise à jour du 25 mars 2019. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573/2021-05-26/*.

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. Mise à jour du 9 décembre 2020. [Consulté le 25 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000033202746/2020-12-09*.

L'ÉCHANTILLON DÉMOGRAPHIQUE PERMANENT

EN 50 ANS, L'EDP A BIEN GRANDI !

Isabelle Robert-Bobée et Natacha Gualbert***

Depuis plus de cinquante ans, l'Insee rassemble des informations socio-démographiques sur un échantillon d'individus représentatifs de la population résidant en France : l'échantillon démographique permanent. Pour chacun de ces individus, l'EDP s'enrichit chaque année de données issues du recensement, de l'état civil, du fichier électoral, et plus récemment, de données d'emploi pour les salariés, et de données fiscales (déclaration de revenus et taxe d'habitation). Suivront prochainement l'ajout de données sur les non salariés.

Actuellement, l'EDP retrace 3,7 millions de trajectoires individuelles, dont plus de 200 000 sur 50 ans. C'est une source unique pour l'étude des mobilités géographiques et sociales sur longue période, mais aussi des évolutions du niveau de vie en lien avec les événements familiaux ou professionnels vécus, comme une séparation ou le passage à la retraite. L'introduction du niveau de vie dans l'EDP, en plus des caractéristiques socio-démographiques habituelles (âge, sexe, catégorie socio-professionnelle, diplôme, situation familiale), a encore étendu le champ des études avec ce panel.

Rare panel d'individus en population générale, l'EDP a su s'adapter aux changements des sources qui l'alimentent. Le passage du recensement exhaustif aux enquêtes de recensement au milieu des années 2000 a été l'occasion d'élargir la taille de l'échantillon et d'intégrer de nouvelles sources exhaustives (données socio-fiscales) ou en panel (déclarations de salaires).

For more than fifty years, the permanent demographic sample (known as "EDP") has been accumulating socio-demographic information for a sample of individuals representative of the population living in France. For each of these individuals, the EDP is enriched each year with data from the census, civil status, the electoral register and, more recently, employment data for salaried employees and tax data (income tax return and housing tax). This will soon be followed by the addition of data on the self-employed, and progress on the health side to evaluate the national health strategy.

The EDP currently tracks 3.7 million individual trajectories, including more than 200,000 over 50 years. It is a unique source for the study of geographical mobility over a long period, but also of changes in living standards in connection with family or professional events experienced, such as couples breakdowns or the transition to retirement. The inclusion of the standard of living in the EDP, in addition to the usual socio-demographic characteristics (age, gender, socio-professional category, diploma, family situation) has further extended the field of study with this panel.

As a rare panel of individuals in the general population, the EDP has been able to adapt to changes in the sources it is based on. The transition from comprehensive census to census surveys in the mid-2000 s was an opportunity to broaden the sample size and integrate new comprehensive (tax data) or panel (employees) sources.

* Cheffe de la division Enquêtes et études démographiques, Insee, *isabelle.robert-bobee@insee.fr*

** Responsable statistique de l'échantillon démographique permanent, Insee, *natacha.gualbert@insee.fr*

L'échantillon démographique permanent (EDP) a été créé en 1968 à l'Insee à partir de la compilation de données des recensements de la population et d'état civil. Il s'agissait de mettre en place un nouvel outil pour l'analyse des mobilités géographiques et des trajectoires sociales (Sautory, 1988), comme les différentiels sociaux de mortalité, les parcours professionnels sur longue période ou la trajectoire des immigrés (mobilité professionnelle, acquisition de la nationalité française par exemple).

Plusieurs approches sont possibles pour constituer des données de trajectoires : enquêter des personnes à un moment donné en leur posant des questions sur leur parcours passé (enquête rétrospective, faisant appel à la mémoire des enquêtés) ; interroger plusieurs fois les mêmes personnes pour recueillir des informations au fil des années (mais avec la difficulté de retrouver les personnes pour les réinterroger, et une attrition qui augmente donc au fil des années) ; ou, comme le panel EDP, rassembler au fil des années des données recueillies par ailleurs.

Dès sa création, l'EDP mobilise des données administratives et des données du recensement de la population, ce qui en fait un dispositif particulièrement économe : pas de coût de collecte, pas de charge de réponse auprès d'enquêtés. L'échantillon peut donc être de grande taille, le tout sans attrition ni biais de mémoire. Un critère d'échantillonnage très simple a été retenu : le jour de naissance. Cela simplifie la mise en œuvre du panel et donc sa pérennité.

Sans effet de mémoire ni attrition, l'EDP est unique également par sa taille (3,7 millions de personnes actuellement), sa profondeur historique (50 ans de données pour plus de 200 000 personnes), et par la diversité de ses sources (état civil, recensement, puis fichier électoral et données socio-fiscales).

En 50 ans, l'EDP a dû s'adapter aux évolutions parfois importantes de ses sources historiques ; le panel a su aussi intégrer de nouvelles données, qui ont enrichi les études, mais ont aussi rendu plus complexe son exploitation.

À L'ORIGINE, ÉTAT CIVIL ET RECENSEMENT POUR LES NATIFS DES 4 PREMIERS JOURS D'OCTOBRE

Les personnes faisant partie de l'EDP sont celles nées certains jours de l'année, dits « jours EDP » et fixés par arrêté depuis 2014[1]. Au départ, il s'agissait des personnes nées les 4 premiers jours d'octobre. La première source retenue pour démarrer le panel a été le recensement de 1968, complétée par des données d'état civil (naissances, décès, mariages).

En 1968, année de la création de l'EDP, 500 000 personnes sont ainsi entrées dans ce panel ***(figure 1)*** :

- 484 200 personnes recensées en 1968 et nées un 1, 2, 3 ou 4 octobre (quelle que soit leur année de naissance ;
- 9 200 personnes nées les 1er, 2, 3 ou 4 octobre 1968 pour lesquelles on disposait du bulletin de naissance dans les fichiers de l'état civil. Elles n'ont pas été recensées en 1968, le recensement ayant eu lieu en mars 1968[2], donc avant leur naissance ;
- 900 personnes ajoutées grâce à l'état civil sur les mariages (6 900 personnes nées un 1er, 2, 3 ou 4 octobre se sont mariées en 1968, et parmi elles, 6 000 ont été recensées en 1968) ;

1. Arrêté du 6 août 2014 portant création d'un traitement automatisé de données à caractère personnel relatif à l'échantillon démographique permanent de l'Insee (voir les références réglementaires en fin d'article).
2. Les recensements généraux de population (jusqu'en 2004) se déroulaient généralement au mois de mars.

- et 500 personnes nées en 1967 mais non recensées en 1968 (sur les 8 300 bulletins de naissance d'enfants nés un 1er, 2, 3 ou 4 octobre 1967, 7 800 concernaient des personnes déjà incluses dans l'EDP grâce au recensement de 1968).

Si le panel avait été limité aux seules personnes nées en 1968, il aurait fallu attendre de nombreuses années avant d'étudier des trajectoires. Pour éviter cet inconvénient, l'échantillon a dès le départ été constitué de personnes de tous âges. Le panel s'est enrichi ensuite chaque année des personnes nées l'un des quatre jours de référence au cours de l'année en France à partir d'informations recueillies dans des actes d'état civil et d'informations consignées dans les bulletins de recensement successifs. Ainsi, la base était proche d'un échantillon représentatif au 1/100e (4 jours / 365 jours) de la population résidant en France (Couet, 2006).

« L'échantillon a dès le départ été constitué de personnes de tous âges. »

Depuis sa création, le renouvellement de l'échantillon est assuré par les naissances ou par la venue en France de nouvelles personnes. Si elles sont nées un jour EDP, ces dernières rejoignent l'EDP à l'occasion d'un recensement ou d'un événement enregistré dans un acte de l'état civil (Couet, 2006). À l'inverse, le suivi d'un individu cesse, de fait, en cas de décès ou de départ à l'étranger. Cependant, la trace de cet individu persiste dans l'échantillon avec le détail des événements démographiques qui ont jalonné son parcours sur le territoire, et l'on peut ainsi comparer à tout moment les trajectoires de différentes cohortes.

PASSER À 16 JOURS POUR S'ADAPTER AU RECENSEMENT CONTINU

Le recensement de la population sert avant tout à établir le nombre d'habitants de chacune des circonscriptions administratives (établissement des populations dites légales), mais c'est aussi une source privilégiée pour décrire la population (sexe et âge, mais aussi diplôme, catégorie sociale, configurations familiales, etc.) : à ce titre, cette source est fondamentale pour l'EDP.

Le recensement a changé de méthode : exhaustif tous les sept à dix ans jusqu'en 1999, il est devenu annuel au début des années deux-mille, et sa collecte repose désormais sur un sondage. Depuis 2004, des enquêtes annuelles de recensement (EAR) sont ainsi menées sur un échantillon d'adresses[3]. Conséquence pour les individus inclus dans l'EDP : ils ne sont donc plus recensés en même temps ; pour une année donnée, seul un individu sur sept[4] est présent dans l'échantillon de l'enquête du recensement.

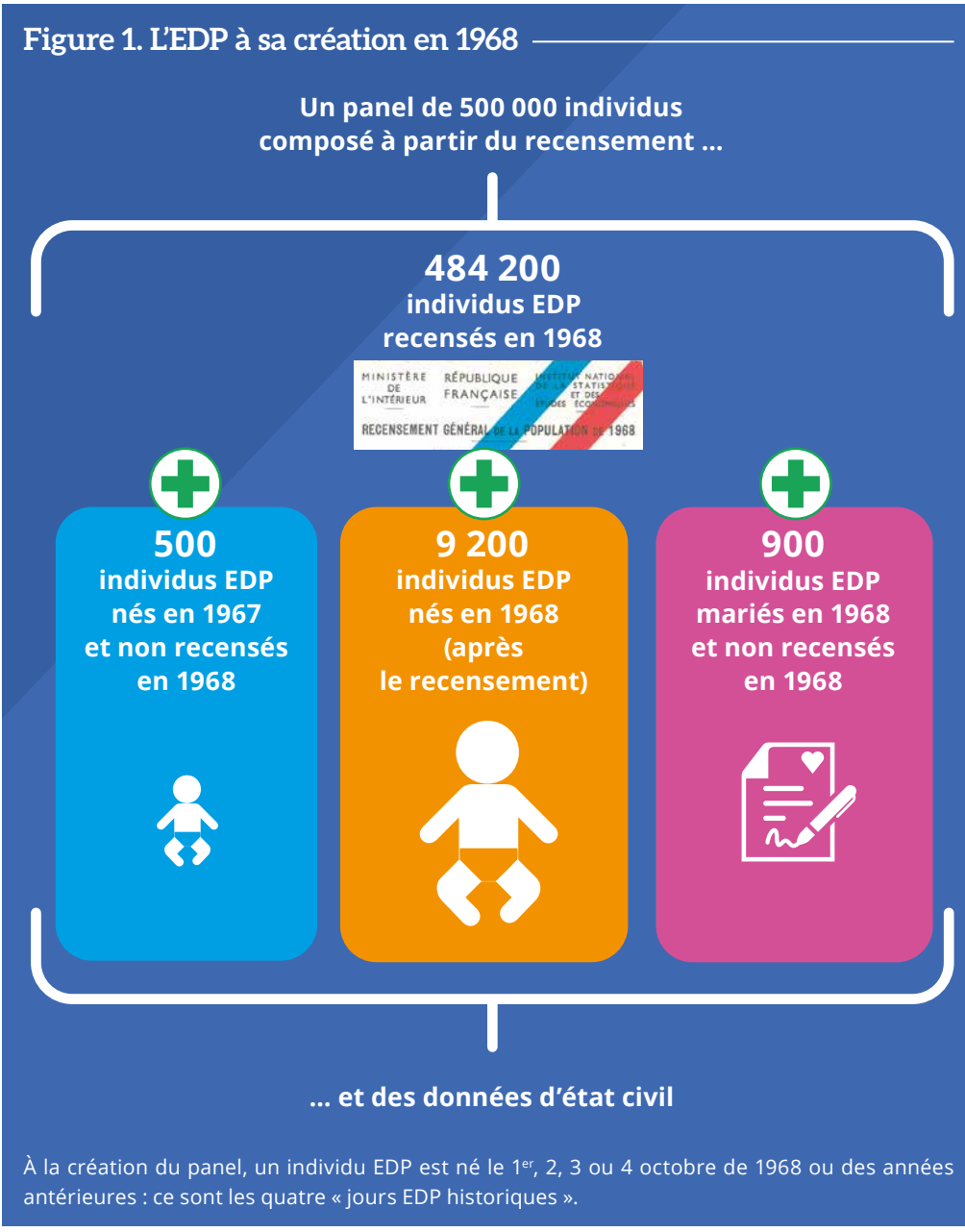
Pour compenser la dégradation de la qualité des estimations du fait de la réduction des données disponibles une même année, la taille de l'échantillon a été considérablement

3. Sur une période de cinq années successives, l'ensemble du territoire est couvert par une collecte du recensement. Les premiers résultats avec la nouvelle méthode sont ceux millésimés 2006, qui combinent les enquêtes de 2004 à 2008 (Godinot, 2005).
4. Une petite commune (moins de 10 000 habitants) sur cinq est recensée exhaustivement chaque année, ainsi que 8 % des adresses en grandes communes. Au total, environ un habitant sur sept est recensé une année donnée.

augmentée : le nombre de jours de référence EDP est passé de 4 à 16 ***(figure 2)***. Cet élargissement n'a pas été rétroactif, les données nominatives n'étant pas conservées dans les fichiers du recensement.

L'extension des « jours EDP » a volontairement été répartie sur l'année, pour améliorer les analyses de trajectoires pouvant être affectées par la saisonnalité des naissances (Cnis, 2006).

Figure 1. L'EDP à sa création en 1968



À la création du panel, un individu EDP est né le 1er, 2, 3 ou 4 octobre de 1968 ou des années antérieures : ce sont les quatre « jours EDP historiques ».

Depuis 2008[5], le suivi du panel EDP porte donc sur les personnes nées les 4 premiers jours de chaque trimestre, avec une subtilité pour éviter le 1er janvier[6]. Les « individus EDP » sont ceux nés un des 16 jours suivants : du 2 et au 5 janvier, du 1er au 4 avril, du 1er au 4 juillet, ou du 1er au 4 octobre (ces 4 derniers jours étant les « jours historiques de l'EDP »).

L'élargissement du panel à 16 jours n'est pas la seule innovation mise en place pour pallier la fin de l'exhaustivité du recensement. L'EDP s'est tournée vers une autre source exhaustive : les données socio-fiscales[7].

AVEC LES DONNÉES FISCALES, L'EDP RETROUVE SON EXHAUSTIVITÉ…

À l'origine, l'EDP compile des données d'état civil et de recensements de la population. Il s'enrichit en 2008 de données du fichier électoral (dates d'inscription sur les listes électorales, dates de radiation, communes d'inscription), puis du panel « tous salariés » (qui décrit le parcours d'emploi salarié et les rémunérations, depuis 1968) et de données socio-fiscales avec Fidéli (Fichiers démographiques sur les logements et les individus) et Filosofi (Fichier localisé social et fiscal) ***(encadré 1)***.

L'intégration des données fiscales a redonné à l'EDP l'exhaustivité qu'il avait perdue lors du changement de méthode du recensement de la population. Les données fiscales apportent annuellement pour tous les individus nés un des jours EDP notamment des informations sur :

- le logement (localisation, caractéristiques du logement) ;
- et sur la situation familiale (car celle-ci influe sur le taux marginal de l'imposition).

> *« L'intégration de données socio-fiscales est sans doute la plus grande avancée de l'EDP ces dernières années. »*

Cela compense donc le fait que ce type d'informations n'est désormais plus récupérable *via* le recensement une année donnée sur la totalité des individus.

L'intégration de données socio-fiscales est sans doute la plus grande avancée de l'EDP ces dernières années. Ces données administratives exhaustives sont issues des déclarations servant à établir l'impôt sur le revenu et la taxe d'habitation, complétées par des données sur les prestations sociales. Elles sont utilisées à des fins statistiques et permettent à l'EDP d'enrichir les domaines couverts, grâce à l'introduction du niveau de vie, variable essentielle à de nombreuses analyses socio-économiques.

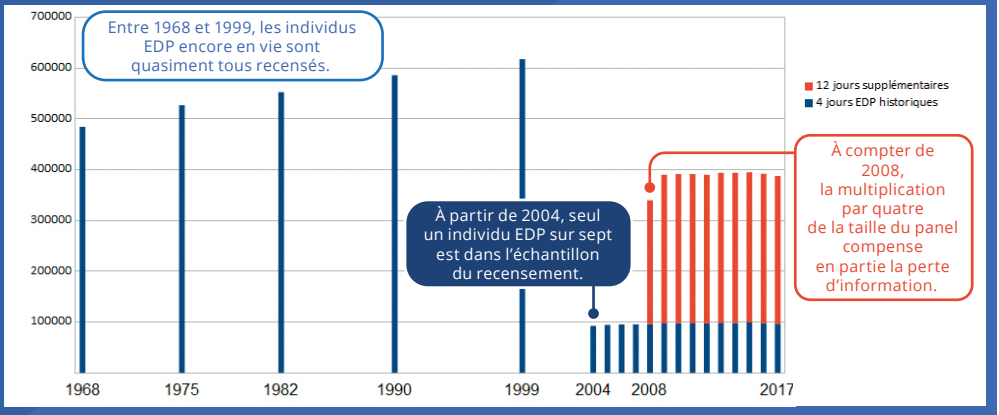
5. En pratique la date de passage varie selon les sources : 2004 pour l'état civil, 2008 pour les EAR, 2002 pour le panel « tous salariés », 2011 pour les données fiscales (revenus perçus en 2010), rétrospectif depuis le début des années 1990 pour les inscriptions électorales.
6. Le 1er janvier a été exclu car c'est trop fréquemment la date retenue lorsque le jour de naissance est inconnu (Insee, 2019).
7. L'intégration de données fiscales à l'EDP répondait aussi à des recommandations du Cnis (Chaleix et Lollivier, 2004).

… ET S'ENRICHIT ANNUELLEMENT DE DONNÉES SUR LES REVENUS ET LE NIVEAU DE VIE

Alors que l'exploitation statistique des données fiscales en dehors de l'EDP ne permet un suivi individuel des niveaux de vie que sur deux années successives[8], l'intégration dans l'EDP de données issues des dispositifs Fidéli et Filosofi offre désormais un suivi sur plus longue période.

Cette panélisation des données fiscales depuis la déclaration fiscale de 2011 (revenus 2010), pour l'échantillon des personnes nées un « jour EDP », ouvre de larges perspectives d'études. Citons à ce titre l'étude récente sur l'évolution du niveau de vie des retraités (Abbas, 2020) : au moment des débats sur la réforme des retraites, cette analyse éclaire non seulement la situation financière des retraités l'année de leur départ à la retraite, mais aussi l'évolution de leur niveau de vie au cours des trois années qui le précèdent et des trois années qui le suivent. Elle met ainsi en lumière la dégradation des conditions

Figure 2. Le passage à 16 jours permet de compenser en partie l'impact du recensement rénové



Nombre d'individus EDP recensés, selon la date de recensement

Lecture : Parmi les 3,7 millions d'individus présents dans l'EDP, 484 200 sont présents au recensement de 1968 et 617 000 sont présents au recensement de 1999. 92 000 sont présents à l'EAR de 2004, effectifs plus réduits du fait du passage de recensement exhaustif à des enquêtes annuelles par sondage. L'extension de l'EDP à 16 jours de naissance au lieu de 4 a permis d'augmenter les effectifs suivis à compter de 2008. Une année donnée, près de 400 000 individus EDP sont recensés dans une EAR. L'effectif est plus réduit en 2008 que pour les années suivantes, la procédure pour identifier les personnes EDP ayant été plus limitée cette année-là. Seules les personnes dont on avait pu retrouver le numéro d'identification au répertoire (NIR) automatiquement au répertoire national d'identification des personnes physiques ont été intégrées, sans traitement complémentaire (Jugnot, 2014).

Source : (Insee, 2019)

8. En dehors d'une étude ponctuelle réalisée avec un suivi de 5 ans (Bonnet, Garbinti et Solaz, 2015).

de vie pour certains en fin de carrière et l'amélioration de leur situation financière après leur départ à la retraite, notamment pour les retraités les moins diplômés. Cette étude inédite n'aurait pas été possible sans l'EDP : elle combine des caractérisations sociales (données du recensement) et la connaissance des ressources au fil des années (données fiscales en panel), disponibles uniquement dans l'EDP.

Les trajectoires individuelles suivies dans l'EDP se sont ainsi étoffées ***(encadré 2)***. L'introduction des données socio-fiscales exhaustives a aussi permis des études inédites sur la situation des familles après une rupture d'union, rares étant les données permettant de telles approches, comme le souligne le Cnis (Thélot *et alii*, 2016). Il en est ainsi de l'étude sur le logement des familles l'année de la rupture d'une union et dans les années qui la suivent (Durier, 2017). Ou des analyses de l'évolution du niveau de vie après une rupture d'union, qui révèlent une forte baisse en moyenne l'année de la rupture pour les femmes et une « récupération » ensuite dans les années qui suivent (Costemalle, 2017), amélioration qui s'observe surtout pour les parents formant rapidement une nouvelle union (Abbas et Garbinti, 2019).

La force de l'EDP réside aussi dans la taille de son échantillon : l'étude sur le niveau de vie après une séparation a ainsi été déclinée pour la région en Auvergne-Rhône-Alpes (Balouzat et Labosse, 2020), ce qui n'aurait pas été possible avec un échantillon national moins important.

À la fois riche par la diversité des sources qui l'alimentent, la finesse géographique potentielle des analyses du fait de la taille de l'échantillon et la profondeur historique d'un suivi sur plus de 50 ans ***(encadré 3)***, l'EDP est une source incontournable pour les études en panel : près de 60 équipes de recherche travaillent actuellement à partir de ces données (CASD, 2021). Mais revers de la médaille, cette richesse s'est accompagnée d'une plus grande complexité pour pouvoir exploiter ces données.

CONSTRUIRE SA POPULATION D'INTÉRÊT…

L'EDP permet de croiser les informations individuelles de plusieurs sources : on peut par exemple étudier la mortalité (données d'état civil) selon à la fois le diplôme (recensement), la catégorie sociale (recensement ou panel tous salariés) et le niveau de vie (données fiscales). La richesse des informations va toutefois de pair avec la complexité de l'usage des données : réfléchir à la manière de construire sa population d'intérêt et à la nécessité ou non de pondérer les données est un préalable incontournable avant de mener une étude à partir de l'EDP. Ces pré-requis font que l'EDP s'adresse à des chargés d'études ou chercheurs à l'aise avec l'exploitation des données et ayant de solides compétences statistiques, pour ne pas introduire de biais dans les résultats et analyses.

> *« C'est en combinant des informations des différentes sources intégrées à l'EDP que chaque chargé d'études construit sa population d'intérêt et les données nécessaires à son étude. »*

C'est en combinant des informations des différentes sources intégrées à l'EDP que chaque chargé d'études construit sa population d'intérêt et les données nécessaires à son étude ***(encadré 1)***. Il doit alors faire appel à différentes tables statistiques de la base études de l'EDP ***(figure 3),*** reliées entre elles par un identifiant commun ***(encadré 4)***.

Encadré 1. Quelles informations sont disponibles dans l'EDP ?

1967 — 2000 — 2011

PAR ANNÉE , DEPUIS 1968 ET 1967 POUR LA NAISSANCE DES INDIVIDUS EDP

- **Bulletin de naissance de l'individu EDP :** sexe, date de naissance, lieu de naissance (en France), date de naissance des parents, lieu de naissance des parents
- **Bulletin de naissance de ses enfants**
- **Bulletin de mariage :** date et lieu, informations sur les conjoints (date de naissance, lieu de naissance, état matrimonial légal antérieur)
- **Bulletin de décès :** date, lieu, état matrimonial légal

RECENSEMENTS 1968, 1975, 1982, 1990, 1999 ET ENQUÊTES ANNUELLES DE RECENSEMENT DEPUIS 2004

- **Individu :** informations sur l'individu EDP et les personnes résidant dans le logement (sexe, date de naissance, lieu de naissance, vie en couple, situation conjugale légale ou de fait, diplôme, lieu d'étude, CS, activité professionnelle, situation sur le marché du travail, lieu de travail, moyen de transport etc.), lieu de résidence antérieur
- **Famille :** situation familiale de individu EDP
- **Ménage-logement :** année d'emménagement, nombre de pièces, nombre d'habitants du logement, statut d'occupation, confort, région, département, commune, zone infra-communale

NB :

- *Les informations intégrées peuvent être différentes selon les années.*
- *Dom depuis 2004*

PANEL TOUS SALARIÉS DEPUIS 1967

Activité salariée, agrégée par année : salaire, nombre d'heures, CS, lieu de résidence, lieu de travail, condition d'emploi, activité économique

NB :

- *Champ variable selon les années : introduction de la fonction publique depuis 1988, Dom et secteur agricole depuis 2002, particuliers employeurs depuis 2009*
- *Prochainement le panel non salariés viendra compléter les données*

INSCRIPTIONS SUR LES LISTES ÉLECTORALES APPROXIMATIF AVANT LES ANNÉES 2000, FIABLE DEPUIS

Commune d ' inscription, date d'inscription, date de radiation

Note : Des données peuvent ne pas être disponibles certaines années ou pour certains individus EDP, notamment pour les années antérieures à l'extension de l'EDP à 16 jours, ou pour d'autres raisons (restriction de traitements des données d'état civil des individus EDP à certains jours de naissance par exemple dans les années 1990 pour certaines données). Le schéma ci-dessus offre un rapide panorama des informations contenues dans l'EDP. La consultation de la documentation détaillée est indispensable pour tout projet d'exploitation.

REVENUS ET COMPOSITION DES FOYERS FISCAUX ET MÉNAGES, PAR ANNÉE

Depuis 2011 (revenus 2010)

- **Individu :** informations sur l'individu EDP et ceux des logements où cet individu est connu au titre de l'impôt sur le revenu ou de la taxe d'habitation (date de naissance, situation conjugale, département commune de l'adresse fiscale, type de déclaration fiscale)
- **Logement :** informations sur le logement au sens de la taxe d'habitation principale (nombre de pièces, confort du logement, année d'emménagement, statut d'occupation)
- **Revenus (Filosofi) :** niveau de vie du ménage, composition du revenu, nombre de personnes
- **Revenus individuels :** revenu déclaré par l'individu EDP

Source : (Insee, 2019).

Par exemple, pour estimer les espérances de vie par niveau de vie, catégorie sociale et, diplôme, étude réalisée pour la première fois en 2018 grâce à l'intégration des données fiscales dans EDP, (Blanpain, 2018a), il a fallu sélectionner des personnes présentes au recensement (diplôme, CS), pour lesquelles on a recherché des données de l'état civil (état vital) et des données fiscales (niveau de vie). La constitution de sa base d'études a nécessité des expertises préalables. L'auteur a comparé l'ampleur des différentiels sociaux de mortalité (Blanpain, 2016) selon que l'on retient la catégorie sociale d'après le recensement ou d'après le panel tous salariés (Costemalle, 2016). Il a imputé un niveau de vie à partir de variables de revenus lorsque l'information sur le niveau de vie n'était pas disponible. Ceci est en effet le cas pour des personnes ne résidant pas en logement ordinaire par exemple, qui sont souvent des personnes âgées, et pouvait donc avoir un impact sur la mesure de la mortalité par niveau de vie (Blanpain, 2018b). L'auteur a également comparé sa population cible à d'autres données pour vérifier que la population d'intérêt qu'elle avait sélectionnée dans l'EDP était bien représentative de l'ensemble de la population étudiée (comparaison des espérances de vie estimées avec l'EDP à celles issues des bilans démographiques (Blanpain, 2018b)) et pouvoir recaler si besoin la population sélectionnée.

Une fois la population sélectionnée dans l'EDP, il faut se demander si elle représente bien la population générale, et se poser donc la question des pondérations.

… SAVOIR PONDÉRER

Le passage du recensement exhaustif aux enquêtes de recensement a introduit une nouvelle pratique pour les études menées à partir de l'EDP : les pondérations. La sélection sur les jours de naissance adoptée dans l'EDP ne biaise pas les analyses (aux limites mentionnées *supra*). Seul un facteur d'échelle était parfois utilisé pour donner des ordres de grandeur des effectifs concernés (par exemple en multipliant les effectifs concernés dans l'EDP par 365/4 ou 365/16 selon les années concernées). Mais estimer des répartitions ou des coefficients de modèles économétriques ne nécessitait pas toujours l'usage de poids : tous les individus avaient le même poids.

Or depuis 2004, le tirage de l'échantillon des enquêtes annuelles de recensement (EAR) marque une différence entre les petites et les grandes communes : ceci rend indispensable l'usage des pondérations associées aux EAR dans l'EDP dès lors qu'elles interviennent dans la définition de la population d'intérêt. Des variables de pondérations ont ainsi fait leur apparition dans l'EDP avec la fin de l'exhaustivité du recensement.

Mais on peut aussi utiliser les EAR sans utiliser ces pondérations, si par exemple cette source sert uniquement à compléter d'autres données en ajoutant des variables complémentaires notamment.

Ainsi, dans l'étude de l'évolution du niveau de vie des personnes parties à la retraite en 2013 (Abbas, 2020), une analyse est menée par diplôme. La population d'intérêt a été définie à partir des données fiscales et de la déclaration de revenus sous forme de pensions. Les informations pour ces personnes ont été complétées par le niveau de diplôme retrouvé dans une des EAR disponibles dans l'EDP. Il n'y a alors pas de raison de prendre en compte dans ce cas les pondérations des EAR, une fois vérifié que la population pour laquelle on a pu associer un niveau de diplôme ne diffère pas de la population cible totale (âge, niveau de vie, etc.).

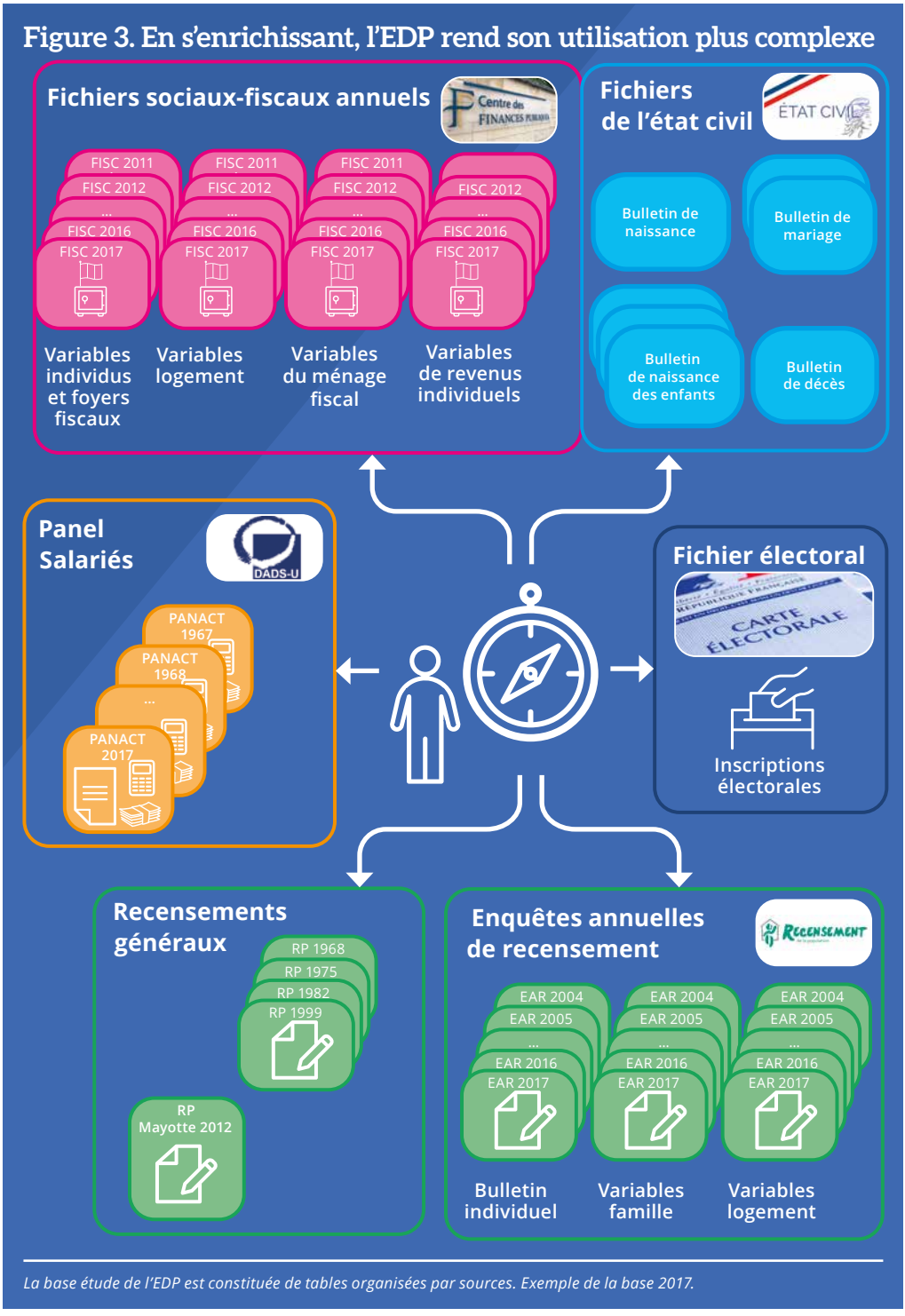
En contrepartie de la richesse des données contenues dans l'EDP, son usage s'est donc complexifié. C'est pourquoi un groupe d'exploitation a été créé en 2015 par l'Insee. Il réunit des chargés d'études et des chercheurs, afin d'échanger sur les nouveautés introduites dans l'EDP et sur les travaux réalisés à partir de ce panel.

Encadré 2. Un exemple fictif de trajectoire individuelle

Source	Informations obtenues
État civil	Un garçon nait un jour EDP en 1966 à Caen dans le Calvados, il est de parents résidant à Courseulles-sur-Mer (Calvados). On connaît leur année de naissance et leur date de mariage.
Recensement général de 1968	En 1968, il vit toujours à Courseulles-sur-Mer.
Recensement général de 1975	En 1975, toujours à Courseulles-sur-Mer, dans une maison de 5 pièces avec ses parents et trois autres enfants âgés de 7 à 16 ans (on ne connaît pas leur âge exact).
Recensement général de 1982	En 1982, il réside toujours à Courseulles-sur-Mer.
Fichier électoral	En 1984, à 18 ans, il s'inscrit sur les listes électorales à Courseulles-sur-Mer.
Recensement général de 1990	En 1990, il vit dans une maison de 2 pièces en location à Hérouville-Saint-Clair (Calvados). On connaît son diplôme et sa profession. Il est célibataire (état matrimonial) et vit en couple avec une femme née en 1968 dans le Calvados.
Fichier électoral	Il se ré-inscrit sur les listes électorales de Hérouville-Saint-Clair (on connaît la date).
État civil	Il a deux jumelles (on connaît leurs date et commune de naissance ainsi que celles de leur mère et la commune de résidence). Les parents ne sont pas mariés.
État civil	Il se marie (dates du mariage et de naissance de sa conjointe, commune de résidence).
Recensement général de 1999	En 1999, il est propriétaire d'une maison de 7 pièces à Hérouville-Saint-Clair et réside dans ce logement depuis 1996. Il y vit avec sa femme et ses filles. On connaît son diplôme, la commune de son lieu de travail et sa profession.
Panel salariés	En 2002 (et jusqu'en 2012), il est employé dans la fonction publique hospitalière de Caen. On connaît son salaire brut.
Enquête annuelle de recensement	En 2008, il réside à Hérouville-Saint-Clair. On connaît sa date d'emménagement. Il vit avec sa femme et trois enfants. On connaît leur date et lieu de naissance, leur emploi, diplôme et commune du lieu de travail.
Enquête annuelle de recensement	En 2013, divorcé, il vit à Hérouville-Saint-Clair dans une maison de 2 pièces en couple avec une femme divorcée et un enfant. On connaît leur date et lieu de naissance, diplôme, profession et commune du lieu de travail.
Fichiers socio-fiscaux Fideli et Filosofi	Divorcé (date du divorce), il a sa résidence fiscale de 2011 à 2013 à Hérouville-Saint-Clair et réside avec une femme (date et le lieu de naissance). Ils ont un enfant à charge (année de naissance). Il verse une pension alimentaire. On connaît son niveau de vie.
État Civil	Il décède le 12 décembre 2013 à Caen.

Note : la finalité de l'EDP n'est pas de suivre un individu, mais d'établir des statistiques sur les trajectoires d'un groupe d'individus définis par des caractéristiques socio-démographiques pour les analyser au regard d'autres groupes par exemple.

Figure 3. En s'enrichissant, l'EDP rend son utilisation plus complexe



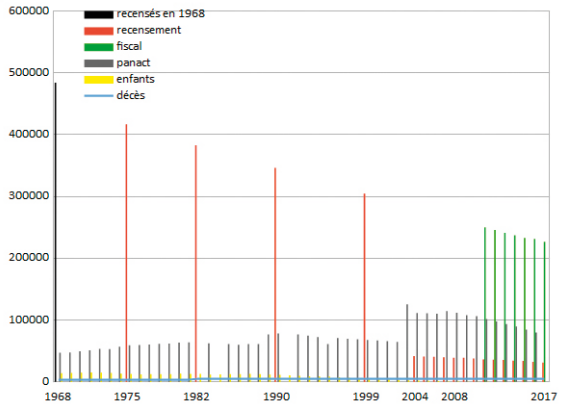
La base étude de l'EDP est constituée de tables organisées par sources. Exemple de la base 2017.

Encadré 3. Que sont devenues les personnes recensées en 1968?

484 200 personnes nées un 1er, 2, 3 ou 4 octobre, quelle que soit leur année de naissance, et recensées en 1968 sont suivies dans l'EDP. Au fil des années, elles sont présentes sur le marché du travail, ont des enfants, sont recensées, et décèdent pour une partie :

- pour 416 900 (soit 86 %) d'entre elles, on dispose d'informations statistiques les concernant au recensement de 1975 ;
- pour 346 200 personnes, on dispose de données du recensement de 1968 et de données du recensement de 1999, pour décrire leurs trajectoires sur 30 ans donc ;
- le nombre de personnes retrouvées aux recensements post-1968 décroît avec les années, car il y a des décès et de potentielles migrations ;
- il baisse fortement avec le passage des recensements exhaustifs aux enquêtes annuelles de recensement, du fait d'un recueil des données désormais sur un échantillon de la population : ainsi on dispose pour 41 600 individus du panel EDP recensés en 1968 d'informations les concernant à l'enquête annuelle de recensement (EAR) de 2004, et pour 31 400 de données de l'EAR de 2017, année la plus récente intégrée à l'EDP à ce jour. Mais on retrouve beaucoup plus de personnes avec des informations issues des données fiscales : **226 700 personnes recensées en 1968 ont aussi dans l'EDP des informations statistiques les concernant dans les données fiscales de 2017, avec un suivi statistique sur près de 50 ans.**

L'approche par cohorte peut aussi se faire par année de naissance. **On peut ainsi suivre de la même manière le devenir des 9 200 personnes nées les 1er, 2, 3 ou 4 octobre 1968 à partir des événements retracés dans l'EDP.**



Champ : individus nés entre les 1er et 4 octobre (quelle que soit l'année) et dont on dispose dans l'EDP d'informations statistiques dans le recensement de 1968 - Source : (Insee, 2019).

Note : l'effectif des naissances des enfants dont au moins un parent est né un jour EDP et a été recensé en 1968 est estimé en multipliant par 2 les effectifs pour les parents EDP nés un 1er ou un 4 octobre, pour tenir compte de trous de collecte (Durier, 2018).

UN ACCÈS AUX DONNÉES TRÈS ENCADRÉ

On dispose dans l'EDP d'une information de plus en plus précise sur les personnes qui composent le panel ***(encadré 1)***, mais pas de données directement identifiantes. Cependant, la répétition d'informations dans le temps rend ces données plus sensibles au risque de non-respect des critères d'anonymisation. En effet, si on connaît une personne dont la date de naissance correspond à un jour EDP, le croisement de ses autres caractéristiques connues avec les informations contenues dans le panel pourrait conduire à l'identifier dans le panel, selon une probabilité plus forte lorsque les informations sont répétées dans le temps que si on dispose uniquement de caractéristiques à une date donnée. Il y a donc un risque d'apprendre plus d'informations sur cette personne que celles que l'on connaît déjà, et donc d'atteinte potentielle à la confidentialité. C'est pour cette raison que la constitution et l'accès aux données de l'EDP sont très réglementés ***(encadré 5)***.

« Pour les chercheurs, l'accès se fait par l'entremise du CASD. »

À l'Insee, les chargés d'études peuvent accéder aux données dans un espace dédié après demande nominative, et ils exploitent les données dans un autre espace, dédié aux traitements.

Pour les chercheurs, l'accès se fait par l'entremise du CASD, Centre d'accès sécurisé aux données (Gadouche, 2019), après avis favorable du comité du secret sur leur projet. Ce mode d'accès, en vigueur depuis 2010, a permis de développer les exploitations de l'EDP, sur des thématiques aussi différentes que les inégalités territoriales, la mobilité géographique des immigrés en France, les parcours professionnels et les transitions de carrières entre secteurs public et privé.

Pour les agents des services statistiques ministériels, la situation varie selon les conditions de sécurité pour les accès aux données au sein du SSM. Au besoin, ils peuvent aussi recourir à l'intermédiaire du CASD.

Encadré 4. Un suivi en panel exigeant un identifiant unique et invariant

Depuis son origine, le suivi longitudinal des personnes présentes dans l'EDP s'est appuyé sur le NIR, numéro d'identification au Répertoire national d'identification des personnes physiques. Cet identifiant est unique et invariant*. Les personnes appartenant à l'échantillon suivi dans l'EDP sont dans un premier temps identifiées à partir de leurs traits d'identité (nom, prénom, sexe, date et lieu de naissance (Jugnot, 2014)) : il s'agit, à partir de ces informations, de retrouver leur NIR, pour ensuite enrichir leurs trajectoires dans le panel. Pour des sources, comme le panel « tous salariés », qui contiennent déjà le NIR, cette procédure d'identification n'est évidemment pas nécessaire*. **Le NIR n'est utilisé qu'à des fins de production de l'EDP, et les noms et prénoms ne sont pas conservés une fois l'identification réalisée.**

Les fichiers mis à disposition des chargés d'études et des chercheurs à des fins statistiques ne contiennent par le NIR, mais uniquement un identifiant de diffusion non signifiant (qui n'apporte donc pas d'information sur la personne). Cet identifiant de diffusion leur permet ainsi de réaliser les appariements entre les différentes bases de l'EDP.

** En toute rigueur, le NIR peut être modifié à de très rares exceptions (changement de genre par exemple, qui modifiera alors le premier chiffre du NIR).*

DES PROJETS EN COURS ET À VENIR

L'intégration des données socio-fiscales a permis de développer des études sur un champ nouveau : l'évolution du niveau de vie suite à un événement, et dans les années qui précèdent ou suivent cet événement. Elle a suscité l'intérêt des chercheurs en économie ou sur les familles : citons par exemple le projet *Big_Stat*, des données statistiques massives pour observer une société mobile (Ined, 2021). Elle a également élargi la possibilité d'analyses régionales (Lacour, 2018, Bertaux *et alii*, 2019, Balouzat et Labosse, 2020, Dherbécourt et Kenedi, 2020).

L'élargissement des sources de l'EDP va se poursuivre prochainement, avec le panel non salariés, pour couvrir de plus larges trajectoires d'emploi : l'EDP comprendra des données annuelles sur l'activité salariée et sur l'activité non salariée, toujours sur longue période, permettant ainsi d'analyser les trajectoires entre différents types d'emploi au fil de la carrière, en lien avec les caractéristiques socio-démographiques des individus (diplôme, famille, etc.). Par le cumul d'informations depuis maintenant plus de 50 ans, l'EDP permet déjà de suivre des trajectoires longues, combinant notamment les parcours de vies familiale et professionnelle[9].

L'EDP s'ouvre également à un nouveau domaine, dans un cadre juridique spécifique : celui de la santé. En 2019, la Drees a apporté à l'EDP les informations du système national des données de santé (SNDS) : cette source, gérée par la Caisse nationale de l'assurance maladie comprend notamment les consommations de soins et les causes médicales de décès, sur 10 ans mais sans informations ni sur les revenus, les milieux sociaux ou les situations professionnelles et familiales. Les données de santé faisant l'objet de procédures spécifiques, ce traitement, autorisé par la Cnil, est limité dans le temps (5 ans) (Drees, 2020). Il s'inscrit dans une finalité délimitée, pour évaluer la stratégie nationale de santé 2018-2022. Il permet de répondre à des questions sur l'évaluation des inégalités sociales de santé, et de compléter les analyses menées avec les données de santé en panel, mais ne comprenant que peu de descripteurs sociaux. La première publication de la Drees avec l'« EDP-santé », offre ainsi un complément d'analyse sur l'interruption volontaire de grossesse, sur la fréquence des IVG selon le niveau de vie (Vilain *et alii*, 2020).

Enfin, l'émergence de nouvelles sources sur les carrières professionnelles et les motifs d'interruption[10] offrent des pistes qui ne demandent qu'à être explorées.

Encadré 5. Le cadre juridique de l'EDP en quelques mots

L'EDP est un traitement de données à caractère personnel mis en œuvre en conformité avec le Règlement général sur la protection des données et la loi Informatique et Liberté. À ce titre, il fait l'objet de règles et de mesures strictes garantissant la sécurité et la confidentialité des données.

Toute personne ayant accès aux données est astreinte au secret statistique. Les chargés d'études et les chercheurs peuvent obtenir communication des données, mais après avis du Comité du secret statistique, en application des dispositions de la loi n°51-711 du 7 juin 1951 sur l'Obligation, la coordination et le secret en matière de statistiques. Les données identifiantes, notamment le NIR, ne sont cependant pas communicables dans ce cadre.

9. Des études longitudinales sur les parcours professionnels peuvent être menées à partir de panels sur les salariés ou les non salariés. Mais l'EDP intègre en plus des données socio-démographiques, sur les situations familiales par exemple.
10. Voir l'article de Christian Sureau et Richard Merlen sur le Répertoire général des carrières unique, dans ce même numéro.

BIBLIOGRAPHIE

ABBAS, Hicham et GARBINTI, Bertrand, 2019. De la rupture conjugale à une éventuelle remise en couple : l'évolution des niveaux de vie des familles monoparentales entre 2010 et 2015. In : *France, portrait social, édition 2019*. [en ligne]. 19 novembre 2019. Insee. Pp. 99-113. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4238447/FPS2019_D1.pdf*.

ABBAS, Hicham, 2020. *Des évolutions du niveau de vie contrastées au moment du départ à la retraite.* [en ligne]. 12 février 2020. Insee Première, n°1792. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4308750*.

BALOUZAT, Bruno, LABOSSE, Aline, 2020. *Lors d'une séparation, les femmes basculent plus souvent dans la pauvreté que leur conjoint.* [en ligne]. Octobre 2020. Insee Analyses Auvergne-Rhône-Alpes, n° 103. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4774341*.

BERTAUX, Frédéric, BOUSSAD, Nadia et SAGOT, Mariette, 2019. *En quinze ans, la moitié des Franciliens résidant dans des espaces « pauvres » ont changé de commune.* [en ligne]. 30 septembre 2019. Insee Analyses Île-de-France, n° 104. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4218834*.

BLANPAIN, Nathalie, 2016. *L'espérance de vie par catégorie sociale et par diplôme – Méthode et principaux résultats.* [en ligne]. 18 février 2016. Insee. Documents de travail, Direction des Statistiques Démographiques et Sociales, n° F1602. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/2022138*.

BLANPAIN, Nathalie, 2018a. *L'espérance de vie par niveau de vie : chez les hommes, 13 ans d'écart entre les plus aisés et les plus modestes*. [en ligne]. 6 février 2018. Insee Première, n°1687. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/3319895*.

BLANPAIN, Nathalie, 2018b. *L'espérance de vie par niveau de vie – Méthode et principaux résultats.* [en ligne]. 6 février 2018. Documents de travail, Direction des Statistiques Démographiques et Sociales, n° F1801. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/3322051*.

BONNET, Carole, GARBINTI, Bertrand et SOLAZ, Anne, 2015. Les variations de niveau de vie des hommes et des femmes à la suite d'un divorce ou d'une rupture de Pacs. In : *Couples et familles*. [en ligne]. 16 décembre 2015. Insee Références, pp. 51-61. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/2017528/COUFAM15.pdf*.

CASD, 2021. EDP : Échantillon Démographique Permanent. In : *site du Centre d'accès sécurisé aux données.* [en ligne]. Les sources de données déjà disponibles au CASD. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.casd.eu/source/echantillon-demographique-permanent/*.

CHALEIX, Mylène et LOLLIVIER, Stéfan, 2004. *Outils de suivi des trajectoires des personnes en matière sociale et d'emploi.* [en ligne]. Juin 2004. Cnis, Mission Panels, note n° 98/B010, class. 1.5.91. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2018/03/RAP_2004_98_outils_trajectoires_personnes_emploi.pdf*.

CNIS, 2006. *Dynamique et trajectoires. Compte rendu de la séance du 3 avril 2006.* [en ligne]. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2018/01/DPR_2006_6e_reunion_GT_inegalites_dynamique_territoire.pdf*.

COSTEMALLE, Vianney, 2016. *Catégorie sociale d'après les déclarations annuelles de données sociales et catégorie sociale d'après le recensement : quels effets sur les espérances de vie par catégorie sociale ?* [en ligne]. 18 février 2016. Insee. Documents de travail, Direction des Statistiques Démographiques et Sociales, n° F1603. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/2022136*.

COSTEMALLE, Vianney, 2017. Formations et ruptures d'unions : quelles sont les spécificités des unions libres ? In : *France, portrait social, édition 2017.* [en ligne]. 21 novembre 2017. Insee. Pp. 95-111. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3197289/FPORSOC17.pdf*.

COUET, Christine, 2006. L'échantillon démographique permanent de l'Insee. In : *Courrier des statistiques.* [en ligne]. Insee. N° 117-119, pp. 5-14. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/metadonnees/source/fichier/echantillon_demograp_permanent_courrier_stat117_119.pdf*.

DHERBÉCOURT, Clément et KENEDI, Gustave, 2020. *Quelle influence du lieu d'origine sur le niveau de vie ?* [en ligne]. 12 juin 2020. France Stratégie. La note d'analyse, n°91. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/fs-2020-na91-niveau-territoire-juin.pdf*.

DREES, 2020. *L'EDP-Santé : enrichissement de l'échantillon démographique permanent par les données du système national des données de santé (SNDS).* [en ligne]. 8 juillet 2020. Direction de la recherche, des études, de l'évaluation et des statistiques. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/ledp-sante-enrichissement-de-lechantillon-demographique-permanent-par*.

DURIER, Sébastien, 2017. *Après une rupture d'union, l'homme reste plus souvent dans le logement conjugal.* [en ligne].17 juillet 2017. Insee focus n°91. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/2896525*.

DURIER, Sébastien, 2018. Une nouvelle source de données sur la famille : l'EDP enrichi de données socio-fiscales. In : *Observer, décrire et analyser les structures familiales.* [en ligne]. Édité par Nicolas Cauchi-Duval. Association internationale des démographes de langue française. Pp. 5-15. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.erudit.org/fr/livres/actes-des-colloques-de-lassociation-internationale-des-demographes-de-langue-francaise/volume-5-observer-decrire-et-analyser-les-structures-familiales/000364li.pdf*.

GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la data science et la recherche scientifique. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254227/courstat-3-7.pdf*.

GODINOT, Alain, 2005. *Pour comprendre le recensement de la population*. [en ligne]. Insee Méthodes hors série mai 2005. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/information/2579979*.

INED, 2021. Big_Stat. In : *site de l'Ined.* [en ligne]. Institut national des études démographiques. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://big-stat.site.ined.fr/*.

INSEE, 2019. *Base études 2017 de l'Échantillon Démographique Permanent, manuel de l'utilisateur.* [en ligne]. 25 avril 2019. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://utiledp.site.ined.fr/fichier/s_rubrique/26440/manuel_edp_be2017.fr.pdf*.

JUGNOT, Stéphane, 2014. *La constitution de l'échantillon démographique permanent de 1968 à 2012.* [en ligne]. 19 septembre 2014. Insee. Documents de travail, Direction des Statistiques Démographiques et Sociales, n° F1406. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/1381113/F1406.pdf*.

LACOUR, Cédric, 2018. *Les séparations : un choc financier, surtout pour les femmes*. [en ligne]. 16 octobre 2018. Insee Analyses Nouvelle-Aquitaine n° 64. Institut national des études démographiques. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/3631116*.

SAUTORY, Olivier, 1988. Plus de la moitié de la population a changé au moins une fois de commune en vingt ans. In : *Économie et statistique.* [en ligne]. Avril 1988. Insee. N° 209, pp. 39-47. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.persee.fr/doc/estat_0336-1454_1988_num_209_1_5197*.

THÉLOT, Claude, BOURREAU-DUBOIS, Cécile, CHAMBAZ, Christine, 2016. *Les ruptures familiales et leurs conséquences : 30 recommandations pour en améliorer la connaissance*. [en ligne]. Mars 2016. Cnis, rapport de groupe de travail. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAPPORT-RUPTURES-FAMILIALES-_-nouvelle-version-29mai2017.pdf*.

VILAIN, Annick, ALLAIN, Samuel, DUBOST, Claire-Lise, FRESSON, Jeanne et REY, Sylvie, 2020. *Interruptions volontaires de grossesse : une hausse confirmée en 2019.* [en ligne]. Septembre 2020. Drees. Études et Résultats, n°1163. [Consulté le 30 mai 2021]. Disponible à l'adresse : *https://www.repere.re/fileadmin/user_upload/DREES_ivg_2019.pdf*.

FONDEMENTS JURIDIQUES

Arrêté du 6 août 2014 portant création d'un traitement automatisé de données à caractère personnel relatif à l'échantillon démographique permanent de l'INSEE. In : *site de Légifrance*. [en ligne]. [Consulté le 30 mai 2021]. Disponible à l'adresse : *http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000029556902*.

LE RÉPERTOIRE DE GESTION DES CARRIÈRES UNIQUE (RGCU)

UN NOUVEAU RÉFÉRENTIEL ET DES PERSPECTIVES POUR L'ANALYSE SOCIALE

Christian Sureau et Richard Merlen***

Le système de retraite français est composé de régimes par grands types de professions : ils gèrent les informations relatives à la carrière de leurs assurés pour, au moment de la retraite, calculer leurs droits et ensuite verser leurs pensions. En 2010, il a été décidé de construire un référentiel unique contenant le détail des carrières de l'ensemble de la population française. Le Répertoire de gestion des carrières unique (RGCU) doit permettre à terme d'optimiser les processus de gestion, en centralisant les données et en améliorant leur complétude et leur qualité.

Constituer un référentiel impose d'abord de structurer les données, quelle qu'en soit l'origine, avec des concepts et une sémantique uniques. L'automatisation des flux d'alimentation constitue ensuite un atout majeur pour l'amélioration continue de la qualité du répertoire.

Le RGCU contiendra alors une information très détaillée sur la carrière des assurés, avec les périodes et les revenus d'activité salariée, ou d'inactivité liée au chômage et à la maladie. Une telle richesse des données en largeur (ensemble de la population) comme en profondeur (totalité de la carrière détaillée en périodes et revenus depuis l'origine des régimes), devrait amener le RGCU à devenir une source précieuse pour des études sociales.

The French pension system is made up of schemes by major types of professions: they manage information on the careers of their insured persons in order to calculate their rights and then pay their pensions when they retire. In 2010, it was decided to build a single repository containing details of the careers of the entire French population. The RGCU (Répertoire de Gestion des Carrières Unique) ***should ultimately optimise management processes by centralising data and improving its completeness and quality.***

Building a repository first requires the data, whatever its origin, to be structured with unique concepts and semantics. The automation of the feed-in flows is then a major asset for the continuous improvement of the quality of the directory.

The RGCU will then contain very detailed information on the career of insured persons, with periods and income from paid employment, or inactivity due to unemployment and illness. Such a wealth of data in terms of both breadth (the entire population) and depth (the entire career detailed in terms of periods and income since the origin of the schemes), should lead the RGCU to become a valuable source for social studies.

* Directeur du programme RGCU, Cnav,
christian.sureau@cnav.fr
** Directeur du programme RGCU pour le régime général, Cnav,
richard.merlen@cnav.fr

LE RÉPERTOIRE DE GESTION DES CARRIÈRES UNIQUE : UNE VOLONTÉ D'AMÉLIORATION DU SERVICE RETRAITE

La vieillesse a longtemps été considérée comme un risque : celui, devenant trop vieux, de ne plus pouvoir subvenir à ses besoins. Si quelques dispositifs pour des catégories professionnelles particulières existaient depuis longtemps, c'est en 1945, avec la création de la Sécurité sociale, que l'ambition a alors été d'apporter une protection à l'ensemble de la population (Damon et Ferras, 2020). L'existence préalable de dispositifs de pensions pour une partie de la population et le souhait des non-salariés de conserver leur organisation propre ont amené à construire un système basé selon les catégories professionnelles. Ainsi existe-t-il aujourd'hui en France de nombreux régimes de retraite de base et de retraite complémentaire[1].

Au moment où l'assuré prend sa retraite, les régimes[2] doivent disposer d'une vision complète et détaillée de la carrière de la personne depuis son tout premier emploi, ainsi que de l'ensemble des informations permettant d'appliquer les différentes règles et de déterminer le montant de la pension à verser.

Or chaque régime dispose de son propre système d'information, avec ses assurés, le détail de leur carrière et les modules permettant de calculer leurs droits. Cette multiplicité complexifie pour l'assuré la visualisation de sa carrière ou l'estimation de ses droits, mais aussi les démarches à faire au moment de préparer sa retraite, puisqu'il devra prendre contact avec les différents organismes auprès desquels il aura cotisé. Enfin, cette diversité ne facilite pas la mise en œuvre de réformes de retraite.

Aussi, en 2010, la loi[3] a instauré le **Répertoire de gestion des carrières unique (RGCU)** et prévu qu'il soit alimenté par les différents régimes de base. Le RGCU a ensuite été étendu en 2014, toujours par la loi, aux régimes complémentaires. La construction de ce référentiel a été confiée à la Cnav (Caisse nationale d'assurance vieillesse).

Le RGCU doit apporter des progrès significatifs dans au moins quatre directions. Il permet tout d'abord d'accroître la performance de la gestion, avec notamment l'amélioration de la complétude et la vérification de la cohérence des carrières au fil de l'eau, la réduction des échanges entre régimes et avec l'assuré, la réduction des délais de reconstitution des carrières et la maîtrise des risques. Il renforce ensuite l'efficience du processus de liquidation de la retraite, grâce à des carrières plus fiables, des informations mieux partagées entre régimes, et une réduction de la durée du cycle de liquidation. Avec ce nouveau dispositif, les démarches de l'assuré et son dialogue avec les régimes se simplifient, grâce au développement d'offres de service innovantes. Enfin, le RGCU permet de construire une brique commune qui facilitera l'intégration des évolutions réglementaires et la mise en œuvre des futures réformes.

1. [N.D.L.R.] Cette variété des systèmes de retraite français a été également abordée dans de précédents articles du Courrier des statistiques, voir (Bellanger et Goujon, 2020) et (Cheloudko et Martin, 2020).
2. Le terme de « régime » désigne, en toute rigueur, un dispositif réglementaire de retraite applicable à une certaine population. Dans la suite de l'article, il désignera les organismes en charge de sa mise en œuvre : recouvrement des cotisations, calcul et paiement de la pension. De même le terme de « carrière » désigne le plus souvent la manière dont la carrière professionnelle d'un assuré est enregistrée dans les systèmes d'information.
3. Article 9 de la loi n° 2010-1330 du 9 novembre 2010 portant réforme des retraites. Voir les références juridiques en fin d'article.

Le terme de RGCU désigne aussi le projet informatique qui a conçu le dispositif et le met progressivement en œuvre : celui-ci possède les caractéristiques des « grands programmes », compte-tenu des enjeux, de la charge en hommes jours et du nombre d'acteurs impactés ***(encadré 1)***. Entamée en juillet 2019 avec un premier régime pilote[4], la mise en production a été étendue en mai 2020 au régime général. Le RGCU contient aujourd'hui les carrières de 80 millions d'assurés, actifs ou déjà retraités.

UN RÉFÉRENTIEL AU SERVICE DE L'ASSURÉ DÈS SON ENTRÉE DANS LA VIE ACTIVE…

Le « métier » de la retraite, celui exercé par les « régimes de retraite », s'organise toujours autour de tout un historique concernant l'assuré. En simplifiant, on peut considérer qu'il y a cinq événements fondamentaux qui déterminent les périodes de cet historique : la naissance, le début d'activité professionnelle, la demande de retraite, la liquidation de la retraite et le décès.

1. La période qui s'écoule **de la naissance au début d'activité professionnelle**[5] ne requiert, pour l'essentiel, pas d'activité particulière de la branche retraite, excepté bien sûr l'immatriculation de l'individu par l'Insee au répertoire national d'identification des personnes physiques (RNIPP), qui l'amène à devenir un assuré et donc un utilisateur futur des services offerts par le RGCU.

2. **La vie active** va du premier emploi à la retraite : outre le recueil des revenus, l'appel des cotisations et le recouvrement pour les organismes concernés, la mission du régime de retraite durant cette période est d'enregistrer, de façon aussi fiable et complète que possible, les éléments relatifs à la carrière, de gérer les événements afférents, et d'informer l'assuré.

3. **Le passage à la retraite** s'étend de la demande de retraite à la liquidation : dans ce bref laps de temps, qui pour de nombreux régimes concentre le cœur du métier, l'organisme complète et vérifie les éléments de la carrière de l'assuré, pour ensuite les traduire en attribution de droits, puis en montants de pensions. Pour ce faire, il échange des informations avec l'assuré et les autres régimes, sur les pièces justificatives reçues et sur le paiement de cotisations, puis il applique les règles de droit.

4. **La retraite**, période plus ou moins longue, va se dérouler ensuite jusqu'au décès : la mission du régime de retraite est de procéder au paiement régulier des pensions, en temps et heure et en cohérence avec le droit, tout en gérant les événements afférents.

5. Dans **la période qui suit le décès**, le service de la retraite s'arrête pour le décédé, mais peut aussi se poursuivre, immédiatement ou plus tard, à la suite d'une demande du conjoint survivant, par le versement d'une pension de réversion ; le traitement administratif nécessaire peut être assez complexe du fait de la coordination entre régimes à établir.

Bien entendu cette vision est simplificatrice, car les périodes se recouvrent : ainsi, on peut avoir une vie active après la retraite, et opérer un cumul emploi – retraite. Par ailleurs, des révisions sont toujours possibles pendant la période de retraite.

4. La CRPCEN, Caisse de retraite et de prévoyance des clercs et employés de notaires, qui compte 240 000 assurés.
5. Ou période assimilée, comme le service national, voir *infra.*

… MAIS AUSSI DES SERVICES POUR LES ORGANISMES CONCERNÉS

Le référentiel contient, bien sûr, les informations relatives à la carrière de l'assuré. Mais il fournit également un ensemble de services pour les régimes ou les autres organismes de protection sociale : il leur permet d'alimenter le RGCU et de contrôler les flux d'alimentation, de restituer les carrières, de les consulter, les saisir, ou de calculer les données nécessaires à l'application des droits de l'assuré.

Durant la vie de l'assuré, les « alimentations » du référentiel sont effectuées soit par les entreprises qui l'emploient, pendant ses périodes salariées, soit par les organismes de protection sociale, tels que Pôle Emploi pour les périodes de chômage ou la Cnam (Caisse nationale d'assurance maladie) pour les périodes de maladie. Le programme RGCU a défini les processus d'alimentation et de contrôle permettant de garantir la qualité des flux de données.

Pendant cette période, mais surtout au moment où l'assuré prend sa retraite, le régime, en lien avec l'assuré, va compléter la carrière et la valider au moyen de services ou d'applications de consultation et de saisie « clé en mains ».

Lors de la validation de la carrière, le régime de retraite va déterminer les droits de l'assuré avec l'application dédiée au calcul des durées d'assurance ou de points et celle effectuant le calcul du salaire annuel moyen nécessaire au calcul des pensions[6] ***(figure 1)***.

Auparavant, pour pouvoir utiliser les données et les services du RGCU, le régime qui intègre le dispositif devra alimenter le référentiel avec son stock de données. C'est le **processus de migration**. Un premier service lui est alors proposé : le « kit de migration » va appliquer les contrôles de structure et de cohérence avant de charger les données dans le répertoire. Ainsi, le régime fournit le fichier de déchargement de sa base sous un même format « pivot » commun à tous.

Encadré 1. Quelques chiffres sur le RGCU

- 80 millions d'assurés
- 7 milliards d'éléments de carrières (périodes et revenus des activités salariées, de chômage, de maladie, etc.)
- Plus de 130 000 jours de travail
- 3 100 unités de calcul (CPU), 850 serveurs
- Première mise en production en 2019

6. Dans le dispositif français des retraites en vigueur depuis 1945, les modalités de calcul des pensions ont évolué, mais nécessitent encore à ce jour, pour les salariés du secteur privé, de s'appuyer sur un salaire annuel moyen.

UNE STRUCTURE DE DONNÉES UNIFIÉE, POUR ACCUEILLIR DES ÉLÉMENTS D'ORIGINES DIVERSES

La structure des données du référentiel doit permettre de disposer des éléments qui retracent la continuité de la carrière de l'assuré, quels que soient sa profession et le régime de retraite auquel il est affilié pendant ses périodes dites « travaillées ». Elle doit aussi être adaptée aux périodes de chômage, de maladie, ou aux périodes de vie ne donnant pas de droits à la retraite mais qui complètent la carrière depuis le début de sa vie professionnelle.

Figure 1. Un référentiel et des services, tout au long de la vie active de l'assuré

L'organisation retenue dans le RGCU comprend trois niveaux ***(figure 2)***.

- La **carrière brute** est composée des éléments détaillés de la carrière avec, pour chacun d'eux, la période concernée et les éléments de revenu, voire de cotisation pour certains régimes. L'information sur les périodes travaillées provient des entreprises ou d'organismes[7]. La carrière brute recouvre aussi les périodes de chômage, de maladie, de maternité, d'invalidité, de service national, de formation, d'apprentissage, de garde d'enfants, etc. ***(figure 3)***. Certaines informations peuvent être spécifiques au type de période ou au régime auquel l'assuré est affilié. Les données sur l'état de validation de la donnée, sa traçabilité, les résultats de l'analyse de la carrière (anomalies, incohérences, etc.) complètent aussi la carrière brute.
- La **carrière étendue** comprend, au-delà de la carrière brute, les périodes de vie qui ne donnent pas lieu à des droits mais permettent de s'assurer qu'il n'y a pas de « trous » dans le parcours d'un assuré (année sabbatique par exemple). Elle comprend aussi les événements qui peuvent donner droit à des majorations de durée d'assurance (trimestres supplémentaires suite à éducation des enfants par exemple). Le concept de carrière étendue retrace ainsi intégralement le parcours professionnel et non professionnel de la vie d'un assuré, depuis son premier emploi salarié jusqu'à sa retraite.
- La **carrière élargie** apporte des informations complémentaires sur les enfants (données déclaratives) ou les affiliations de l'assuré aux différents régimes et les dates de liquidation par régime de retraite. La carrière élargie contiendra aussi les informations de « valorisation » : elles restituent les différentes durées d'assurance par régime et tous régimes confondus (nombre de trimestres par an).

Le contenu d'une carrière va ainsi présenter l'ensemble des activités d'un assuré avec par exemple ses emplois d'été ou de stagiaire lorsqu'il était étudiant, ses périodes d'activité professionnelle, avec le salaire, l'identification de l'employeur[8], l'origine de l'information, mais aussi les périodes de chômage ou de maladie.

« Le RGCU contient l'historique de mise à jour de toutes les carrières. »

Pour chaque nouvelle mise à jour d'un élément existant, un nouvel élément est créé en historisant le précédent. Ainsi, le RGCU contient l'historique de mise à jour de toutes les carrières.

Au-delà d'un objectif de complétude, le RGCU vise à disposer de données précises relatives aux principaux types de période.

Les **périodes professionnelles** concernent les activités des salariés du secteur privé, mais également des fonctionnaires civils et militaires, les stages de formation professionnelle, l'apprentissage, les emplois à domicile ou les activités intermittentes, et enfin, bien sûr, des périodes d'activité indemnisées ou pas. Pour ces types de période, le référentiel renseigne les dates de début et de fin de période, l'exercice juridique, le statut catégoriel (cadre, non cadre, etc.), la profession ou catégorie socio-professionnelle des emplois salariés des secteurs privés et publics (selon la nomenclature de l'Insee), le nombre de jours ou d'heures travaillés, le revenu, le taux cotisé, etc.

7. Acoss/Urssaf (Agence centrale des organismes de sécurité sociale/Unions de recouvrement des cotisations de sécurité sociale et d'allocations familiales) pour les Cesu (chèque emploi service universel) par exemple.
8. Numéro Siret d'inscription au répertoire Sirene.

Les **périodes assimilées**, par exemple le service national, les périodes de chômage (indemnisé ou pas) ou de maladie, disposent aussi des dates de début et de fin de période et du revenu perçu. Les périodes de maladie détaillent la typologie (congé maladie, accident du travail, invalidité, etc.), les catégories d'invalidité (totale, partielle, etc.) et l'origine de l'invalidité (par accident du travail ou autre).

Pour les **périodes informatives** qui permettent de compléter la carrière de l'assuré mais ne sont pas prises en compte dans le calcul des droits, les informations disponibles sont les dates de début et de fin ainsi que quelques données spécifiques.

UNE NORMALISATION DES CONCEPTS POUR L'ENSEMBLE DES ALIMENTATIONS

Le RGCU ayant un caractère de référentiel, une attention particulière doit être apportée à la modélisation des objets métiers, leur sémantique et les nomenclatures utilisées par les régimes qui intègrent le dispositif.

Figure 2. Une structure des données à trois niveaux

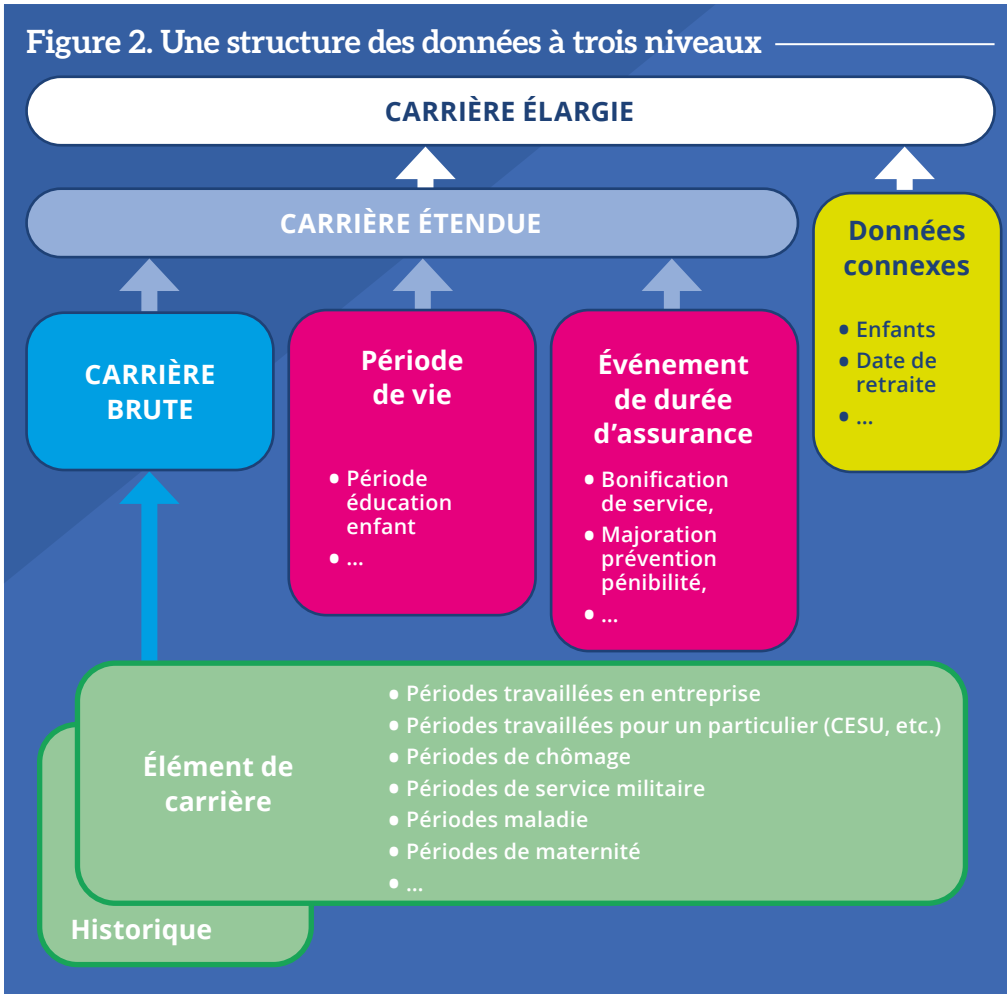


Figure 3. Un exemple de carrière retracée dans le RGCU

La carrière d'un assuré est composée d'éléments correspondant à...

- des périodes travaillées...
- des périodes assimilées...
- et des périodes informatives.

01/01/1993	31/12/1993	Emploi de droit privé	Entreprise Z	Validé	83 519,78	FRF	DADS
01/04/1992	31/12/1992	Emploi de droit privé	Entreprise Z	Validé	48 261,85	FRF	DADS
02/01/1992	01/04/1991	Période assimilée	Chômage	Validé			
02/11/1991	01/01/1992	Période assimilée	Chômage	Validé			
02/01/1991	01/11/1990	Emploi de droit privé	Entreprise Y	Validé	65 648,62	FRF	DADS
01/01/1990	01/01/1991	Emploi de droit privé	Entreprise Y	Validé	9 563,93	FRF	DADS
09/03/1989	03/04/1989	Période assimilée	Maladie				
01/01/1989	31/12/1989	Emploi de droit privé	Entreprise Y	Validé	72 498,96	FRF	DADS
02/01/1988	31/12/1988	Emploi de droit privé	Entreprise Y	Validé	69 564,37	FRF	DADS
02/01/1987	01/01/1988	Emploi de droit privé	Entreprise Y	Validé	64 432,02	FRF	DADS
01/01/1986	01/01/1987	Emploi de droit privé	Entreprise Y	Validé	53 987,45	FRF	DADS
01/09/1985	31/12/1985	Emploi de droit privé	Entreprise Y	Validé	48 734,67	FRF	DADS
01/08/1985	30/08/1985	Période assimilée	Chômage	Validé			
01/08/1984	31/07/1985	Période assimilée	Service National	Validé			
03/07/1983	30/09/1983	Emploi de droit privé	Entreprise X	Non validé	13 564,56	FRF	DAS

TYPES D'ÉLÉMENTS DE CARRIÈRE ÉTENDUE

01 - Prestation en montant (C001)
02 - Période étude valorisée (C002)
03 - Période rachat (C003)
04 - Période à l'étranger (C004)
05 - Cotisation 30-46 (C005)
06 - Période assurance volontaire (C006)
07 - Elément carrière globalisé (C007)
08 - Activité indemnisée (C008)
09 - Activité non salariée (C009)
10 - Période validée gratuite (C010)
11 - Période activité équivalente (C011)
12 - Assurance vieillesse des parents au foyer (C014)
13 - Activité cultuelle (C015)
14 - Rente régime intégré (C016)
15 - Période salarié du secteur privé ou agent non titulaire de l'État et des collectivités publiques (CSE1)
16 - Période fonctionnaire civil ou militaire (CSE2)
17 - Stage de formation professionnelle salarié (CSE3)
18 - Apprentissage (CSE4)
19 - Période assimilée divers (CPA1)
20 - Période assimilée famille (CPA2)
21 - Période assimilée maladie-AT (CPA3)
22 - Période assimilée chômage (CPA4)
23 - Activité intermittente (CAA1)
24 - Emploi à domicile (CAA2)
25 - Période de vie religieuse (P001)
26 - Période éducation enfant (P002)
27 - Période informative assuré (P003)
28 - Période de prise en charge personne (P004)
29 - Bonification de service (E001)
30 - Contrat apprentissage (E002)
31 - Choix MDA enfant (E003)
32 - Durée d'assurance acquise (E004)
33 - Majoration prévention pénibilité (E005)
34 - Report SNGC (CMG1)
35 - Majoration SNGC (CMG2)
36 - Service validé (CSEV)
37 - Période organisation internationale et institution européenne (C017)
38 - Elément de droit (CDRO)
39 - Information professionnelle (IPRO)
40 - Période SNGC (CMG3)

TYPES DE PÉRIODES INFORMATIVES

01 - Période non travaillée validée gratuitement
02 - Période avec salaire issu d'une validation gratuite loi 26/12/1964
03 - Période validée gratuitement au titre d'ancien rapatrié pour l'année 1962
04 - Validation Congé Sans Solde (CSS) parental éducation (article 7)
05 - Congé parental d'éducation avec ou sans cotisations Invalidité Vieillesse Décès (IVD)
07 - Activité cultuelle
08 - Périodes de salariat effectuées avant le 01/07/1939 et/ou avant le 01/07/1951 pour celles effectuées en Algérie ou en Alsace-Moselle
09 - Temps alterné gratuit
10 - Congé parental alterné gratuit
11 - Congé de présence parentale
12 - Validation au titre de l'amiante
13 - Congé Sans Solde (CSS) pour élever un enfant de moins de 8 ans né après le 01/07/2008
14 - Congé parental enfants nés après le 01/07/2008
15 - Congé Sans Solde (CSS) enfant handicapé recueilli entre 8 et 20 ans
17 - Validation Congé Sans Solde (CSS) parental éducation enfants nés avant le 01/07/2008
18 - Période de salariat en Algérie 1938-1962
19 - Majoration 72

Le travail de définition du modèle d'objets a été mené et validé avec tous les régimes de retraite. Il apparaissait en effet essentiel de disposer d'objets métiers avec une sémantique parfaitement définie et stable. Sur cette base, lorsqu'un nouveau régime de retraite intègre le RGCU, chaque type de donnée est soumis à une analyse sémantique, afin de l'intégrer dans l'objet correspondant du référentiel. Ce faisant, on s'assure de stocker les données relatives aux mêmes concepts dans les mêmes objets.

Le RGCU fait l'objet d'une norme d'échange qui a été construite autant pour alimenter le référentiel que pour en restituer les données. Cette norme s'appuie sur :

- le modèle de données « de la norme », avec l'ensemble des objets et de leurs cardinalités ;
- les modèles de message pour alimenter ou restituer les données du RGCU ;
- la définition des données, qu'il s'agisse de données métier ou données de gestion, afin de partager leur sémantique ;
- les contrôles à appliquer lors de l'alimentation du RGCU, contrôles de structure et de cohérence (voir *infra*) ;
- une documentation précisant la structure des données, les messages et les règles de contrôle qui leur sont appliquées ;
- et une gestion des versions : mode de maintenance de la norme, plan du versionnage, du contenu et de la date d'effet de chaque version. Pour cela, la Cnav a développé un outil générique de gestion de norme, Saturne[9], permettant de définir la norme et de générer automatiquement documentations et outils de contrôle pour chaque version.

« La base « carrière » partagée doit disposer d'une nomenclature « unifiée » qui ne doit pas être une addition des nomenclatures issues de chaque régime de retraite. »

La rigueur appliquée à la définition des concepts, aux objets et aux données du RGCU doit aussi concerner les nomenclatures. Ces dernières sont des listes de valeurs (par exemple les natures d'emploi) qui se caractérisent par un code et un libellé. La base « carrière » partagée doit disposer d'une nomenclature « unifiée » qui ne doit pas être une addition des nomenclatures issues de chaque régime de retraite.

Les nomenclatures se basent autant que possible sur celles existant chez les partenaires gérant les risques concernés (DSN, etc.). Toutefois, pour tenir compte de spécificités de certains régimes, elles peuvent avoir été complétées de nouvelles valeurs.

UN PROCESSUS ET DES OUTILS QUI PERMETTENT DE GARANTIR LA QUALITÉ D'ALIMENTATION

Une fois les concepts des objets et nomenclatures normalisés, les processus d'alimentation doivent ensuite garantir la qualité des informations qui intègrent le référentiel. Les processus de contrôle de la qualité interviennent à trois moments ***(figure 4)*** :

- tout au long de l'alimentation de la carrière, avec les différents contrôles associés à la norme évoquée *supra* ;

9. Saturne est utilisé de la même manière pour la norme de la DSN dans toutes ses versions, voir (Cnav, 2021 ; net-entreprise.fr, 2021).

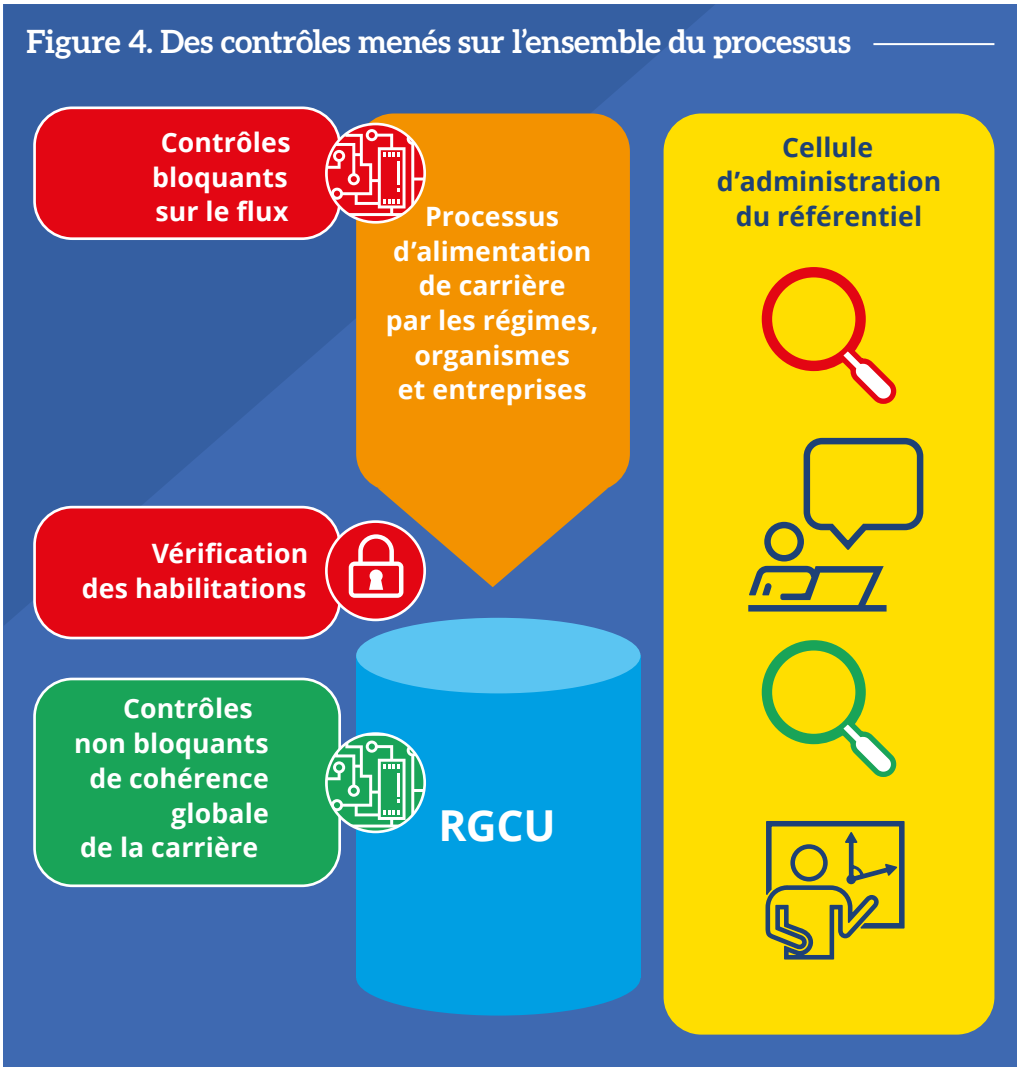
- lors de l'analyse de la carrière par un technicien à l'initiative d'un régime ou suite à des demandes de l'assuré ;
- de manière périodique, par une cellule spécifique en charge d'analyser les indicateurs de qualité du référentiel et de définir les plans d'actions pour améliorer les données *a posteriori*.

Lors de l'alimentation, des contrôles bloquants et des contrôles d'habilitations sont réalisés pour déterminer si les données peuvent être intégrées dans le RGCU.

Les **contrôles bloquants** sont appliqués directement sur le flux d'alimentation. Ils permettent de vérifier le respect de la structure du flux, avec les règles syntaxiques ou sémantiques des données qu'il transporte. Voici deux exemples de règles :

- (Contrôle de structure) Si la rubrique *Type de Prestation en montant* est égale à « 01 - Complément indemnités journalières maladie », alors les rubriques suivantes sont obligatoires : *Salaire soumis à cotisation* et *Salaire brut* ;

Figure 4. Des contrôles menés sur l'ensemble du processus



- (Contrôle de cohérence) La rubrique *Date de fin d'activité non salariée* doit être postérieure ou égale à la rubrique *Date de début d'activité non salariée*.

Les **contrôles d'habilitation** vérifient les droits de l'« alimenteur » de créer, modifier, supprimer chaque type d'élément de carrière. Par exemple, un régime ne pourra pas supprimer ou modifier un élément alimenté par un autre organisme si celui-ci ne l'y a pas autorisé.

Lors du processus d'alimentation de la carrière, après l'écriture dans le RGCU ou lors d'opérations spécifiques d'analyse de la carrière, des **contrôles non bloquants** sont appliqués et leurs résultats stockés dans le référentiel. Les contrôles non bloquants n'empêchent donc pas l'alimentation des données dans le RGCU : ils signalent des incohérences potentielles. Ils relèvent de trois types :

- les **contrôles métiers** que l'on peut appliquer sur un élément de carrière, sans avoir à analyser sa cohérence avec la totalité de la carrière existante. Ils interviennent sur le flux d'alimentation ou bien sur le stock *a posteriori*. On y trouve des contrôles de validité de la date (par exemple, la date de début de l'élément de carrière doit être strictement postérieure au 01/01/1946). Cet ensemble recouvre aussi des contrôles faisant intervenir des référentiels autres que le RGCU (par exemple, s'assurer que la date de début d'un élément de carrière est postérieure à la date de naissance de l'assuré nécessite de récupérer des données du référentiel identification (fichier miroir du RNIPP)) ;
- les **contrôles de cohérence référentielle** consistent à vérifier la cohérence entre des éléments de carrière transmis et la carrière déjà présente dans le RGCU pour le même assuré. C'est typiquement le cas des contrôles de non-chevauchement entre périodes : par exemple, on transmet dans le flux des périodes d'activité, et on va vérifier que ces périodes ne se superposent pas avec d'autres périodes existantes. La spécificité des contrôles de cohérence référentielle (qui les différencient des contrôles métier), est qu'ils font intervenir au moins deux éléments de carrière ;
- les **contrôles d'atypie** ont une portée globale et concernent l'ensemble de la carrière de l'assuré. Ils en vérifient la conformité relativement à une situation que l'on prend comme référence. Une atypie est un élément intrinsèquement juste mais atypique par rapport à ceux qui l'entourent (exemple d'un salaire anormalement élevé par rapport aux autres sur une période donnée).

Les signalements issus des contrôles métiers et des contrôles de cohérence référentielle ont une portée qui dépasse la seule gestion des données : ils empêchent l'attribution de droits de retraite à l'assuré pour les éléments de vie concernés. Ils nécessitent donc d'être corrigés avant la liquidation de la carrière.

En complément des contrôles automatisés, il a été créé au sein de la Cnav, une équipe dédiée[10] appelée « **Cellule d'administration du référentiel** », qui définit, produit et analyse des tableaux de bord. En fonction du niveau des indicateurs (sur le pourcentage d'éléments rejetés sur les flux par exemple), des plans d'actions sont mis en œuvre avec les différentes parties prenantes (les organismes qui alimentent le RGCU, l'équipe du programme RGCU, etc.). La cellule participe ainsi à la boucle qui permet de garantir une amélioration continue de la qualité du référentiel.

10. La cellule compte de 5 à 6 personnes en 2021.

UNE RICHESSE D'INFORMATIONS POUR LES ASSURÉS ET POUR LES RÉGIMES DE RETRAITE

Ce qui fait la « richesse » des informations du RGCU, c'est d'abord le nombre d'assurés que le référentiel couvre : 80 millions d'individus y figurent, actifs ou déjà retraités. Le détail des données et la profondeur historique apportés par l'ensemble des « alimenteurs » ***(encadré 2)*** constituent également des atouts. Or la qualité de ces données, notamment leur complétude, dépendent de facteurs bien identifiés :

- le nombre de régimes de retraite qui ont migré leurs carrières sur le RGCU ;
- la qualité de leur stock de données au moment de la migration ;
- la complétude et la qualité des flux d'alimentation ;
- et enfin le statut de l'assuré, retraité ou non.

> *« Environ 50 % des éléments de carrière détaillés tous régimes confondus sont déjà intégrés dans le RGCU, ce qui représente maintenant 7 milliards d'éléments. »*

Depuis 2020, 80 millions d'assurés ont donc été « chargés » dans le RGCU avec les données détaillées de trois régimes : le régime général, le régime des clercs et employés de notaire, le régime des métiers du culte, et le régime complémentaire AGIRC-ARRCO. Le RGCU contient aussi le nombre annuel de trimestres cotisés pour les régimes dits de base, et le revenu annuel pour les salariés agricoles et les travailleurs indépendants. Au total, environ 50 % des éléments de carrière détaillés tous régimes confondus sont déjà intégrés dans le RGCU, ce qui représente maintenant 7 milliards d'éléments.

Le champ des assurés et leurs éléments de carrière seront complétés lors des opérations de migration des autres régimes de retraite, selon un calendrier prédéfini ***(figure 5)***. Tant qu'un régime n'a pas encore effectué la migration de ses données vers le RGCU, le référentiel n'intègre que des données consolidées[11] sur les éléments qu'il couvre. Celles-ci seront remplacées par les données détaillées lors de la migration.

Comme indiqué *supra*, durant la vie de l'assuré, la grande majorité des informations de la carrière sont intégrés sous forme de flux informatiques, en provenance des employeurs[12], de Pôle Emploi pour les périodes de chômage, de la Cnam pour les périodes de maladie, de la Cnaf[13], de l'Urssaf Caisse nationale[14] et enfin des régimes de retraite eux-mêmes. La plupart des flux émis par ces organismes sont intégrés au RGCU, le reliquat ne concernant qu'un nombre limité de cas (chômage non indemnisé involontaire, etc.).

A contrario, les autres périodes de vie (congé sabbatique, etc.) peuvent ne pas être intégrées au moment où elles se produisent. C'est lors de la liquidation de la retraite que l'échange va avoir lieu entre le régime de retraite et l'assuré, pour vérifier sa carrière et la compléter au besoin avec la fourniture des justificatifs manquants.

Après cette dernière étape, la carrière de l'assuré sera complète et vérifiée, et donc totalement fiable pour le calcul des droits à la retraite.

11. Selon les régimes, il s'agit de trimestres ou de revenus par année (voir *supra*).
12. *Via* les DADS-U (Déclaration des Données Sociales Unifiées) qui cèdent maintenant majoritairement la place aux DSN (Déclarations Sociales Nominatives) pour les entreprises du secteur privé (Humbert-Bottin, 2018).
13. Caisse nationale d'allocations familiales, pour les prestations d'Avpf (assurance vieillesse du parent au foyer).
14. Anciennement Agence centrale des organismes de sécurité sociale (Acoss), pour les chèques emplois services et les prestations d'accueil du jeune enfant (PAJE).

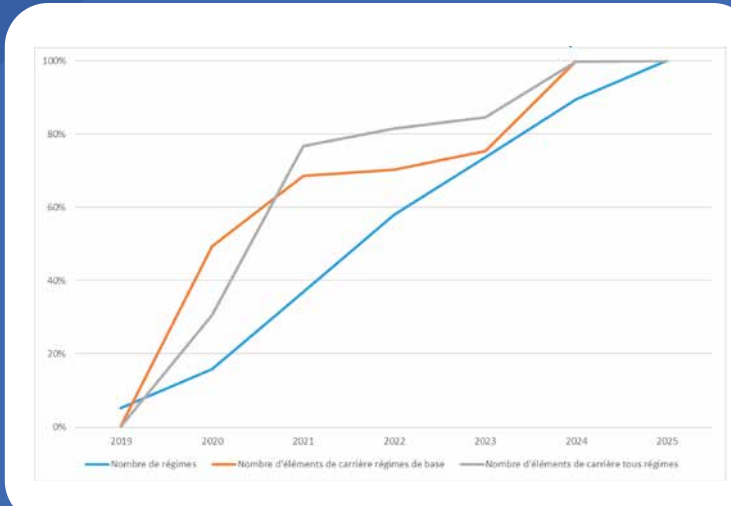
L'HISTOIRE ET L'INFORMATISATION DES FLUX, FACTEURS DÉTERMINANTS DE LA QUALITÉ DES DONNÉES

La complétude et le niveau de détail des données du RGCU sont très largement le fruit de l'histoire de la Sécurité sociale et de l'informatisation de la transmission des données sociales.

Pour les entreprises de droit privé, par exemple, les premières informations sur les périodes travaillées datent de 1930 : jusqu'en 1935, elles étaient saisies par les techniciens des régimes lorsque les assurés souhaitaient prendre leur retraite. Pour ces années, une seule période a été définie, contenant l'ensemble des cotisations. Ensuite, de 1936 à 1947, les cotisations ont été saisies année par année. À partir de 1947, ce sont les salaires (et non plus les cotisations) qui seront saisis par année.

Les données contenues dans le RGCU, pour ce qui relève du stock, sont issues des migrations successives de systèmes d'informations précédents. Les données du RGCU sont ainsi issues du SNGC (Système national de gestion des carrières) créé en 1998, lui-même successeur du FNCI (Fichier national des comptes individuels) construit au début des années 1980.

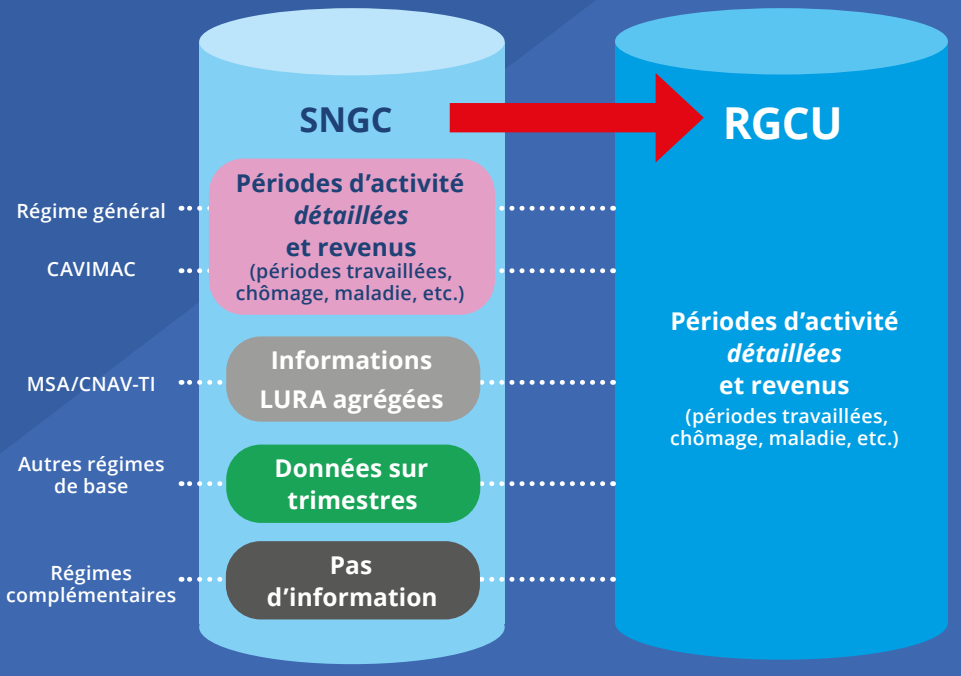
Figure 5. Une montée en charge progressive sur les 5 années à venir



Dans le FNCI, toutes les carrières avant 1970 ont fait l'objet de saisies à partir des informations communiquées par les caisses régionales[15] sans notion de période détaillée. Pour ces années, les salaires sont intégrés année par année, mais sans distinguer les employeurs. Les salaires après 1970 et jusqu'en 2000 sont ensuite reportés par année et par employeur. C'est à partir de l'année 2000 que les périodes ont été alimentées de date à date avec les salaires associés. Pour les années précédentes, les périodes ont été interprétées pour avoir la même structure de date à date ***(figure 6)***.

Le SNGC contenait les données détaillées des différentes périodes pour les employés des entreprises de droit privé relevant du régime général et pour les métiers du culte. Les informations pour les autres régimes de base étaient fournies en trimestre par année. Depuis 2016, les deux régimes de retraite pour les agriculteurs et pour les travailleurs indépendants alimentaient le SNGC avec leurs revenus par année, pour répondre à la mise en œuvre du dispositif Lura[16]. Les données des autres régimes que le régime général et celui des métiers du culte seront détaillées au fur et à mesure de leur intégration dans le RGCU.

Figure 6. Du SNGC au RCGU



SNGC : Système national de gestion des carrières créé en 1998
CAVIMAC : Caisse d'assurance vieillesse, invalidité et maladie des cultes
MSA : Mutualité sociale agricole
CNAV-TI : désigne le régime des travailleurs indépendants
LURA : liquidation unique des régimes alignés

15. Les caisses régionales d'assurance maladie (CRAM) dont les activités s'exerçaient surtout dans l'assurance vieillesse et les risques professionnels. En 2010, la loi les transforme en caisses d'assurance retraite et de la santé au travail (CARSAT), sauf en Alsace et dans le département de la Moselle, ainsi qu'en Île-de-France.
16. Liquidation unique des régimes alignés, permettant aux assurés de ces régimes de faire une seule demande de retraite et de ne percevoir qu'une seule pension.

Encadré 2. Liste des professions et des régimes associés prévus à terme

Statut		Régime de base
Salariés du secteur privé (hors agriculture et aviation civile)		CNAV (Caisse Nationale d'Assurance Vieillesse)
Agents non titulaires de la fonction publique		
Salariés de l'aviation civile		
Salariés du secteur privé agricole		**MSA-SA** (Mutualité Sociale Agricole)
Non salariés : Exploitants agricoles		**MSA-NSA** (Mutualité Sociale Agricole)
Artistes Auteurs		**CNAV** *avec les affiliations gérées par l'AGESSA* (*Association pour la Gestion de la Sécurité Sociale des Auteurs d'œuvres cinématographiques, musicales, photographiques et télévisuelles*)
Salariés relevant de régimes spéciaux	**- Employés des Mines**	
	- Clercs et employés de notaires	
	- Industries électriques et gazières	
	- RATP	
	- SNCF	
	- Les marins	
	- Banque de France	
	- Ouvriers de l'état	
	- Comédie Française	
	- Opéra de Paris	
	- Port autonome de Strasbourg	
Agents de la fonction publique territoriale et hospitalière		CNRACL (Caisse Nationale de Retraites des Agents des Collectivités Locales)
Fonctionnaires civils et militaires		SRE (Service des Retraites de l'Etat)
Non salariés : artisans, commerçants, industriels		
Non-salariés : professions libérales (hors avocats)	**- Chirurgiens-dentistes et des sages-femmes**	CNAVPL (Caisse Nationale d'Assurance Vieillesse des Professions Libérales) *La gestion est assurée par les caisses complémentaires*
	- Médecins de France	
	- Infirmiers, masseurs-kinésithérapeutes, pédicures-podologues, orthophonistes et orthoptistes	
	- Vétérinaires	
	- Agents généraux et des mandataires non-salariés de l'assurance et de la capitalisation	
	- Experts-comptables et des commissaires aux comptes	
	- Officiers ministériels, des officiers publics et des compagnies judiciaires	
	- Pharmaciens	
	- Autres professions libérales	
	- Notaires	
Non salariés (professions libérales) : avocats		
Non salariés : ministres des cultes		CAVIMAC (Caisse d'Assurance Vieillesse, Invalidité et MAladie des Cultes)
Le régime parlementaire du Sénat avec la caisse autonome de Sécurité sociale du Sénat		
Le régime parlementaire de l'Assemblée Nationale		

dans le RGCU

Régime complémentaire	
AGIRC-ARRCO (Association Générale des Institutions de Retraite des Cadres - Association pour le Régime de Retraite COmplémentaire des salariés)	
IRCANTEC (Institution de retraite complémentaire des agents non titulaires de l'État et des collectivités publiques)	
CRPN (Caisse de Retraite complémentaire du Personnel Naviguant)	
AGIRC-ARRCO	
MSA-RCO (Mutualité Sociale Agricole - Régime COmplémentaire)	
IRCEC (Institution de retraite complémentaire de l'enseignement et de la création)	**RAAP** (Régime de retraite des Artistes et Auteurs Professionnels)
	RACD (Régime de retraite des Auteurs et Compositeurs Dramatiques)
	RACL (Régime de retraite des Auteurs et Compositeurs Lyriques)
CANSSM (Caisse Autonome Nationale de la Sécurité Sociale dans les Mines)	
CRPCEN (Caisse de Retraite et de Prévoyance des Clercs et Employés de Notaires)	
CNIEG (Caisse Nationale des Industries Electriques et Gazières)	
CRPRATP (Caisse de Retraites du Personnel RATP)	
CPRPSNCF (Caisse de prévoyance et de retraite du personnel de la Société nationale des chemins de fer français)	
ENIM (Établissement National des Invalides de la Marine)	
BDF (Banque de France)	
FSPOEIE (Fonds spécial des pensions des ouvriers des établissements industriels de l'État)	
CRPCF (Caisse de Retraite du Personnel de la Comédie Française)	
CROPERA (Caisse de retraites des personnels de l'Opéra national de Paris)	
Régime de retraite des salariés du port autonome de Strasbourg	
RAFP (Retraite Additionnelle de la Fonction Publique)	
RAFP	
SSI (Sécurité Sociale des Indépendants)	
CARCDSF (Caisse Autonome de Retraite des Chirurgiens Dentistes et des Sages-Femmes)	
CARMF (Caisse Autonome de Retraite des Médecins de France)	
CARPIMKO (Caisse autonome de retraite et de prévoyance des infirmiers, masseurs kinésithérapeutes, pédicures-podologues, orthophonistes et orthoptistes)	
CARPV (Caisse de retraite pour les vétérinaires)	
CAVAMAC (Caisse d'Allocation Vieillesse des Agents Généraux et des Mandataires non-salariés d'Assurance et de Capitalisation)	
CAVEC (Caisse d'assurance vieillesse des experts comptables)	
CAVOM (caisse d'assurance vieillesse des officiers ministériels, des officiers publics et des compagnies judiciaires)	
CAVP (Caisse d'Assurance Vieillesse des Pharmaciens)	
CIPAV (Caisse interprofessionnelle de prévoyance et d'assurance vieillesse)	
CPRN (Caisse de Prévoyance et de Retraite des Notaires)	
CNBF (Caisse Nationale des Barreaux Français)	
AGIRC-ARRCO	
CASS SENAT (Caisse autonome de Sécurité sociale du Sénat)	
FSS Assemblée nationale (Fonds de Sécurité sociale de l'Assemblée Nationale)	

Les périodes militaires ont fait l'objet d'une reprise du passé en 2007, grâce aux informations fournies par le Bureau central des archives administratives militaires (BCAAM).

En complément des périodes travaillées, les périodes assimilées correspondent à des périodes non travaillées pouvant donner des droits à la retraite (voir *supra*). Elles ont été reportées dans le FNCI puis dans le SNGC, sous forme de trimestres acquis par année. Ce n'est qu'en 2001 pour le chômage, et en 2008 pour les périodes de maladie, invalidité et de rentes AT (arrêt de travail), que ces périodes assimilées ont été reportées de date à date dans le SNGC. Plus récemment, certaines dispositions nouvelles ont été prises en compte, comme les périodes assimilées pour les sportifs de haut niveau (2013) ou les points pénibilité (2016). Et pour ajouter à la complexité, certaines informations ne peuvent encore aujourd'hui être alimentées qu'en montants forfaitaires année par année (comme les Avpf, assurance vieillesse pour les parents au foyer).

Pour pouvoir assurer sa mission, le RGCU doit également s'adapter à l'évolution de la législation, nécessitant des informations complémentaires et parfois plus détaillées pour la bonne application des textes de loi.

UNE VOLONTÉ D'OUVERTURE POUR LA MISE À DISPOSITION DES DONNÉES DU RGCU

Le RGCU est en production depuis 2019. Il contient aujourd'hui les carrières détaillées des salariés du privé, du régime spécial des clercs et employés de notaire et les métiers des cultes, pour 80 millions d'assurés actifs ou déjà partis à la retraite. Il va être complété d'ici le début de l'année 2022 avec les salariés agricoles et les travailleurs indépendants.

C'est une source de données d'une extraordinaire richesse par le nombre de personnes concernées et par le détail, la complétude et la profondeur des informations contenues. À terme, toutes les personnes ayant commencé leur vie professionnelle seront intégrées dans le RGCU quelles que soient leurs professions. L'ensemble de la carrière de ces individus y est détaillé avec les périodes d'activité, les employeurs, les revenus, mais aussi les périodes de chômage, de maladie, etc. ainsi que les périodes sans activité.

Le RGCU constitue de ce fait une photo très détaillée de l'activité de l'ensemble de la population, mais il contient surtout le « film » : pour chaque individu depuis le début de la vie professionnelle jusqu'à la date de retraite.

Le RGCU offre ainsi de nombreuses pistes d'analyse, qu'il s'agisse de l'évolution des revenus, des activités exercées, des accidents de parcours selon les types de profession, et ce non pas sur un échantillon de population mais sur son ensemble, et en pouvant raisonner sur une longue période. Il représente potentiellement une mine d'informations nouvelle et prometteuse pour des études sociales.

C'est la raison pour laquelle, au-delà de l'apport sur l'efficience générale apportée à la gestion de la retraite, la Cnav et la direction de la Sécurité sociale souhaitent ouvrir l'utilisation du référentiel aux travaux de recherche. Ces données seront mises à disposition *via* le CASD (Centre d'accès sécurisé aux données[17]), de manière totalement pseudonymisée, d'ici le début de l'année 2022.

17. Voir (Gadouche, 2019) pour plus de détails sur le fonctionnement du CASD.

BIBLIOGRAPHIE

BELLANGER, Bryan et GOUJON, Samuel, 2020. Prisme, du régime général au régime universel, la microsimulation comme outil d'aide à la décision. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. N° N5, pp. 95-113. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008705/courstat-5-7.pdf*.

CHELOUDKO, Pierre et MARTIN, Henri, 2020. Une décennie de modélisation du système de retraite. La genèse du modèle de microsimulation TRAJECTOiRE. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. N° N4, pp. 23-41. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497060/courstat-4-3.pdf*.

CNAV, 2021. *RGCU- Cahier technique de la norme R.* 4 mai 2021. Caisse Nationale d'Assurance Vieillesse - MOA Normes. Version 3.21.06.

DAMON, Julien et FERRAS, Benjamin, 2020. *La Sécurité sociale*. 16 septembre 2020. Collection Que sais-je ? Tome n°4035. ISBN 978-2-7154-0431-1.

GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la data science et la recherche scientifique. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254227/courstat-3-7.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 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.

NET-ENTREPRISES.FR, 2021. *Norme DSN (NEODeS) et documentation technique.* [en ligne]. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.net-entreprises.fr/declaration/norme-et-documentation-dsn/*.

FONDEMENTS JURIDIQUES

Décret n° 2018-154 du 1er mars 2018 relatif au répertoire de gestion des carrières unique. In : *site de Légifrance*. [en ligne]. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000036666210*.

Loi n° 2010-1330 du 9 novembre 2010 portant réforme des retraites. In : *site de Légifrance*. [en ligne]. Mise à jour le 22 janvier 2014. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000023022127/*.

Loi n° 2014-40 du 20 janvier 2014 garantissant l'avenir et la justice du système de retraites. In : *site de Légifrance*. [en ligne]. Mise à jour le 25 décembre 2016. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000028493476/*.

UN OUTIL D'APPARIEMENT SUR IDENTIFIANTS INDIRECTS

L'EXEMPLE DU SYSTÈME D'INFORMATION SUR L'INSERTION DES JEUNES

*Loïc Midy**

Le service statistique du ministère de l'Éducation Nationale (la Depp) réalise depuis longtemps deux enquêtes annuelles d'insertion dans la vie active, portant sur les sortants de l'apprentissage et de la voie professionnelle scolaire. Mais elles ne permettent pas de publier des statistiques au niveau des établissements, comme requis depuis 2018 par la loi pour la Liberté de choisir son avenir professionnel. Afin de répondre à ce besoin, la Depp et la Dares ont construit un nouveau dispositif, appelé InserJeunes, qui apparie des sources administratives, principalement sur identifiants indirects. Cette problématique était centrale pour la réussite du dispositif, qu'il s'agisse des choix méthodologiques, du paramétrage des algorithmes ou des développements informatiques. Le processus choisi comporte, classiquement, cinq étapes : normalisation des données, indexation, calcul de similarités, classification supervisée et évaluation de la qualité. Le choix des méthodes adaptées est présenté à travers un cas réel de production : si elles ont été implémentées à travers un outil d'appariement développé spécifiquement pour InserJeunes, elles restent transposables dans des environnements similaires.

The French Statistical Office of the National Education Ministry (the DEPP) carry out two surveys on the labour market integration of the students who just finished their study as apprentice or in vocational school path. But they don't enable to publish statistics at the establishment level as required by the 2018 Act for the Liberty to choose one professional future. So the DEPP and the DARES (Statistical Office of the Labour Ministry) have designed a new information system, InserJeunes, based on the record linkage of administrative data sources. Record linkage is central in this device, from the methodological, the algorithmic and the IT development standpoints. In InserJeunes, the record linkage process has five steps: data normalisation, indexing, similarities calculation, supervised classification and quality evaluation. The methods are presented through a real production example from the InserJeunes information system. They were implemented through a record linkage tool developed by the InserJeunes team, which can be reused for other record linkage processes.

* Directeur du projet Mesure de l'insertion des jeunes, Depp,
loic.midy@education.gouv.fr

MIEUX CONNAÎTRE L'EFFICACITÉ DES ÉTABLISSEMENTS EN MATIÈRE D'INSERTION DES JEUNES

L'orientation des élèves se construit tout au long de la scolarité avec des étapes clés en fin de troisième, seconde et terminale. Ainsi, l'orientation en voie professionnelle peut commencer dès la fin de troisième avec un choix entre apprentissage ou voie professionnelle scolaire. L'insertion dans l'emploi étant la première finalité de la formation professionnelle, connaître les taux d'insertion des formations initiales permet d'éclairer les choix des jeunes et de leur famille.

La direction de l'Évaluation, de la prospective et de la performance (Depp) réalise, depuis le début des années 1990, deux enquêtes d'insertion annuelles[1] permettant de suivre l'entrée dans la vie active des sortants d'apprentissage et de voie professionnelle scolaire. Ces opérations apportent des informations précieuses, mais ne permettent pas de publier des statistiques au niveau établissement, compte tenu des taux de réponse observés[2].

Or, la loi du 5 septembre 2018 pour la Liberté de choisir son avenir professionnel[3] prévoit la publication de statistiques par établissement sur le parcours scolaire et l'insertion dans l'emploi des jeunes en formation professionnelle. Afin de répondre à ce besoin, la Depp et la direction de l'Animation de la recherche, des études et des statistiques (Dares) ont construit un nouveau dispositif[4] : InserJeunes est basé sur l'appariement de sources administratives exhaustives, relatives à la scolarité des élèves et apprentis, à la réussite aux examens, aux contrats d'apprentissage et aux contrats salariés de la déclaration sociale nominative (DSN[5]). Les premiers résultats ont été diffusés début février 2021.

> *« InserJeunes est basé sur l'appariement de sources administratives exhaustives, relatives à la scolarité des élèves et apprentis, à la réussite aux examens, aux contrats d'apprentissage et aux contrats salariés de la déclaration sociale nominative. »*

Côté Éducation nationale, les bases sur les élèves peuvent s'apparier sur un identifiant spécifique, l'identifiant national élève (INE)[6]. Mais il n'existe pas d'identifiant commun pour apparier les bases « scolarité » des élèves et apprentis, avec les contrats salariés de la DSN. Ces appariements ne sont donc possibles qu'indirectement, et réalisés à partir des cinq variables que sont les noms, prénoms, date et lieu de naissance et le sexe. La mise en place d'un outil d'appariement sur identifiants indirects performant et de qualité est donc un enjeu central d'InserJeunes. Cette problématique bien connue des statisticiens fait l'objet d'une vaste littérature (voir par exemple (Kilss et Alvey, 1985)).

Cet article présente la démarche d'ensemble retenue, les sources principales utilisées, le cadre juridique à respecter ainsi que les choix effectués entre les différentes méthodes et outils informatiques d'appariement sur identifiants indirects.

1. Les enquêtes *Insertion dans la vie active* (IVA) pour les sortants de la formation professionnelle des lycées, et *Insertion professionnelle des apprentis* (IPA).
2. De l'ordre de 60 %.
3. Voir les références juridiques en fin d'article.
4. InserJeunes a bénéficié d'un financement du fonds pour la transformation de l'action publique.
5. Pour plus d'information sur la DSN, voir (Humbert-Bottin, 2018).
6. L'INE, mis en place en 2017, est un identifiant unique de chaque élève.

LES PRINCIPES DU DISPOSITIF INSERJEUNES

Le processus principal s'articule autour de plusieurs phases ***(figure 1)***. Dans un premier temps, pour une année scolaire donnée, le champ des élèves en année terminale de formation est calculé en mobilisant trois bases de données administratives « scolarité »[7], chacune couvrant une partie du champ d'InserJeunes : l'apprentissage, la voie professionnelle scolaire dans un établissement du ministère de l'Éducation nationale et celle dans un établissement du ministère de l'Agriculture. Ces bases contiennent les variables indirectement identifiantes, ainsi que l'INE, et des informations sur l'établissement et la formation suivie.

Dans une deuxième phase, on établit le champ des élèves sortants, c'est-à-dire ceux qui ne sont plus en formation. Pour ce faire, on recherche, principalement sur l'INE, si ces élèves sont encore présents l'année scolaire suivante dans l'ensemble des bases de données élèves disponibles c'est-à-dire les trois bases déjà mobilisées dans la phase précédente ainsi que trois bases supplémentaires[8] afin d'être le plus exhaustif possible[9]. Tout élève retrouvé est noté comme étant toujours en étude, les autres sont appelés les sortants de formation[10]. Cela permet d'établir le **taux de poursuites d'études**.

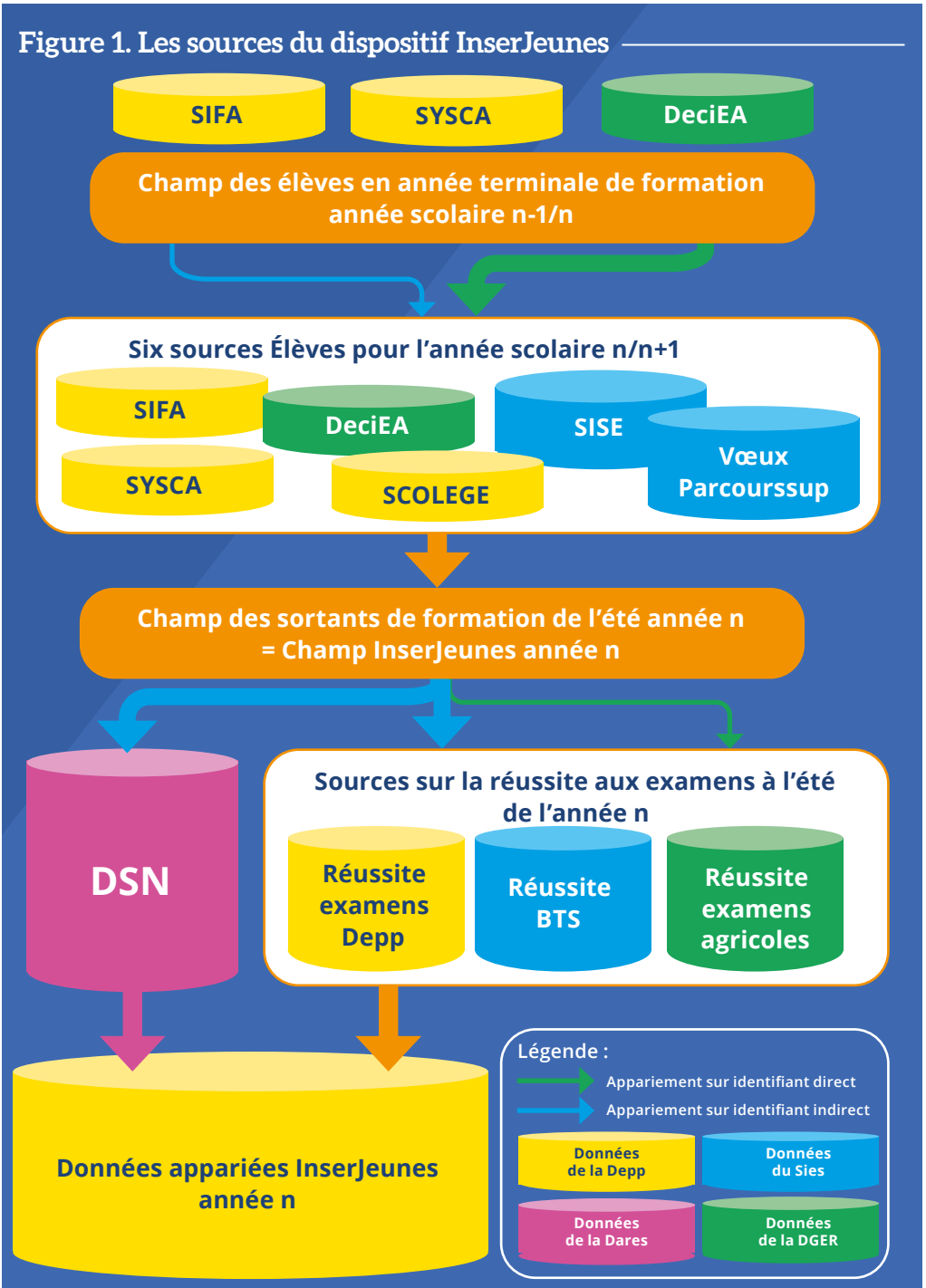
Dans la troisième phase, les bases élèves/apprentis sont enrichies avec leur réussite aux examens (selon les cas cet appariement est réalisé sur l'INE ou sur identifiants indirects), ce qui permet de calculer le **taux d'interruption en cours de formation**.

Enfin, lors de la quatrième phase, les bases élèves/apprentis sortants sont appariées sur identifiants indirects avec la DSN[11], ce qui permet de mesurer un **taux d'emploi salarié en France des sortants** puis la **valeur ajoutée de l'établissement sur ce taux d'emploi**[12]. La DSN contient des informations détaillées sur les contrats salariés (type de contrat, salaire, quotité de travail, catégorie socioprofessionnelle, etc.) ainsi que sur l'établissement employeur (secteur, commune d'implantation, etc.) : de ce fait, InserJeunes pourra permettre également de réaliser des études statistiques, par exemple, sur l'adéquation formation/emploi.

Dans InserJeunes, le taux d'appariement ne donne aucune indication du niveau de qualité du processus. Par exemple, lorsqu'un sortant n'est pas apparié avec la DSN, il n'est pas possible de savoir si c'est parce qu'il n'est pas en emploi salarié ou en raison d'une erreur dans le processus d'appariement. Mais le dispositif comporte un appariement sur identifiants indirects appelé « appariement qualité », annuel, pour lequel le taux théorique est de 100 % : il s'agit du rapprochement du fichier recensant les apprentis au 31 décembre ayant un contrat d'apprentissage actif[13] avec la DSN. Ainsi, le taux d'appariement réel obtenu constitue un indicateur de la qualité du processus d'appariement.

7. SIFA (Système d'information de la formation des apprentis) pour les apprentis, SYSCA (Système d'information statistique consolidé académique) pour les élèves de voie professionnelle scolaire du ministère de l'Éducation nationale et DeciEA pour les élèves de voie professionnelle scolaire du ministère de l'Agriculture.
8. SIFA, SYSCA, DeciEA plus les élèves du secteur privé hors contrat avec la source SCOLEGE et le supérieur *via* les enquêtes SISE (Système d'information sur le suivi des étudiants) et les vœux validés *via* Parcoursup dans un institut de formation en soins infirmiers.
9. En particulier en prenant en compte les poursuites dans l'enseignement supérieur.
10. En réalité quelques sortants de formation peuvent en fait être encore en étude : par exemple on ne repère pas les poursuites d'études à l'étranger.
11. Pour être tout à fait précis, on utilise la source DMMO (Déclaration de mouvement de main d'œuvre) basée sur la DSN (Déclaration sociale nominative).
12. La notion de valeur ajoutée est un concept largement développé dans un précédent article, voir (Evain, 2020).
13. En dehors de la fonction publique, car les emplois publics ne sont pas encore intégrés en DSN.

Figure 1. Les sources du dispositif InserJeunes



SIFA : système d'information de la formation des apprentis
SYSCA : système d'information statistique consolidé académique
DeciEA pour les élèves de voie professionnelle scolaire du ministère de l'Agriculture
SCOLEGE : scolarité léger, application web développée par la Depp
SISE : système d'information sur le suivi des étudiants

> *Le processus statistique InserJeunes comporte au total dix appariements sur identifiants indirects pour chaque année scolaire.*

Le processus statistique InserJeunes comporte au total dix appariements sur identifiants indirects pour chaque année scolaire. La problématique des appariements sur identifiants indirects est donc centrale. Cela implique de mettre au point un processus d'appariement général, puis d'en faire une implémentation informatique générique et rapide.

AU CŒUR DU SYSTÈME D'INFORMATION : UN PROCESSUS D'APPARIEMENT EN CINQ ÉTAPES

Dans InserJeunes, chaque appariement sur identifiants indirects est réalisé entre deux tables individuelles[14] sans double compte. Le processus d'appariement retenu pour InserJeunes ***(figure 2)*** comporte cinq étapes successives, comme c'est aussi le cas dans la présentation de Peter Christen ***(figure 3)*** (Christen, 2012).

Les données sont tout d'abord normalisées. Puis vient l'étape d'indexation, qui consiste à établir une liste de taille raisonnable de paires « potentiellement intéressantes ». Une paire correspond au croisement d'une ligne de la première table avec une ligne de la seconde table. Chaque paire comporte donc un/des noms, un/des prénoms, une date de naissance, un lieu de naissance et une variable sexe provenant de chacune des deux tables qu'on apparie. En troisième lieu, une similarité est calculée pour chacun des cinq couples d'identifiants indirects de chaque paire (par exemple couple de noms, couple de dates de naissance). Quatrièmement, chaque paire est classifiée : les paires supposées relever du même individu (i.e. lorsque les cinq similarités calculées à l'étape précédente sont suffisamment élevées) sont acceptées et les autres sont rejetées. Enfin, la qualité du processus d'appariement est évaluée.

NORMALISER DES DONNÉES

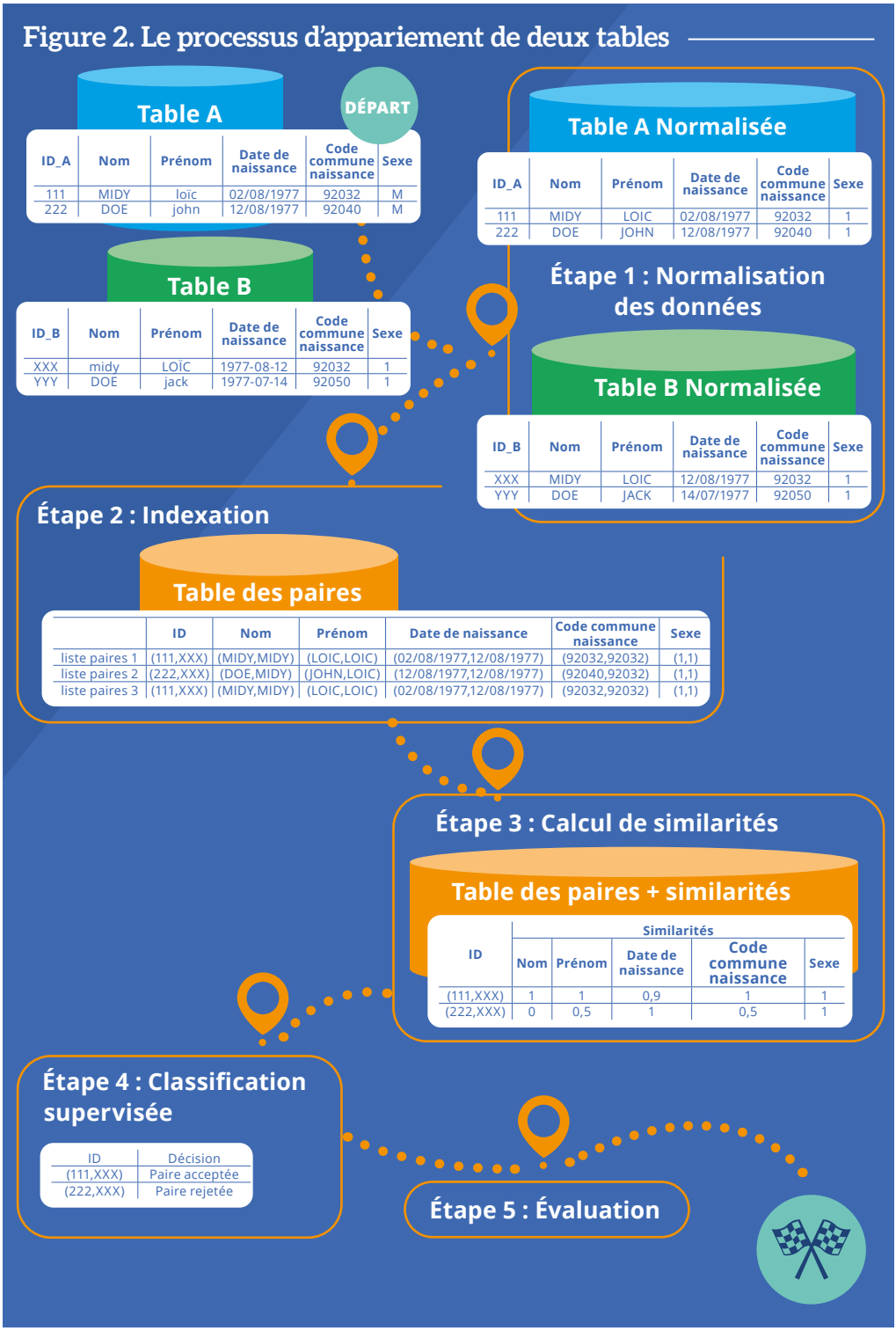
Les identifiants indirects utilisés dans l'appariement se présentent sous des formats hétérogènes dans les différentes sources mobilisées dans InserJeunes. La normalisation des données, première étape du processus d'appariement, consiste à les recoder selon une structure commune afin de faciliter les traitements ultérieurs.

Pour les noms et les prénoms, les traitements principaux suivants sont réalisés :

- mettre en majuscule[15] ;
- éliminer les caractères spéciaux ;
- remplacer les lettres spéciales par leur valeur canonique[16] ;
- éliminer les noms/prénoms d'une lettre et certains noms/prénoms de deux ou trois lettres afin de supprimer les termes peu informatifs comme DE et LE.

14. C'est-à-dire que l'unité d'observation est l'individu (ici élève ou apprenti).
15. Ce qui supprime au passage les accents.
16. Par exemple : Ç devient C, Ï devient I.

Figure 2. Le processus d'appariement de deux tables



Par ailleurs, jour, mois et année des dates de naissance sont stockés dans des variables différentes afin de pouvoir mener des calculs sur les dates de naissance même lorsqu'elles ne sont que partiellement renseignées.

Les sources mobilisées dans InserJeunes sont de bonne qualité. En effet, il n'y a pas de doublons dans les sources principales, il y a très peu de valeurs manquantes sur les identifiants indirects, et le code COG[17] de la commune de naissance, information plus précise que le libellé de la commune, est généralement fourni. Cela est dû au fait que pour tous les élèves, l'immatriculation au Répertoire national des identifiants élèves, étudiants et apprentis a déjà nécessité que l'ensemble des variables identifiantes soient fournies. De même, chaque salarié dans la source DSN fait l'objet d'une procédure de certification du NIR[18] associé, ce qui assure également un niveau de qualité élevé des variables identifiantes.

INDEXER LES DONNÉES : L'APPROCHE NAÏVE

Pour la deuxième étape, celle de l'indexation des données, une première approche naïve consiste à analyser tous les croisements possibles entre les deux tables. Mais le temps de traitement augmente de manière quadratique avec le nombre d'observations des tables à apparier et donc en pratique, cette méthode n'est plus applicable au-delà d'un certain seuil.

Dans le cas de l'appariement qualité, environ 315 000 apprentis sont rapprochés des 7,5 millions de salariés ayant un contrat actif en décembre de l'année considérée. L'analyse exhaustive de chacune des 2,3 billions de paires possible prendrait plusieurs jours voir plusieurs dizaines de jours[19].

> *« La méthode d'indexation retenue doit conjuguer deux objectifs en apparence contradictoires : élaborer une liste de paires la plus petite possible tout en veillant à obtenir le plus possible de paires relatives au même individu. »*

Par ailleurs, analyser l'ensemble des paires ne présente aucun intérêt. En effet, lorsque les noms ou les prénoms ou les dates de naissance sont très différents, il est extrêmement peu probable que la paire soit acceptée.

L'étape d'indexation a pour objectif d'établir une liste de paires « potentiellement intéressantes » de taille raisonnable. La méthode d'indexation retenue doit conjuguer deux objectifs en apparence contradictoires : élaborer une liste de paires la plus petite possible tout en veillant à obtenir dans cette liste le plus possible de paires relatives au même individu.

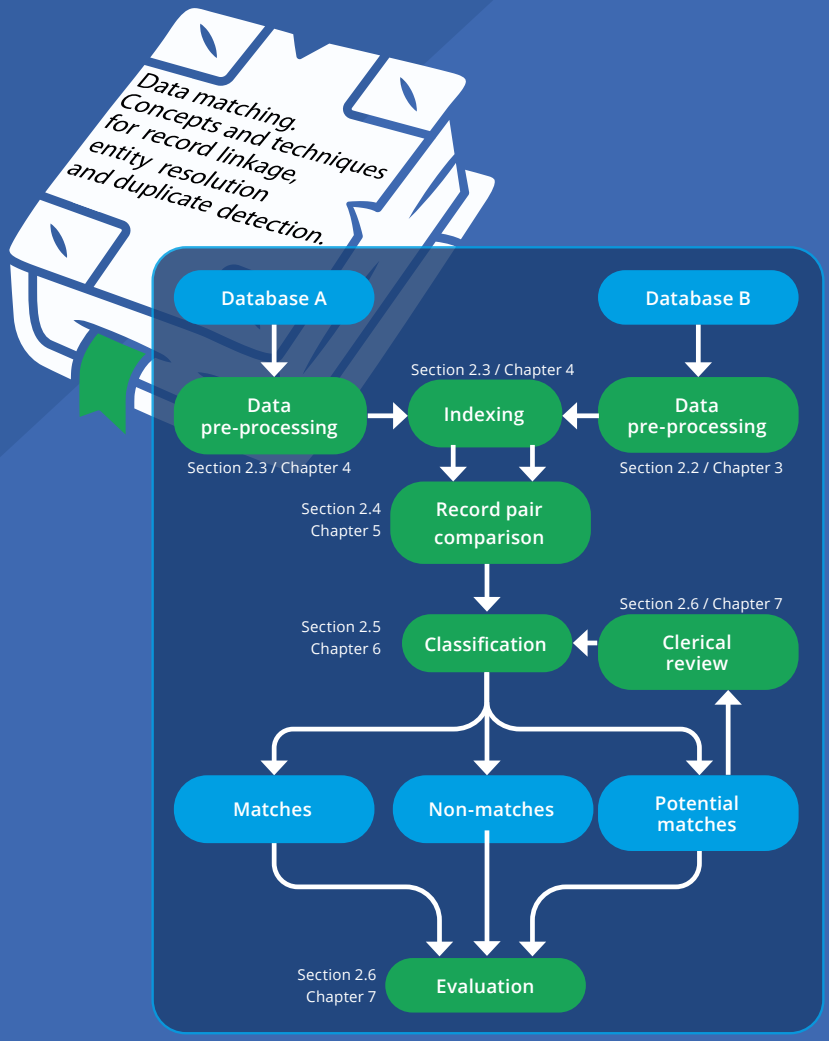
17. Le Code officiel géographique (COG) identifie chaque commune de France.
18. Numéro d'inscription au répertoire national d'identification des personnes physiques (RNIPP).
19. Si chaque ligne était analysée très rapidement, mettons en un cent millième de seconde, le temps de traitement total serait de 273 jours. Même en parallélisant les calculs sur un certain nombre de cœurs, cela resterait trop long.

INDEXER LES DONNÉES : L'APPROCHE CLASSIQUE PAR CLÉ DE BLOCAGE

La méthode la plus souvent utilisée pour indexer les données consiste à ne conserver que les paires qui partagent la même modalité d'une ou de plusieurs variables indirectement identifiantes, qu'on appelle les **clés de blocage** (Christen, 2012 ; Jabot et Treyens, 2018). Par exemple si la clé de blocage est le code de la commune de naissance, cela veut dire que seules les paires partageant ce code seront retenues.

Cette méthode n'a pas été choisie pour InserJeunes, car elle présente plusieurs inconvénients. Tout d'abord, elle produit une liste de paires « potentiellement intéressantes » qui reste encore trop importante. En outre, elle conduit à écarter à tort certaines paires. Par exemple,

Figure 3. Les cinq étapes définies par Christen



si la clé de blocage est le code de la commune de naissance et si cette variable est mal renseignée pour un individu, alors il ne sera jamais apparié. Enfin, il n'est pas possible d'appliquer de clé de blocage sur les noms ou les prénoms car à la moindre faute de frappe ou d'orthographe, les modalités de la clé de blocage seront différentes. Pour résoudre ce problème, on peut certes remplacer les noms et prénoms par leur version phonétisée. Il existe pour ce faire de nombreux algorithmes phonétiques ***(encadré 1)***. Prenons l'exemple d'une paire avec les prénoms « christina » et « kristina ». Ces deux prénoms auront la même version phonétisée avec *Phonex*, (soit *c623)*. *A contrario*, « peter » et « pedro » ont la même version phonétisée avec *Soundex* (soit *p360*), ce qui veut dire qu'on conservera, selon les algorithmes, une liste trop importante de paires « potentiellement intéressantes ». De plus, ces algorithmes phonétiques ont été initialement développés pour la langue anglaise et ils n'ont pas tous été adaptés pour la langue française.

INDEXER LES DONNÉES : L'APPROCHE RETENUE DANS INSERJEUNES

Compte tenu des inconvénients de l'approche classique par clé de blocage, une méthode d'indexation spécifique a été développée pour InserJeunes.

Dans un premier temps, un appariement exact entre les deux tables est réalisé sur l'ensemble des champs suivants : le premier nom, le premier prénom, le jour, le mois et l'année de naissance, le code de la commune de naissance et le sexe. Dans le cadre de l'appariement qualité, cette étape permet d'apparier environ 84 % des apprentis avec la source DSN. Une fois cette étape franchie, il ne reste à apparier qu'environ 50 000 apprentis avec 7,2 millions de salariés ayant un contrat actif en décembre. Le volume de travail est ainsi déjà divisé par un facteur six[20].

Dans un second temps, l'union (sans doublons) des trois listes de paires suivantes est établie :

- les paires qui ont une distance faible entre les premiers noms, une distance faible entre les premiers prénoms, même département de naissance et même année de naissance ;
- les paires qui ont même date de naissance et même département de naissance ;
- les paires qui ont même premier nom et même premier prénom.

InserJeunes utilise la **distance de Levenshtein** pour les noms et les prénoms. Celle-ci correspond au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'un nom/prénom à l'autre. L'union des différentes requêtes permet de bien couvrir tous les cas rencontrés fréquemment et ainsi de conserver presque toutes les paires « potentiellement intéressantes ». De plus, comme chaque requête est relativement précise, le nombre de paires « potentiellement intéressantes » retenues n'est pas trop élevé. Pour l'appariement qualité, cette méthode d'indexation conduit à retenir 1 million de paires soit un nombre raisonnable qui pourra être traité suffisamment rapidement lors des étapes ultérieures du processus.

Cette méthode d'indexation fonctionne très bien sur la volumétrie d'InserJeunes mais ne serait peut-être pas adaptée dans le cas de l'appariement de tables de plusieurs dizaines de millions de lignes.

20. Soit 315 000/50 000.

CALCULER LES SIMILARITÉS

La troisième étape consiste à enrichir chacune des paires déterminées lors de l'indexation, de cinq variables « similarités » sur nom, prénom, date et commune de naissance et sexe. Chaque similarité est une mesure du degré de ressemblance des identifiants indirects considérés. Trois natures de similarités différentes sont utilisées dans InserJeunes selon la nature des variables mobilisées comme identifiants indirects.

> *« Chaque similarité est une mesure du degré de ressemblance des identifiants indirects considérés. »*

Tout d'abord, la **similarité de Jaro-Winkler** est mise en œuvre pour les noms et prénoms. Cette dernière est une adaptation de la similarité de Jaro développée par le statisticien Winkler qui ajoute une « bonification » lorsque les deux chaînes que l'on compare débutent par un préfixe commun. L'algorithme de calcul de la similarité de Jaro entre deux chaînes de caractères est le suivant :

- on commence par calculer un *facteur éloignement* qui est égal à la longueur de la chaîne la plus longue divisée par 2 moins 1 (exemple : si on compare DWAYNE et DUANE, l'éloignement est de 2) ;
- puis, on établit la *liste des caractères correspondants* c'est-à-dire les caractères qu'on retrouve dans les deux chaînes avec un éloignement inférieur ou égal à la valeur calculée précédemment (si on compare **D**W**A**Y**NE** et **D**U**ANE**, les caractères correspondants sont D, A, N et E) ;
- ensuite il faut calculer le *nombre de transpositions entre les caractères correspondants*, c'est-à-dire le nombre de fois (divisé par deux) où le $i^{ème}$ caractère correspondant de la première chaîne est différent du $i^{ème}$ caractère correspondant de la seconde chaîne (dans l'exemple qui précède, le nombre de transpositions est de 0) ;
- enfin, on calcule **la similarité de Jaro**, somme pondérée des trois termes suivants :
  - nombre de caractères correspondants divisé par la longueur de la première chaîne de caractères (soit 4/6 dans notre exemple) ;
  - nombre de caractères correspondants divisé par la longueur de la seconde chaîne de caractères (soit 4/5 dans notre exemple) ;
  - nombre de caractères correspondants moins nombre de transpositions, divisé par le nombre de caractères correspondants (soit 1 dans notre exemple). La similarité de Jaro vaut donc ici 1/3 x 4/6 + 1/3 x 4/5 + 1/3 x 1 = 0,822.

Il existe de nombreuses autres mesures de la similarité entre noms et prénoms. Par exemple, certaines sont basées sur la comparaison de *bigrammes ou trigrammes*[21] entre chaînes de caractères, comme la similarité de Jaccard. Cependant il semble qu'il n'en existe pas qui donne des résultats nettement meilleurs que la similarité de Jaro-Winkler, sur les noms et prénoms (Christen, 2006).

Ensuite, une similarité spécifique à InserJeunes a été élaborée pour les dates de naissances. Par exemple, lorsque deux dates de naissance ne diffèrent que sur le jour de naissance, la similarité est de 0,9 si la différence ne porte que sur un des deux chiffres et de 0,8 sinon.

21. Par exemple les bigrammes de DWAYNE sont DW, WA, AY, YN et NE et les trigrammes de DUANE sont DUA, UAN et ANE.

Si les deux dates de naissance ont la même année et que le jour d'une date correspond au mois de l'autre et réciproquement (exemple : **01**02**02**2005 et **02**01**01**2005) alors la similarité est de 0,65.

Enfin, pour la variable sexe, une similarité binaire est utilisée. Pour la commune de naissance, la similarité est de 1 lorsque les codes COG sont identiques. S'ils sont différents, la similarité est de 0,5 si le code département est identique et de 0 sinon.

CLASSIFIER LES PAIRES: UNE AFFAIRE DE *MACHINE LEARNING?*

La quatrième étape consiste à statuer sur chaque paire, en utilisant les similarités calculées à l'étape précédente. Lorsque les similarités sont élevées, c'est-à-dire proches de 1, la paire est acceptée. Dans le cas contraire, la paire est rejetée.

Encadré 1. Introduction aux algorithmes phonétiques

Un algorithme phonétique est un algorithme conçu pour indexer les mots selon leur prononciation. Par exemple l'algorithme **Soundex** procède comme suit :

1. Conserver la première lettre de la chaîne.

2. Supprimer toutes les occurrences des lettres : a, e, h, i, o, u, w, y (à moins que ce ne soit la première lettre du nom).

3. Attribuer une valeur numérique aux lettres restantes de la manière suivante (dans la version pour les noms en anglais) :

- 1 = B, F, P, V
- 2 = C, G, J, K, Q, S, X, Z
- 3 = D, T
- 4 = L
- 5 = M, N
- 6 = R

4. Si deux lettres (ou plus) avec le même nombre sont adjacentes dans le nom d'origine, ou s'il n'y a qu'un h ou un w entre elles, alors on ne retient que la première de ces lettres.

5. Renvoyer les 4 premiers éléments. S'il y a moins de 4 éléments compléter par des zéros.

Voici des exemples d'application d'algorithmes phonétiques :

Nom de départ	Nom phonétisé			
	Soundex	Phonex	NYSIIS	Double Metaphone
christina	c623	c623	chra	krst
kristina	k623	c623	cras	krst
peter	p360	b360	pata	ptr
pedro	p360	b360	padr	ptr

(Sources : https://fr.wikipedia.org/wiki/Algorithme_phon%C3%A9tique et https://fr.wikipedia.org/wiki/Soundex).

Une première approche simple consiste à calculer une similarité globale pour chaque paire, fonction strictement croissante des similarités des différents champs. Les paires dont la similarité globale est supérieure à un certain seuil sont acceptées, les autres étant rejetées. La fonction et le seuil retenus sont choisis de manière empirique *via* l'analyse d'un échantillon de paires dont le statut (accepté ou rejeté) a été annoté manuellement. Cette méthode présente l'avantage de la simplicité mais le choix de la fonction et du seuil demeurent arbitraires, donc rien ne garantit que ces choix soient optimaux.

Compte tenu des limites de la première approche, des classifications mobilisant des **algorithmes de *machine learning* supervisés** ont été testées (forêts aléatoires et machines à vecteurs de support ou séparateurs à vaste marge (SVM) ***(encadré 3)***). La démarche consiste à entraîner l'algorithme sur un échantillon de paires dont le statut a été renseigné manuellement, puis à l'appliquer sur les autres paires. Les paramètres optimaux de chaque algorithme sont déterminés par validation croisée en maximisant la **métrique *f-measure***.

Dans le cas de l'appariement qualité, l'approche simple, les forêts aléatoires et les SVM ont tous donné des résultats similaires et excellents donc, au final, la classification artisanale a été retenue pour InserJeunes.

Comment expliquer ce résultat, *a priori* surprenant ? Une façon de représenter notre problème est de considérer que chaque paire est un point dans un espace à 5 dimensions, celles des 5 similarités (nom, prénom, date de naissance, commune de naissance et sexe). La variable sexe étant très peu discriminante, elle pourrait être éliminée de l'analyse ce qui restreindrait l'espace à 4 dimensions. Dans chaque dimension, les similarités prennent des valeurs entre 0 et 1.

Le problème consiste donc à trouver une frontière de séparation entre les points/paires acceptés et les points/paires rejetés dans un espace $[0 ; 1]^4$, soit un espace de toute petite taille. De plus, la zone « proche » du point (1,1,1,1) correspond à la zone dans laquelle se trouvent presque toutes les paires qu'on doit accepter. Ainsi, les points correspondant aux paires qu'il faut accepter ne sont pas trop « mélangés » avec les points correspondant aux paires qu'il faut rejeter. Il est donc relativement facile de résoudre ce type de problème, ce qui explique que toutes les méthodes testées donnent de manière équivalente de très bons résultats.

La classification probabiliste, développée originellement par (Fellegi et Sunter, 1969) est une autre méthode développée spécifiquement dans le cadre des appariements sur identifiants indirects et fréquemment citée dans la littérature. Cette méthode n'a pas été investiguée par l'équipe InserJeunes, car il n'en existe pas, à notre connaissance, d'implémentation rapide dès que le volume de données est assez important.

ÉVALUER POUR VALIDER LES CHOIX OPÉRÉS

Étant donné le caractère central des appariements sur identifiants indirects dans InserJeunes, il convenait d'évaluer leur qualité. C'est donc la cinquième et dernière étape du processus.

L'évaluation nécessite de disposer d'un échantillon de paires dont le statut (accepté ou rejeté) a été annoté manuellement et qui n'a pas été utilisé lors de l'étape de classification. Sur cet échantillon, la prédiction issue de la classification supervisée est comparée avec le véritable statut de la paire, c'est-à-dire celui établi manuellement, ce qui permet d'obtenir dans un premier temps quatre grandeurs :

- les vrais positifs (VP) ;
- les faux positifs (FP) ;
- les vrais négatifs (VN) ;
- et les faux négatifs (FN).

Par exemple, une paire faux négatif est une paire rejetée par l'algorithme de classification mais acceptée par l'humain qui a réalisé l'annotation. À partir de ces quatre grandeurs il est possible d'établir plusieurs mesures de la qualité globale.

La mesure la plus connue est l'*accuracy*, soit (VP+VN)/ nombre total de paires. Mais comme toute mesure qui utilise les vrais négatifs, elle n'est pas adaptée. Pourquoi ? Parce que les données sont déséquilibrées : il y a beaucoup de paires dont le vrai statut est rejeté et peu de paires dont le vrai statut est accepté. Dans le cas de l'appariement qualité, environ 40 000 paires sont acceptées sur 1 million de paires donc au minimum 950 000 paires ont pour véritable statut « rejeté ». Un classifieur naïf qui rejette 100 % des paires a donc une *accuracy* d'au moins (0+950 000)/ (1 000 000) soit 95 %.

Encadré 2. Les démarches juridiques

Les appariements sur identifiants indirects se font souvent sur des données à caractère personnel (DCP). Or leur usage est encadré depuis 2018 par le règlement général sur la protection des données (RGPD)*.

Le dispositif InserJeunes a donc fait l'objet d'une déclaration au registre des traitements suivi par le délégué à la protection des données du ministère de l'Éducation nationale, conformément à l'article 30 du règlement. Une analyse d'impact relative à la protection des données (AIPD) a également été réalisée. La réalisation d'une AIPD est obligatoire d'une part pour certains types de traitements (cette liste ayant été établie par la Cnil**) et d'autre part lorsqu'au moins deux critères parmi une liste de neuf s'appliquent au traitement.

InserJeunes remplit les trois critères suivants :

- collecte de données personnelles à large échelle ;
- croisement de données ;
- et personnes vulnérables (patients, personnes âgées, enfants, etc.).

Schématiquement, une AIPD comporte trois parties :

- une description du traitement mis en œuvre ;
- l'évaluation de la nécessité et de la proportionnalité de collecte de DCP ;
- une analyse des risques de sécurité ainsi que leur impact potentiel sur la vie privée ;

Le RGPD impose de limiter au strict nécessaire, compte tenu des finalités du traitement, la collecte et la conservation de DCP.

L'équipe InserJeunes a également fait auprès du Cnis des demandes d'accès aux sources de la Direction générale de l'enseignement et de la recherche*** et de la Dares utilisées dans le dispositif, au titre de l'article 7bis de la loi de 1951*.

** Voir références juridiques en fin d'article.*
*** Voir https://www.cnil.fr/sites/default/files/atoms/files/liste-traitements-aipd-requise.pdf.*
**** Direction du ministère de l'Agriculture.*

Trois autres mesures ont donc été retenues dans le dispositif InserJeunes :

- la **précision**, qui vaut VP/(VP+FP). Par exemple, si la précision est de 80 % alors cela veut dire que 80 % des paires acceptées le sont à bon escient ;
- le **rappel**, qui vaut VP/(VP+FN) : si le rappel est de 90 % alors cela veut dire que 90 % des vraies paires ont été détectées par l'algorithme de classification ;
- et la ***f-measure*** qui est la moyenne harmonique de la précision et du rappel : elle vaut donc 2 x (précision x rappel)/(précision+ rappel).

Dans le cas de l'appariement qualité, la précision vaut 95 % et le rappel 99 %. Au total, 97 % des apprentis sont appariés dans l'appariement qualité (84 % *via* l'appariement direct et 13 % *via* l'appariement approché) soit un taux d'appariement proche du taux théorique de 100 %.

LA MISE EN ŒUVRE INFORMATIQUE : LE CHOIX D'UN OUTIL SPÉCIFIQUE

Plusieurs logiciels d'appariement ont été testés dans le cadre du projet InserJeunes : *FEBRL* de Peter Christen, *matchID* développé au ministère de l'Intérieur[22] et deux librairies R. En R, la meilleure implémentation semble être la librairie R *fastlink*, mais d'après la documentation[23] elle met environ 8 heures pour traiter l'appariement de tables de 300 000 lignes et elle s'appuie sur la classification probabiliste de Fellegi et Sunter. Seul l'outil *matchID* répondait aux besoins mais sa mise en œuvre s'est avérée relativement complexe.

L'équipe InserJeunes a également étudié la documentation concernant d'autres logiciels. L'institut national de statistiques italien Istat a développé un outil nommé RELAIS[24] qui implémente notamment la classification probabiliste ; mais à partir de 100 000 observations, le temps de traitement est d'environ 1h15 (Eurostat, 2009). Aux États-Unis, le bureau du Census a de son côté conçu l'outil *bigMatch* spécifiquement pour traiter de gros volumes, mais il semble ne réaliser que la phase d'indexation des données et il est écrit en C, ce qui complique son intégration avec des outils de *machine learning*[25].

Suite à ce travail de comparaison, il a été décidé de développer un outil spécifique qui réponde à quatre grands besoins d'InserJeunes :

- les appariements doivent être **rapides** : l'outil InserJeunes réalise l'appariement qualité en 15 minutes ;
- l'outil d'appariement doit être **générique** c'est-à-dire facilement adaptable pour tous les cas d'appariements sur identifiants indirects. Pour ce faire, la spécification de chaque appariement (les champs comparés, la méthode de similarité choisie pour chaque champ, etc.) est décrite dans le langage de balisage XML qui est ensuite interprété par l'outil. Cela implique que le statisticien ou le *data scientist* qui produit le XML respecte une grammaire formelle : il doit décrire son appariement en respectant un formalisme et un vocabulaire spécifiques à l'outil[26] mais qui est très fortement inspiré par celui présenté dans l'ouvrage de Peter Christen (Christen, 2012). Cette façon de procéder permet également d'assurer une traçabilité complète de chaque appariement, les spécifications XML étant toutes sauvegardées ;

22. Dans le cadre d'un projet du programme Entrepreneurs d'Intérêt Général.
23. Voir (Enamorado, Fifield et Imai, 2019) page 362 figure 3 *Running Time Comparison*.
24. Pour plus de détail, voir (Istat, 2020).
25. L'équipe projet n'a pas connaissance de librairie de *machine learning* écrite en C, et faire cohabiter des briques écrites dans des langages différents demande un investissement plus conséquent.
26. Un langage spécifique a ainsi été créé à cette occasion pour le domaine appariement.

Encadré 3. Brève introduction aux classifications supervisées en *machine learning*

Les algorithmes de classification supervisée de *machine learning* sont, dans un premier temps, entraînés sur des données étiquetées c'est-à-dire pour lesquelles la variable à prédire est connue (dans le cas d'InserJeunes une variable qualitative binaire). On retient le paramétrage général de l'algorithme qui maximise une grandeur statistique à déterminer et qui dépend du problème traité.

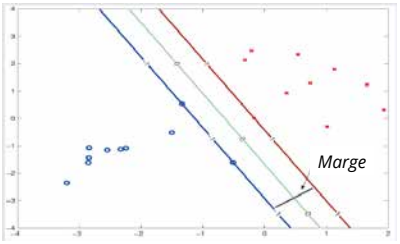


Figure A

En l'occurrence, c'est la métrique *f-measure* qui est maximisée. Ensuite, le modèle précédemment entraîné est appliqué sur de nouvelles données pour lesquelles la variable à prédire est inconnue. L'enjeu méthodologique principal consiste à s'assurer que l'algorithme a « bien appris » pendant l'entraînement afin qu'il fasse ensuite des prédictions correctes sur les nouvelles données.

Le processus de classification supervisée est illustré avec l'algorithme *support vector machine* (SVM) appliqué sur un exemple simplifié, dans lequel il n'y a que deux dimensions ***(figure A)***.

La première approche consiste à choisir la frontière (les lignes bleues et rouge) dont la marge est la plus large possible, c'est-à-dire qu'on veut que tous les points d'une couleur soit d'un côté de la frontière et les points de l'autre couleur soient de l'autre côté et on veut également maximiser la taille du *no man's land* entre les deux lignes c'est-à-dire la zone dans laquelle il n'y a aucun point. Cette approche s'appelle aussi « séparateur à vaste marge ».

Mais pour certains jeux de données ce type de frontière n'existe pas. De plus, si on impose à l'algorithme de séparer 100 % des points, alors ce dernier « collera » trop aux données d'apprentissage. Il risque alors de « sur-apprendre » et de mal généraliser sur de nouvelles données ***(figure B).***

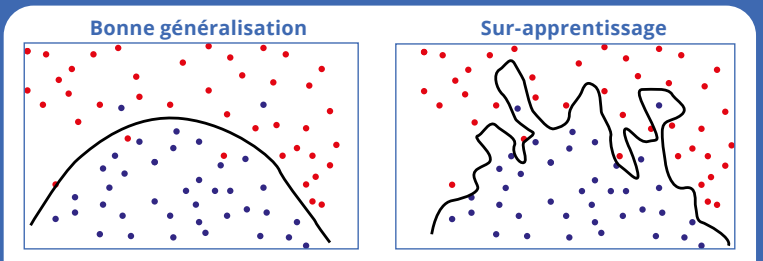


Figure B

Pour éviter cet écueil il faut accepter que le SVM ne classe pas correctement un petit pourcentage des paires. Il faut également évaluer la qualité de l'algorithme sur des données étiquetées qui n'ont pas été utilisées lors de la phase d'apprentissage, ce qui permet de vérifier qu'il n'y a pas eu de sur-apprentissage.

- étant donné que l'évaluation du processus d'appariement est fondamentale et que cela nécessite de disposer d'un échantillon de paires annotées manuellement, une interface d'annotation de paires **ergonomique** a été développée ;
- l'outil s'appuie sur plusieurs librairies *open source* ce qui a permis d'accélérer son développement (le cœur de l'outil a été développé en deux semaines[27]) et d'en faciliter la maintenance.

L'outil d'appariement d'InserJeunes sera mis à disposition en *open source* à l'été 2021. Cependant, pour des projets d'appariement sur des tables nettement plus volumineuses, l'outil *matchID* semble plus adapté notamment parce qu'il réalise l'indexation *via* des requêtes *elastic search* et non pas *via* des requêtes SQL.

PARTAGER L'EXPÉRIENCE ACQUISE AU COURS DU PROJET

À la lumière de l'expérience acquise au cours du projet InserJeunes, quels enseignements tirer ?

En premier lieu, il est manifeste que la qualité globale du processus d'appariement dépend très fortement de la qualité des variables identifiantes indirectes. Le contexte d'InserJeunes était favorable, car les variables sont bien renseignées dans les bases mobilisées. Le respect de chacune des étapes se révèle effectivement indispensable à la réussite de l'opération : bien normaliser les données afin de faciliter les traitements ultérieurs, consacrer du temps à déterminer la meilleure façon de calculer les similarités (ce travail s'appelle le « *feature engineering* » en *machine learning*) afin d'augmenter sensiblement la qualité de la classification, aucune de ces activités n'est superflue. L'évaluation, quand elle s'appuie sur un échantillon de paires annotées manuellement n'ayant pas été utilisées dans l'étape de classification, permet de garantir la qualité globale du processus et en particulier de vérifier qu'il n'y a pas eu de sur-apprentissage. Pour InserJeunes, pouvoir réaliser un « appariement qualité » annuel est une chance, mais tous les systèmes d'informations ne s'y prêteront pas.

Du point de vue informatique le fait de s'appuyer sur plusieurs librairies *open source* a permis de réaliser les développements dans des délais courts. Le choix de décrire chaque spécification d'appariement en XML, respectant un langage spécifique d'appariement a permis de spécifier puis d'intégrer rapidement dans la chaîne de production tous les appariements nécessaires à InserJeunes. Globalement, ces choix ont permis d'achever la mise en œuvre du cœur d'InserJeunes fin 2020, puis de diffuser début février 2021 les premiers résultats sur les jeunes sortants de voie professionnelle scolaire et par apprentissage à l'été 2018 et 2019[28].

27. L'indexation est réalisée en langage SQL sur une base de données PostgreSQL, en mobilisant le module fuzzystrmatch, le calcul des similarités de Jaro-Winkler et de Levenshtein mobilise la librairie Python jellyfish et les algorithmes de *machine learning* sont réalisés avec la librairie Python scikit-learn.
28. Voir (Collin et Marchal, 2021a ; 2021b ; 2021c).

BIBLIOGRAPHIE

CHRISTEN, Peter, 2006. *A Comparison of Personal Name Matching: Techniques and Practical Issues.* [en ligne]. Septembre 2006. The Australian National University Research Publications. Joint Computer Science Technical Report Series, TR-CS-06-02. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://openresearch-repository.anu.edu.au/bitstream/1885/44521/3/TR-CS-06-02.pdf*.

CHRISTEN, Peter, 2012. *Data matching. Concepts and techniques for record linkage, entity resolution and duplicate detection.* 4 juillet 2012. Springer. ISBN 978-3-642-31163-5.

COLLIN, Christel et MARCHAL, Nathalie, 2021a. *Six mois après leur sortie en 2019 du système éducatif, 41 % des lycéens professionnels sont en emploi salarié.* [en ligne]. Février 2021. DEPP-MENJS. Note d'information n°21.06. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://education.gouv.fr/six-mois-apres-leur-sortie-en-2019-du-systeme-educatif-41-des-lyceens-professionnels-sont-en-emploi-309320*.

COLLIN, Christel et MARCHAL, Nathalie, 2021b. *Six mois après leur sortie en 2019 du système éducatif, 62 % des apprentis de niveau CAP à BTS sont en emploi salarié.* [en ligne]. Février 2021. DEPP-MENJS. Note d'information n°21.07. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://education.gouv.fr/six-mois-apres-leur-sortie-en-2019-du-systeme-educatif-62-des-apprentis-de-niveau-cap-bts-sont-en-309329*.

COLLIN, Christel et MARCHAL, Nathalie, 2021c. *Des lycéens professionnels et des apprentis mieux insérés 12 mois après leur sortie d'études en juillet 2020 que 6 mois après, malgré la crise.* [en ligne]. Mai 2021. DEPP-MENJS. Note d'information n°21.24. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://education.gouv.fr/des-lyceens-professionnels-et-des-apprentis-mieux-inseres-12-mois-apres-leur-sortie-d-etudes-en-323294*.

ENAMORADO, Ted, FIFIELD, Benjamin et IMAI, Kosuke, 2019. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. In : *American Political Science Review*. [en ligne]. N° 113, 2, pp. 353-371. [Consulté le 27 mai 2021]. Disponible à l'adresse : *http://imai.fas.harvard.edu/research/files/linkage.pdf*.

EUROSTAT, 2009. *Insights on Data Integration Methodologies*. [en ligne]. ESSnet-ISAD workshop, Vienne, 29-30 mai 2008, page 53. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.istat.it/it/files/2015/04/Insights-on-Data-Integration-Methodologies.pdf.*

EVAIN, Franck, 2020. Indicateurs de valeur ajoutée des lycées. Du pilotage interne à la diffusion grand public. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 74-94. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008703/courstat-5-6.pdf*.

FELLEGI, Ivan P. et SUNTER, Alan B., 1969. A theory for record linkage. In : *Journal of the American Statistical Association.* Décembre 1969. Taylor & Francis Ltd.. Volume 64, n°328, pp. 1183-1210.

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 27 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.

ISTAT, 2020. RELAIS (Record Linkage At Istat). In : *site de Istat.* [en ligne]. 19 novembre 2020. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/relais*.

JABOT, Patrick et TREYENS, Pierre-Eric, 2018. Appariement de l'enquête Care par identification du plus proche écho. In : *site des 13es Journées de méthodologie statistique de l'Insee (JMS).* [en ligne]. 12-14 juin 2018. [Consulté le 27 mai 2021]. Disponible à l'adresse : *http://www.jms-insee.fr/2018/S20_1_ACTEv2_TREYENS_JMS2018.pdf*.

JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor et TIBSHIRANI, Robert, 2013. *An introduction to statistical learning with applications in R*. Springer. ISBN 978-1-4614-7138-7.

KILSS, Beth et ALVEY, Wendy, 1985. *Record Linkage Techniques – 1985.* [en ligne]. 1er décembre 1985. Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9-10, 1985. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://rosap.ntl.bts.gov/view/dot/13855*.

FONDEMENTS JURIDIQUES

Loi n° 2018-771 du 5 septembre 2018 pour la liberté de choisir son avenir professionnel. In : *site de Légifrance*. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000036847202/*.

Loi n° 51-711 du 7 juin 1951 sur l'Obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance*. [en ligne]. Mise à jour du 25 mars 2019. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573/*.

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. In : *site EUR-Lex.* [en ligne]. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32016R0679&from=FR*.

UN SERVICE AU CŒUR DE LA QUALITÉ DES BASES DE DONNÉES

PRÉSENTATION D'UN PROTOTYPE D'ATMS

*Isabelle Boydens**, *Gani Hamiti*** *et Rudy Van Eeckhout****

Les bases de données doivent idéalement pouvoir être adaptées aux évolutions de leur environnement selon leurs usages. La prise en compte de ces changements revêt un impact stratégique pour la qualité des données administratives, et de ce fait pour les systèmes d'information statistiques qui les utilisent. Afin de gérer au mieux les transformations issues du réel observable touchant les données, il existe désormais une approche innovante, opérationnelle et généralisable à tout système de gestion de base de données relationnel.

Grâce aux avancées de la recherche en matière de qualité de données, l'étude des anomalies et de leurs traitements donne le jour à un prototype original, appelé ATMS* (Anomalies & Transactions Management System). *Ce service permet un suivi des anomalies et des traitements, en support à la méthode dite du* back tracking : *dans une approche préventive de la qualité des données, la méthode est destinée à améliorer structurellement la qualité à la source, et son implémentation révèle un retour sur investissement important. Les caractéristiques du prototype d'ATMS sont mises en relation avec le recours aux* data quality tools *en usage dans les approches curatives, offrant de nouvelles perspectives pour les systèmes d'information statistiques.

Databases must ideally evolve over time along with the change of their environment according to their uses. Taking these changes into account has a strategic impact on administrative data quality, and therefore on the statistical information systems using them. In order to help manage the transformations resulting from the observable reality affecting data, this article proposes an innovative, operational approach generalizable to any relational database management system.

Thanks to the advances of research in data quality, the study of anomalies and their management has given rise to an original prototype, called ATMS (Anomalies & Transactions Management System). This service allows the tracking of anomalies and processing, supporting the back tracking method: in a preventive approach of data quality, the method is intended to structurally improve quality at the source, and its implementation provides a significant return on investment. The characteristics of the ATMS prototype are combined with the use of data quality tools in curative approaches, offering new perspectives for statistical information systems.

* Professeur, Université libre de Bruxelles ; *Data Quality Expert*, Smals,
Isabelle.Boydens@ulb.be
** *Data Quality Analyst*, Smals,
Gani.Hamiti@smals.be
*** *Database R&D*, Smals,
Rudy.Van-Eeckhout@smals.be

« L'évolution est inséparable de la structure parce que l'ensemble lui-même est moins un système arrêté que la fixation provisoire d'un mouvement, l'ordre intelligible d'une tendance [...] *».*

Raymond Aron (La philosophie critique de l'histoire, 1969)

« L'HYPOTHÈSE DU MONDE CLOS » AU CŒUR D'UN RÉEL FLUCTUANT

Les statisticiens font de plus en plus appel à des données administratives et transactionnelles, en soi ou en complément d'autres sources (enquêtes, etc.) (Hand, 2018). L'examen de leur qualité est dès lors également stratégique pour la statistique, qu'elle soit publique ou pas.

Une base de données doit idéalement évoluer avec l'interprétation des réalités qu'elle permet d'appréhender. Les réalités normées sont en effet mouvantes. Ainsi, après les attentats des années 2015 et suivantes, les fichiers de police en la matière, tant en France (Chapuis, 2018) qu'en Belgique (Agence Belga, 2018), ont-ils connu de nombreuses anomalies[1] : doublons potentiels, « faux actifs » en dépit d'un non-lieu, effacements anticipés, difficultés d'interprétation, etc. Celles-ci se sont accumulées suite à l'émergence de catégories de menaces inédites, mais aussi l'urgence dans laquelle ces données sensibles ont dû être traitées. De telles évolutions sont constamment à l'œuvre au cœur des bases de données administratives, sources importantes pour le statisticien.

> *« Toute base de données opérationnelle bien conçue repose sur une hypothèse, celle du « monde clos ». »*

Toute base de données opérationnelle bien conçue repose sur une hypothèse, celle du « monde clos » : des domaines de définition spécifient l'ensemble des valeurs admises au sein du modèle ou du schéma de la base de données (les contraintes d'intégrité) ; les « règles métier » peuvent aussi se décliner dans le code applicatif et contribuer ainsi à la définition des données. Dès lors, une valeur non incluse dans le domaine de définition est considérée comme fausse et doit être rejetée de la base.

Or à l'échelle de millions d'enregistrements, de centaines de champs et de flux d'information, les phénomènes émergents sur le terrain ne sont pas immédiatement pris en compte au sein des bases de données. L'information s'y construit progressivement, au fil de l'interprétation humaine et en l'absence de référentiel absolu.

En dehors du domaine de définition de la base répondant à l'hypothèse du monde clos, la « réalité normée » évolue de manière continue et imprévisible, faisant fi de toute règle d'explication causale déterministe. Et quand la base de données est un instrument d'action sur le réel (dans les secteurs administratifs, médicaux, environnementaux, militaires, etc.), ces questions sont fondamentales et affectent la qualité des données.

1. Par anomalie, nous entendons ici une erreur formelle (par exemple : valeur obligatoire non complétée) mais aussi une présomption d'erreur demandant une interprétation humaine (par exemple : présomption de doublons entre enregistrements fortement similaires, émergence d'une nouvelle catégorie d'activité non prise en compte dans les tables de référence, etc.). Une typologie des anomalies est proposée plus loin.

Il est cependant possible de mieux prendre en compte ces phénomènes sur le plan opérationnel, au cœur du système d'information. Et ce, en particulier, à travers l'interprétation des anomalies et de leurs traitements.

La recherche en « qualité de données » s'intéresse à l'univers des anomalies à des fins opérationnelles. Leur analyse s'est concrétisée par la mise en place d'un service appelé **ATMS, *Anomalies & Transactions Management System.*** Ce système permet le suivi dans le temps de l'historique des anomalies et de leurs traitements sur la base d'indicateurs jugés stratégiques et variables selon le contexte d'usage. Il s'agit d'un concept innovant déjà éprouvé dans la pratique, qui s'appuie sur une méthode dite de *back tracking*.

DES ENJEUX IMPORTANTS S'AGISSANT DES BASES DE DONNÉES ADMINISTRATIVES

De nombreux systèmes d'information administratifs d'envergure sont concernés par la qualité des données, par exemple, en France, la DSN (Déclaration Sociale Nominative). Au sein de celle-ci, tout contrôle jugé essentiel est bloquant (Renne, 2018) mais certains contrôles sont « non bloquants » afin de ne pas ralentir le processus de recueil ou de collecte des déclarations. Ces derniers demandent un traitement manuel ultérieur que les gestionnaires de la base s'efforcent de rationaliser et soulève un arbitrage « coût-qualité » (Renne, 2018). L'exemple de la DSN montre également que la qualité de la base requiert que soit portée une attention particulière à la maîtrise des changements réglementaires (Humbert-Bottin, 2018)[2].

En Belgique, la base de données LATG[3] à la fin des années 1990, puis, son héritière modernisée, la DmfA[4] au début des années 2000, furent des « cas d'étude » des travaux de recherche en matière de qualité de données (Boydens, 1999 ; 2018) vu leur ampleur. La DmfA permet en effet actuellement le prélèvement et la redistribution annuels de 65 milliards d'euros de cotisations et prestations sociales à l'échelle de la Belgique. Elle fait l'objet de modifications législatives trimestrielles et constitue depuis 2001 un système d'information d'envergure intégré, doté de caractéristiques proches de la DSN sur le plan fonctionnel. Elle représente un des socles de nombreuses statistiques en matière d'emploi et de salaires en Belgique. Depuis 20 ans, les recherches ont également porté sur la qualité de nombreuses autres bases de données transactionnelles à la source de productions statistiques[5].

La qualité des données se pose aussi dans le cadre de registres administratifs transversaux tels que ceux prévus en Allemagne. Ceux-ci reposent en effet sur de nombreuses interconnexions susceptibles de soulever d'importantes questions sémantiques[6].

2. [N.D.L.R.] Ce point est régulièrement évoqué dans les articles du Courrier des statistiques, voir également l'article de Christian Sureau et Richard Merlen dans ce même numéro.
3. Base de données relatives aux salaires et aux temps de travail (*Loon en ArbeidsTijdsGegevensbank*).
4. Déclaration multifonctionnelle (*Multifunctionele Aangifte*).
5. Tous ces travaux sont effectués dans le respect de la réglementation européenne du RGPD (Règlement général pour la protection des données).
6. Voir (Bens et Schukraft, 2019). Les auteurs citent à titre d'exemple les grandeurs monétaires, comme les revenus et les chiffres d'affaires (p. 15) ou la nécessité d'un identifiant unique pour les personnes et les entreprises, abstraction faite d'un usage donné (fiscalité, Sécurité sociale, commerce, etc.) (pp. 14-15).

Le rythme des adaptations à apporter à un système d'information varie en fonction des objectifs poursuivis : bases de données administratives ou médicales, par exemple, en tant qu'outils d'action sur le réel, d'une part ou systèmes d'information statistique, d'autre part, en tant qu'instruments d'observation ou d'aide à la prise de décision. Pour les premières, le rythme de modification sera idéalement rapide. Pour les seconds, le rythme d'évolution structurelle est beaucoup plus lent voire inexistant (dans le cas de résultats d'enquêtes, par exemple) afin d'assurer une comparaison sur le long terme alors que l'actualité et la qualité des sources qui les alimentent seront importantes.

Dans un tel contexte, le statisticien est souvent confronté à la question du meilleur moment auquel prendre « la photo » afin d'extraire les données issues d'un système d'information administratif, vis-à-vis duquel il vit une forme de perte de maîtrise (Rivière, 2018). Face à cela, les acquis des travaux en matière de qualité de données offrent des perspectives constructives.

LES TRAVAUX ACADÉMIQUES EN MATIÈRE DE QUALITÉ DE DONNÉES

« La qualité d'une base de données désigne son adéquation relative aux usages pour lesquels elle a été conçue, sous contrainte de budget. »

La qualité d'une base de données désigne son adéquation relative aux usages pour lesquels elle a été conçue, sous contrainte de budget. Les recherches dans ce domaine se sont déployées dans les années 1980 (Madnick *et alii*, 2009) avec la nécessité pour les entreprises de disposer d'adresses et de coordonnées adéquates dans leurs fichiers de clients. C'est ainsi que sont apparus les *data quality tools* ***(encadré 1)***, domaine qui s'est très vite développé au niveau international et qui reste très actif (Hamiti, 2019).

En tant qu'approche « curative » ***(figure 1)***, ceux-ci ont pour objet, sur la base de milliers d'algorithmes régulièrement enrichis, de détecter les problèmes de qualité formellement identifiables (présomptions de doubles, etc.) déjà présents dans les bases de données et d'y remédier *a posteriori* de manière semi-automatique. Ces outils permettent également la gestion automatisée des cas problématiques « en ligne », lors de la saisie dans un portail, par exemple.

Cependant, si l'on se contente d'agir en aval, on ne résout pas structurellement la cause de ces problèmes qui vont sans cesse se reproduire. Ceux-ci peuvent en effet puiser leur source dans des défauts de conception, dans l'évolution du réel représenté ou, encore, dans les flux et procédures qui alimentent les bases de données (par exemple, processus inutilement redondants générant systématiquement des doublons). Dès lors, sans action complémentaire en amont, les *data quality tools* sont destinés à être mobilisés *ad infinitum*. Ceux-ci restent néanmoins indispensables, car l'utilisateur n'a pas nécessairement accès aux flux et procédures qui ont produit les données qu'il exploite.

Dès lors, en complément des interventions « curatives », des approches « préventives » sont indispensables afin d'identifier et de résoudre structurellement à la source les causes des anomalies ***(figure 1)***.

L'Université libre de Bruxelles dédie un enseignement spécifique à la qualité des données depuis 2006, présentant les deux types d'approches (Boydens, 2021). Les travaux de recherche se situent pour une part dans l'optique des méthodes préventives (Boydens, 1999 ; 2010 ; 2012 ; 2018 ; Bade, 2011 ; Radio, 2014 ; Dierickx, 2019). Les travaux actuels en matière de *data quality research* sont également axés sur les algorithmes de type *record linkage* (Batini et Scannapieco, 2016). Ces derniers sont mobilisés par les *data quality tools* dans le cadre des opérations de « comparaison et dédoublonnage » ***(encadré 1)***.

Nombreuses sont les études qui adoptent une vision déterministe : elles envisagent l'écart de la base de données au réel en termes d'inexactitude/exactitude formelle (Srivastava *et alii*, 2019). Or, il n'existe aucune projection biunivoque nécessaire entre le réel empirique et sa représentation au sein d'une base de données.

Encadré 1. *Data Quality Tools*

En parallèle de l'approche préventive de la qualité des données, décrite dans cet article et appuyée sur l'ATMS, une approche curative existe. Celle-ci est destinée à l'amélioration semi-automatique de la qualité des données à leur entrée dans le système ou déjà présentes dans une base de données préexistante ou un ATMS. Le plus souvent, l'approche curative va mobiliser des outils gratuits ou commerciaux développés par une tierce partie. Généralement, ces outils couvrent une à trois de ces grandes familles de fonctionnalités :

- **profilage** : analyser qualitativement et quantitativement des données pour en évaluer la qualité et, souvent, débusquer des problèmes inattendus. Exemple : distribution de la longueur des valeurs d'une colonne, inférence de type, vérification ou découverte de dépendances fonctionnelles ;
- **standardisation** : conformer les données à un standard défini avec le maître d'ouvrage ou à un référentiel existant, pouvant être fourni avec l'outil. Exemple : nettoyage et uniformisation de la représentation des numéros de téléphone, correction et enrichissement d'adresses postales ;
- **comparaison et dédoublonnage** : détecter les doublons et incohérences dans les enregistrements au sein d'un jeu de données ou entre plusieurs (issus potentiellement de bases de données distinctes, en vue d'une intégration, par exemple). La comparaison se base sur des colonnes discriminantes et des algorithmes tolérants à l'erreur (mesure de la distance d'édition, comparaison de l'empreinte phonétique, etc.), déterminés avec le maître d'ouvrage qui apporte sa connaissance du métier. Les outils les plus avancés permettent ici de conserver et lier les enregistrements originaux pertinents sans les écraser et d'en construire un qui représente chaque grappe ainsi repérée. Cet enregistrement sera alors le « survivant », utilisé pour dédoublonner les jeux de données si nécessaire.

Typiquement, ces outils interviennent en *batch*, c'est-à-dire en ciblant, en différé, un ou plusieurs jeux de données déjà existants. Certains permettent cependant également d'intervenir plus en amont, en exposant ces fonctionnalités sous la forme d'une API* que l'application peut appeler au cas par cas au moment où les données entrent dans le système. Ce mode d'action permet de standardiser ou de dédoublonner les données avant leur écriture dans la base et même, si besoin, de conditionner cette écriture par la réussite des opérations qui la précèdent. L'outil implémente ainsi effectivement un pare-feu de données complémentaire au système de détection d'anomalies déjà mis en place par l'application.

** Application Programming Interface*

QUAND L'ÉTUDE DES ANOMALIES AMÉLIORE LA QUALITÉ DES DONNÉES

Outre son intérêt évident pour la qualité des données, l'étude des anomalies est importante en raison de leur pourcentage élevé qui affecte structurellement les systèmes d'information : jusqu'à 10 % selon (Boydens, 2012) et selon d'autres sources (Van Der Vlist, 2011). Or, quand les enjeux (sociaux, financiers, médicaux, etc.) le demandent, ces anomalies doivent faire l'objet d'un examen semi-automatique, voire manuel, souvent lent et fastidieux.

D'où viennent les anomalies, quelle en est la typologie et de là, comment les gérer au mieux ? Afin de répondre à ces questions, il convient de revenir préalablement sur la notion de donnée telle que nous l'envisageons ici et qui fut récemment étudiée du point de vue du statisticien (Rivière, 2020).

DONNÉES DÉTERMINISTES ET DONNÉES EMPIRIQUES

Dans le monde des bases de données (Hainaut, 2018), une donnée est un triplet (i, d, v) composé des éléments suivants :

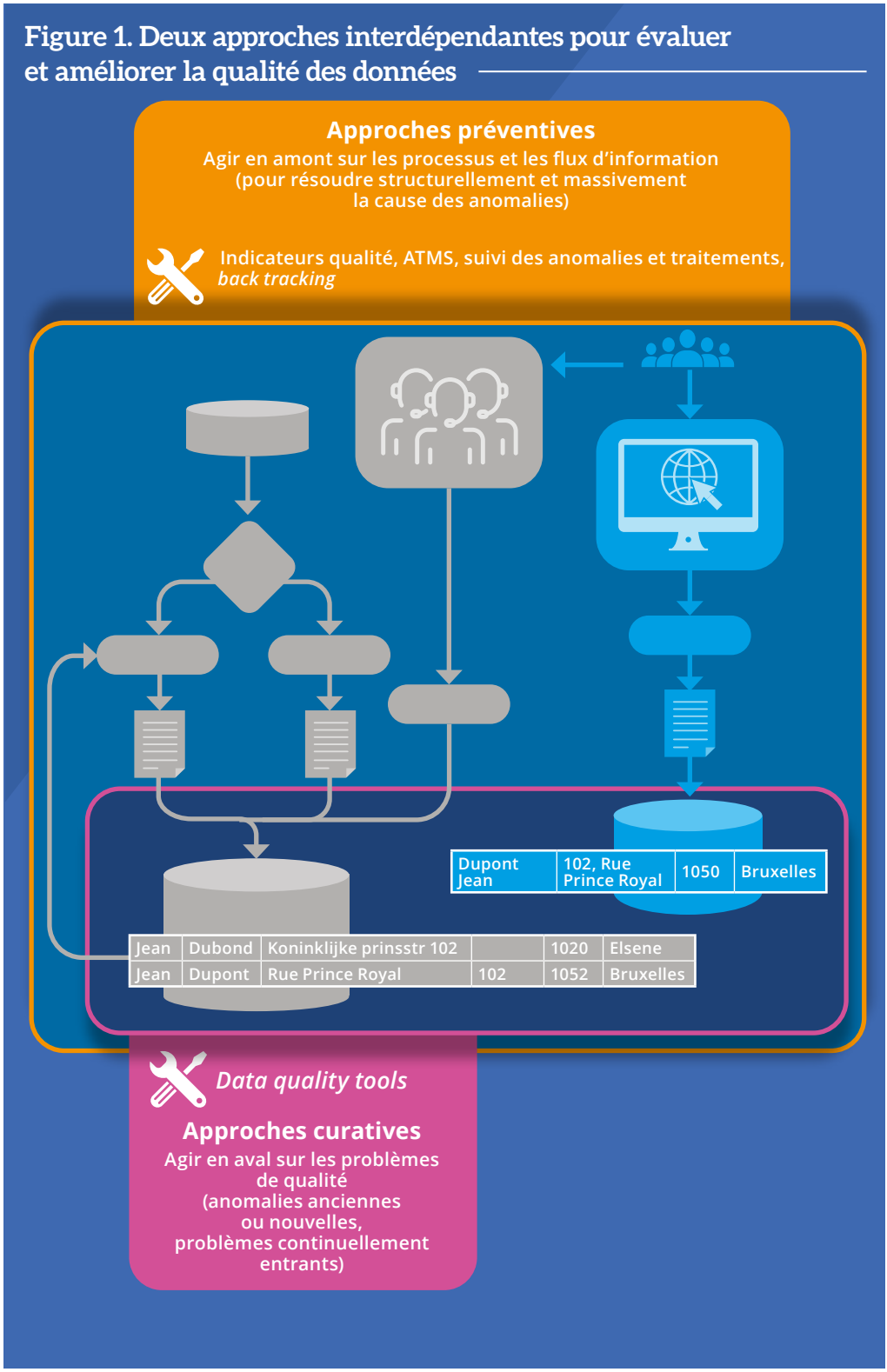
- un intitulé (i), renvoyant à un concept (une *catégorie d'activité administrative*, par exemple) ;
- un domaine de définition (d), composé d'assertions formelles spécifiant l'ensemble des valeurs admises dans la base pour ce concept (une liste contrôlée de valeurs alphabétiques d'une longueur maximale l, par exemple), complétées éventuellement de règles métier se trouvant dans le code applicatif ;
- et enfin, une valeur (v) à un instant t (le *secteur de la chimie*, par exemple).

On distingue alors les *données déterministes* des *données empiriques* (Boydens, 1999). Les premières se caractérisent par le fait que l'on dispose à tout moment d'une théorie qui permet de décider si une valeur v est correcte ou pas. Ainsi en est-il d'une opération algébrique simple portant sur un objet lui-même déterministe, comme la somme de valeurs relatives à tel champ numérique d'une base de données à un instant t. Les règles de l'algèbre pas plus que l'objet évalué n'évoluant dans le temps, on peut savoir à tout moment si le résultat d'une telle somme est correct ou pas. On dispose en effet d'un référentiel stable à cette fin.

> *« En ce qui concerne les données empiriques, sujettes à l'expérience humaine, la norme évolue dans le temps avec l'interprétation des valeurs qu'elle permet d'appréhender. »*

En revanche, en ce qui concerne les données empiriques, sujettes à l'expérience humaine, la norme évolue dans le temps avec l'interprétation des valeurs qu'elle permet d'appréhender. Ainsi en est-il par exemple du domaine médical (où la théorie évolue au fil des observations sur les patients atteints par une pathologie, comme en témoignent les recherches actuelles sur le coronavirus) mais aussi des domaines juridiques et administratifs où l'interprétation des concepts légaux se transforme avec l'évolution continue de la réalité traitée et avec celle de la jurisprudence. Comment en évaluer la validité en l'absence de référentiel absolu à cette fin ?

Figure 1. Deux approches interdépendantes pour évaluer et améliorer la qualité des données



CAPTURER ET INTERPRÉTER L'ÉVOLUTION DU RÉEL OBSERVÉ

Les données empiriques s'apparentent à des concepts « mobiles » au cœur des bases de données. En d'autres termes, la signification des concepts évolue avec l'interprétation des valeurs qu'ils permettent d'appréhender et ce, en l'absence de référentiel absolu et stable.

Approfondissant ce constat épistémologique apparemment sans issue opérationnelle, il est pourtant possible de bâtir une méthode généralisable permettant d'évaluer et d'améliorer la qualité de telles informations. La question n'est plus uniquement « les données sont-elles correctes ? » mais surtout « comment les données se construisent-elles progressivement ? ».

> *« La question n'est plus uniquement « les données sont-elles correctes ? » mais surtout « comment les données se construisent-elles progressivement ? »*

Ainsi, avec la mondialisation, de nouveaux cas non prévus initialement dans les tables de référence et dans la législation peuvent se présenter à une échelle nationale. Cela peut se produire par exemple dans le domaine de l'activité énergétique : la production d'énergie géothermique est très variable d'un endroit à l'autre du globe, certaines entreprises étrangères pourront donc utiliser une classification de leurs unités d'exploitation (« énergie renouvelable ») moins précise que celle potentiellement exigée, après examen, par la législation du pays d'exploitation (« énergie géothermique »), catégorie qui, dans notre exemple, n'est pas encore prise en compte dans la table de référence, laquelle devra faire l'objet d'une adaptation, comme expliqué ci-dessous ***(figure 2)***.

Dans ce cas, il n'est pas possible de vérifier le caractère correct des valeurs de la base de données de manière déterministe. En effet, lorsqu'une incohérence apparaît entre une telle valeur saisie au sein de la base et les tables de référence permettant d'en tester la validité, il peut s'avérer indispensable, lorsque les enjeux sont stratégiques[7], de procéder à une vérification manuelle, en contactant le citoyen ou l'entreprise concernée, par exemple. Une telle intervention peut aussi souvent être mobilisée si la catégorie attendue dans le fichier de référence pour un employeur donné ne correspond pas à la catégorie déclarée, car il se peut que l'employeur ait changé de catégorie depuis son immatriculation sans que cela n'ait été enregistré (car il ne l'a pas signalé par exemple).

C'est là, entre autres, que résidera l'intérêt d'un ATMS, en vue d'enregistrer et d'historiser les anomalies et transactions, permettant un suivi continu de celles-ci et une modification éventuelle ultérieure du domaine de définition pour l'adapter à une réalité nouvellement observée.

Illustrons ce mécanisme ***(figure 2)*** avec un autre exemple concret. En 2005, la catastrophe de l'ouragan Katrina a fait plus de 1 800 morts aux USA. Les instruments de mesure destinés à alerter les citoyens afin qu'ils quittent la zone existaient. Mais *a posteriori*, on s'est rendu compte que les bases de données qui les alimentaient n'étaient pas conçues pour intégrer l'évolution de certains phénomènes qui, alors sous-estimés, se sont révélés pourtant déterminants : il s'agissait de la montée des eaux dans les océans suite au réchauffement climatique, ainsi que de la sur-construction qui ne permet plus l'écoulement rapide de l'eau dans les sols. L'évacuation de la population fut dès lors beaucoup trop tardive. Ces mutations du réel sont encore à l'œuvre de nos jours dans les domaines hydrologiques et climatiques (Boydens, Hamiti et Van Eeckhout, 2020).

7. Par enjeux stratégiques, nous entendons des enjeux fondamentaux au regard du domaine d'application et des objectifs poursuivis : dans le domaine de la Sécurité sociale, par exemple, il peut s'agir du calcul des cotisations dues par un employeur (les taux variant avec la catégorie d'activité) ou bien des droits sociaux du travailleur (accès aux soins de santé, droit au chômage, etc.), lesquels peuvent dépendre entre autres de la validité des données signalétiques de l'employeur et de l'interprétation des anomalies associées, « stratégiques » elles aussi.

ESSAI DE TYPOLOGIE DES ANOMALIES

Une typologie des anomalies se profile alors, en fonction de leur cause potentielle et de la manière de les envisager :

- **erreur formelle certaine** due à l'intervention humaine lors de la mise à jour (champ obligatoire non complété, par exemple) ;
- **présomptions d'erreurs formelles** : présomptions de doubles ***(figure 1)*** par exemple dues à des processus redondants en amont ou incohérence avec une table de référence dont on ignore si elle a été mise à jour ;
- **erreur indétectable formellement *a priori***[8] : par exemple, omission d'une mise à jour.

Les deux derniers cas de figure peuvent quant à eux dénoter d'anomalies dues à l'évolution dans le temps du domaine empirique représenté et à l'émergence de nouveaux concepts non pris en compte ***(figure 2)***.

Selon les besoins du métier, on décidera de considérer ces anomalies comme :

- **bloquantes** : elles sont rejetées de la base de données en vertu de l'hypothèse du monde clos précédemment évoquée ;
- **non bloquantes** : les valeurs sont tout de même intégrées selon des modalités variables au sein du système d'information avec l'enregistrement correspondant, pour deux familles de raisons :

  - les rejeter du système ralentirait le processus métier (par exemple, le prélèvement des cotisations sociales) et elles ne sont pas considérées comme « stratégiques » (voir *supra*) ;

  - les prendre en considération dans le système d'information est indispensable, car elles sont considérées comme stratégiques et sont liées à des données empiriques dont la définition est potentiellement évolutive. À partir d'un certain seuil à évaluer par les spécialistes du domaine, leur traitement demande une interprétation humaine, car elles peuvent dénoter de l'émergence de phénomènes qu'il importera de prendre en considération dans le système d'information ***(figure 2)***, moyennant une gestion de versions. En outre, elles trouvent potentiellement leur origine dans les flux alimentant la base de données, problématique qui, une fois identifiée, pourra être structurellement résolue, comme nous le verrons plus loin avec le *back tracking*.

La décision consistant à identifier les anomalies empiriques « non bloquantes » est sensible en ce qu'elle relève d'une connaissance prévisionnelle des réalités traitées à un instant t, élément lui-même évolutif susceptible de faire l'objet d'une adaptation concertée au sein du système d'information. Ceci nous renvoie à la question épistémologique de la « boucle herméneutique »[9] (Boydens, 1999 ; 2012).

Comment prendre en considération les « anomalies non bloquantes » et leurs traitements, sans affecter ni la performance, ni l'intégrité des données en production ?

Avec l'ATMS, ou *Anomalies and Transactions Management System,* on passe de « l'hypothèse du monde clos » à celle d'un « monde ouvert » sous contrôles automatisés.

8. Ces cas peuvent être uniquement cernés indirectement, *via* des moyens latéraux, dont l'ATMS (voir *infra*).
9. La démarche herméneutique consiste à envisager les phénomènes empiriques en termes d'interactions par rapport à un cadre conceptuel plus général construit en vue de leur conférer un sens. Cependant, toute démarche interprétative soulève un paradoxe : celui du « cercle herméneutique » (Aron, 1969). Chaque observation ne prend sens que confrontée à un ensemble, à une « précompréhension ». Or, la sémantique de l'ensemble repose elle-même sur l'interprétation des éléments qui le constituent. Le processus de construction que suppose l'herméneutique est par nature toujours inachevé.

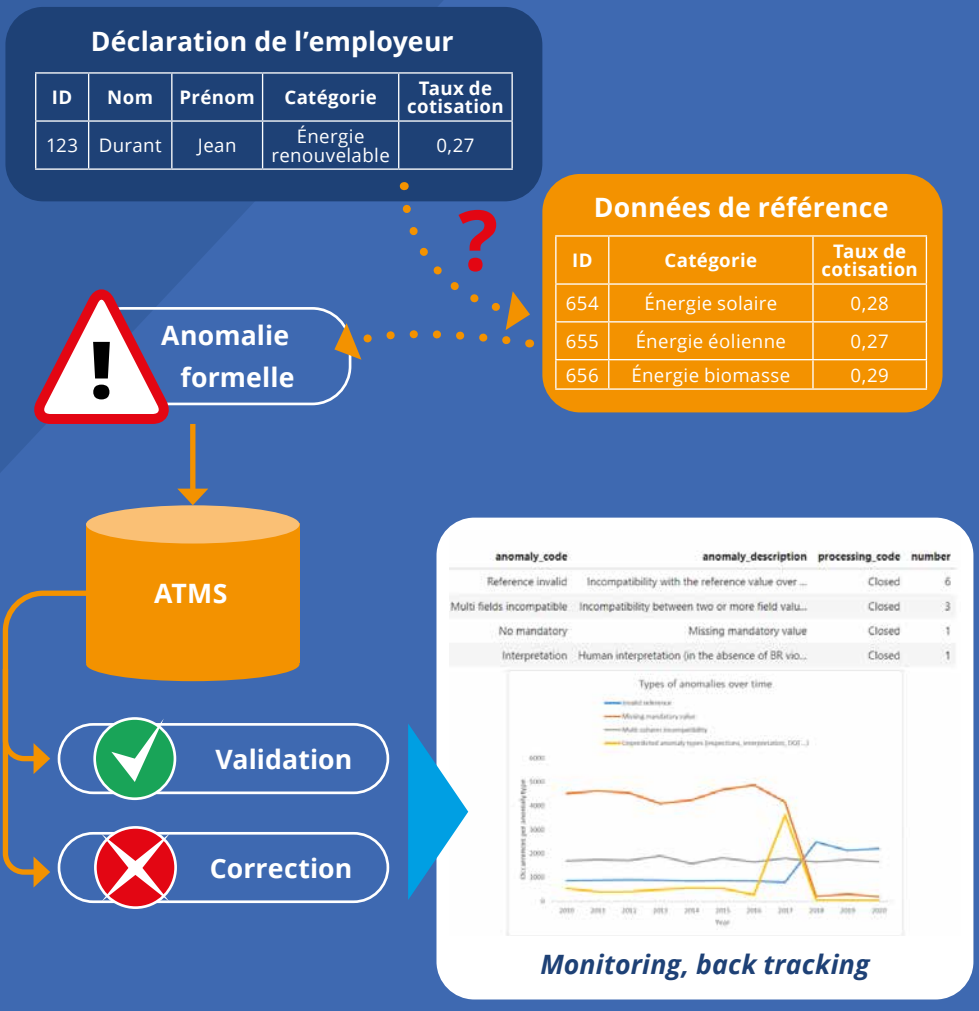
TROIS ÉCHELLES TEMPORELLES EN INTERACTION CONTINUE

« Solidaires, mais asynchrones. »

L'évolution de la norme, les transformations opérées au sein des bases de données, et la mouvance des « phénomènes » observables sur le terrain sont solidaires. Solidaires, mais asynchrones. Elles opèrent, suivant leur nature, au sein d'échelles de temps différentes.

Figure 2. Violation de « l'hypothèse du monde clos » dans un domaine empirique

Exemple : un test d'intégrité avant l'entrée des données dans la base de données principale détecte une anomalie formelle. Le traitement de l'anomalie (validation ou correction) est stocké dans l'ATMS et alimente un tableau de bord qui aidera à la prise de décision en vue d'améliorer la qualité des données.



Appliquant la notion de « temporalités étagées » de l'historien Fernand Braudel (Braudel, 1949) à l'étude des systèmes d'information administratifs en vue d'en améliorer la qualité, on peut distinguer :

- « **le temps long** » des normes (juridiques et plus largement empiriques), dont le rythme d'évolution est relativement plus lent (ainsi, dans le domaine de la Sécurité sociale belge, les modifications législatives impliquant autant de versions de schéma sont trimestrielles) ;
- le « **temps intermédiaire** » de la gestion des bases de données relativement plus rapide, ne fut-ce qu'en raison des évolutions technologiques ;
- et le « **temps court** » du réel observable, celui des citoyens ou des entreprises assujettis à l'administration, dont l'évolution est continue (Boydens, 1999).

Le concept de « temporalités étagées », théorisé par Fernand Braudel est ainsi une construction permettant d'identifier une hiérarchie entre plusieurs séquences de transformation inter-agissantes. Cette approche peut être complétée par les travaux du philosophe allemand Norbert Elias et sa notion de « continuum évolutif » (Elias, 1986). En effet, les interactions entre temporalités ne sont ni déterministes, ni unidirectionnelles, ce que laisserait entendre le modèle de Braudel seul, au sein duquel les séquences les plus lentes déterminent les plus rapides. Illustrons ces interactions par un exemple concret.

> *« Les interactions entre temporalités ne sont ni déterministes, ni unidirectionnelles. »*

En 1986, une équipe de scientifiques britanniques, spécialistes de l'étude du globe, signala la chute des taux d'ozone dans la stratosphère. Sur la base de cette observation, des chercheurs de la Nasa réexaminèrent leurs bases de données stratosphériques distribuées de par le monde ; ils découvrirent que depuis une décennie déjà, le phénomène de la baisse des taux d'ozone était resté occulté du fait que les valeurs faibles correspondantes avaient été systématiquement considérées comme des erreurs de mesure. En effet, la théorie scientifique de l'époque, modélisée dans leurs bases de données, ne leur avait pas permis de concevoir que de telles valeurs puissent être valides. Par la suite, le domaine de définition de la base a été adapté afin de considérer comme valides des taux faibles antérieurement en état d'anomalie (Boydens, 1999). De nos jours, ces phénomènes continuent d'évoluer.

D'un point de vue dynamique, une base de données idéale devrait donc calquer le rythme de ses mises à jour sur la répartition – imprévisible – en « temporalités étagées » des évolutions de la réalité qu'elle appréhende. À ce qui ressemble à une gageure s'ajoute la nécessité, toujours révélée *a posteriori*, d'intégrer des observations imprévues, interdites *a priori* par l'hypothèse du monde clos et se révélant notamment à travers les anomalies évoquées plus haut.

ANOMALIES AND TRANSACTIONS MANAGEMENT SYSTEM : PRÉSENTATION FONCTIONNELLE

L'ATMS ou *Anomalies and Transactions Management System*[10] aide à repérer l'émergence et les augmentations de « validations » d'anomalies jugées stratégiques lors de la phase de traitement manuel. Une opération de validation signifie qu'après examen, un agent a estimé que l'anomalie correspondait dans les faits à une valeur pertinente. L'opérateur peut alors « forcer » le système à accepter la valeur sans affecter l'intégrité de la base de données principale ; le dispositif doit prévoir notamment un système de gestion de versions. Selon les droits d'accès, les agents ont accès à la fois à la base de données principale et à l'ATMS : ils peuvent donc à tout moment visualiser toutes les données, qu'elles soient ou pas en état d'anomalie potentiel, et quel que soit leur stade de traitement (correction, validation, etc.) ***(figure 3)***.

Si le taux de telles validations d'anomalies est élevé et récurrent ou si l'anomalie validée est stratégique, la possibilité existe que le domaine de définition de la base lui-même ne soit plus pertinent. *A priori* l'approche s'intéresse aux cas systématiques, mais elle peut également couvrir des cas peu nombreux touchant des types d'anomalies sensibles pour le métier (émergence d'une pathologie rare, par exemple).

Un algorithme peut alors émettre un « signal » destiné aux gestionnaires de la base afin qu'ils examinent si une modification structurelle de son domaine de définition, voire une révision de la norme correspondante (législation, théorie, etc.) sont requises. Les fluctuations des données environnementales ou administratives illustrées plus haut exemplifient les cas où ce mécanisme est à l'œuvre. En outre, il s'agit de conserver l'historique de leur traitement (une même anomalie pouvant être corrigée ou validée à plusieurs reprises suite à des inspections de terrain ou à l'interprétation des réglementations).

En l'absence d'une telle intervention, l'écart entre la base de données et le réel se creuserait. En effet, si l'on omet d'adapter le schéma, les anomalies correspondant à ces cas vont continuer de croître, nécessitant un examen manuel potentiellement lourd, susceptible de ralentir le traitement des dossiers et d'affecter la qualité des données avec des impacts financiers ou sociaux.

En amont, l'ATMS aide à améliorer la qualité des bases de données « source » et fournit divers indicateurs sur l'état du traitement des anomalies, aux personnes impliquées dans la gestion du système d'information (maître d'œuvre, maître d'ouvrage). Il permet par exemple :

- d'identifier les « pics » d'anomalies, de corrections et de validations (pouvant entraîner quant à elles une restructuration du schéma de la base) ;
- d'identifier les anomalies qui ne seraient jamais traitées (ni corrigées, ni validées) ;
- de déterminer le temps de stabilisation de la base de données, au fil des traitements des anomalies spécifiées, en fonction des besoins, et le moment le plus opportun en vue d'en tirer une photographie pour l'exploiter à d'autres fins.

10. Conceptuellement, logiquement et physiquement, l'ATMS est une base de données. Le terme de « système » désigne un cadre plus large, comprenant les processus, les applicatifs et l'équipe de gestion de la base, de traitement des anomalies, de même que les fournisseurs de l'information.

Certains de ces indicateurs pourraient s'avérer utiles également pour mieux exploiter les sources administratives à des fins statistiques. En support au *back tracking*, l'ATMS est un instrument plus puissant encore d'amélioration de la qualité des données, dans une perspective préventive ***(figure 1)***.

La méthode du *back tracking* est généralisable à tout domaine d'application[11]. Elle fut initiée par Thomas Redman, sous l'appellation *data tracking*.

LE *DATA TRACKING* DE THOMAS REDMAN

Le *data tracking* proposé par Thomas Redman d'AT&T Labs aux USA (Redman, 1996) vise à évaluer quantitativement la validité formelle des valeurs introduites dans une base de données et à en améliorer structurellement le traitement (Redman écarte explicitement les questions d'interprétation des données qu'il juge trop complexes).

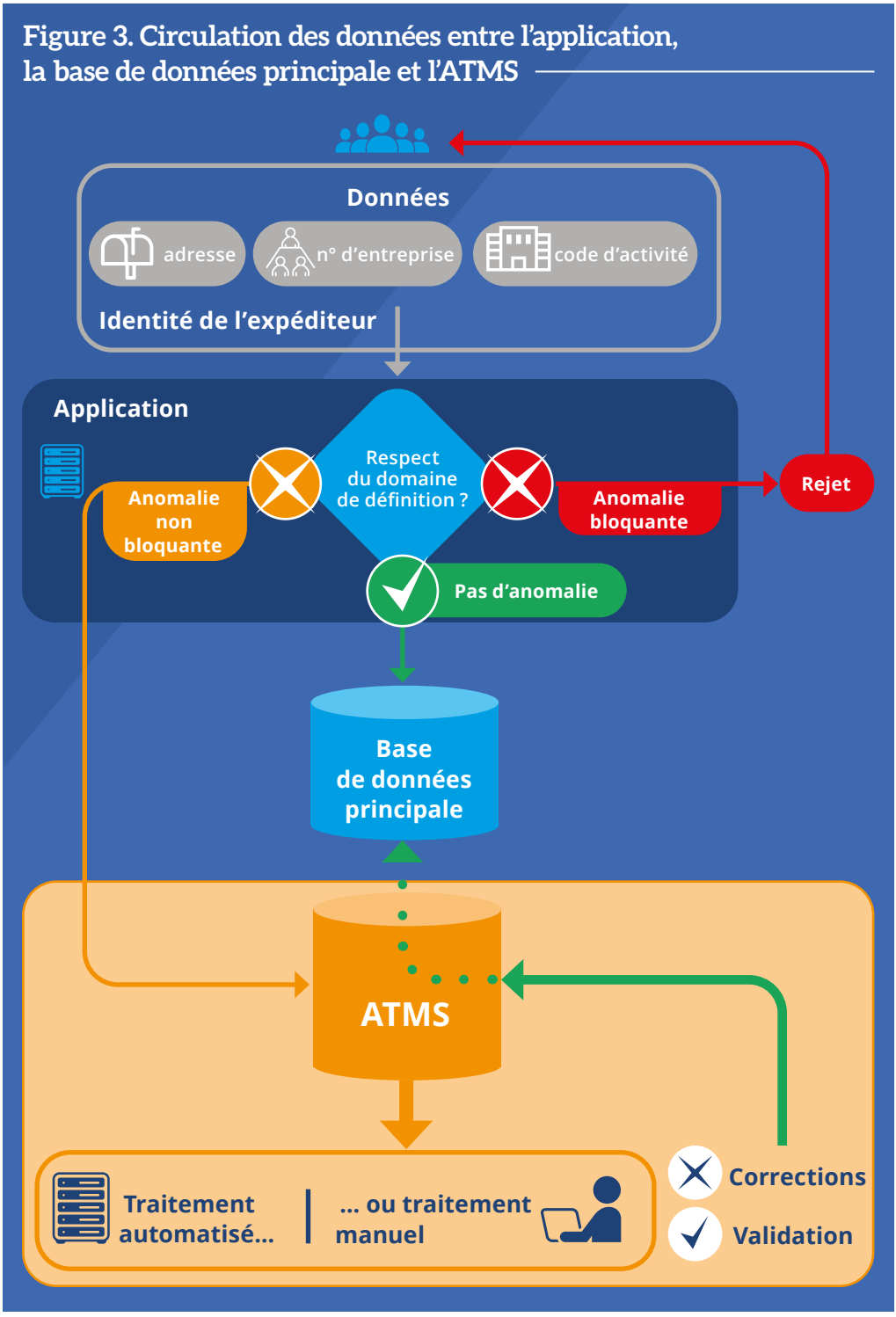
> *« Évaluer quantitativement la validité formelle des valeurs introduites dans une base de données et en améliorer structurellement le traitement. »*

Une base de données s'apparente à un lac, selon Redman. Au lieu de nettoyer ponctuellement le fond du lac continûment alimenté par des flux et courants externes (comme le préconise le *data cleansing*, méthode de correction automatique), Redman propose, sur la base d'un échantillon aléatoire de données prélevé en entrée (amont) du système d'information, d'analyser méthodiquement les processus et les flux permettant l'assemblage des données. Le but final consiste à déterminer les causes des erreurs formelles identifiées afin d'y remédier structurellement à la source (bogues ou erreurs de programmation, par exemple).

L'opération repose sur le principe selon lequel un petit nombre de flux, processus ou pratiques sont à l'origine d'un pourcentage important d'erreurs formelles. L'approche fait référence au principe de Pareto également appelé « principe 80/20 ». Elle est ainsi basée sur l'hypothèse selon laquelle une part importante des cas problématiques (environ 80 %) est engendrée par environ 20 % des causes possibles.

11. Par domaine d'application, on entend, en modélisation de bases de données, le segment du réel observable que l'on tente de représenter à travers la base.

Figure 3. Circulation des données entre l'application, la base de données principale et l'ATMS



LE *BACK TRACKING* : DE L'ORTHOGONALITÉ À L'INTERACTION

Dans le courant des années 2000, l'application à grande échelle de la méthode de Redman (Boydens, 2010) a abouti à une méthode originale dénommée *back tracking* (Boydens, 2018). En se basant toujours sur le principe de Pareto posé par Redman, le *back tracking* enrichit sa méthode sur cinq aspects importants :

- le modèle de la base de données est étendu et relié à un historique des anomalies et de leurs traitements selon les principes exposés *supra* ;
- un suivi continu des cas jugés les plus stratégiques est mis en place à partir de l'ATMS, de façon à faciliter la gestion de la qualité de la base de données. Le suivi du traitement des anomalies permet de détecter, dans les domaines d'application fortement évolutifs, l'émergence de nouveaux phénomènes observables demandant une adaptation ponctuelle du domaine de définition de la base de données, voire des normes associées, en vue de diminuer le nombre d'anomalies fictives à traiter ;
- l'échantillon de cas retenus n'est pas aléatoire comme chez Redman, puisque l'on dispose d'une connaissance *a priori* concernant la totalité des données jugées problématiques (*via* l'historique exhaustif des anomalies et de leurs traitements). L'approche permet une sélection plus précise et représentative des cas à investiguer dès le début de l'opération, réduisant ainsi l'inévitable marge d'erreur d'un échantillon aléatoire ;
- au-delà de l'erreur formelle (déterministe et détectable *via* un algorithme, cas uniquement pris en compte par Redman), les questions d'interprétation des données au fil de l'évolution de la législation (ou de toute théorie empirique mobilisée) et des réalités appréhendées sont également abordées ;
- il s'agit d'un *tracking* arrière (ou *back tracking*) : on part, en synergie avec les fournisseurs de la base de données, de la situation finale (base de données principale et ATMS) pour revenir, étape par étape, à chaque source et processus qui en a permis l'élaboration, jusqu'à l'identification des causes à l'origine de cas problématiques. L'objectif est d'éviter le traitement de données ou de flux inutiles et de travailler de manière plus économe, renforçant les gains de l'opération. En effet, la recherche des origines structurelles des anomalies prend fin dès que toutes leurs causes par type[12] ont été détectées, sans que tous les flux ne soient inutilement parcourus (dans le cas du *data tracking* de Redman, tous les flux doivent être parcourus, ce qui constitue une perte de temps, ceux-ci pouvant inclure des dizaines, voire des centaines de processus).

> *« L'opération de back tracking repose ainsi sur un suivi préalable des anomalies et transactions, lui-même mis en place après spécification des indicateurs de qualité stratégiques. »*

L'opération de *back tracking* repose ainsi sur un suivi préalable des anomalies et transactions, lui-même mis en place après spécification des indicateurs de qualité stratégiques. Elle permet ensuite d'identifier, au sein des processus et des flux de données, en partenariat avec le fournisseur de l'information et le gestionnaire de la base[13], les éléments à l'origine

12. Comme évoqué *supra*, les anomalies sont spécifiées à un instant t par référence à un domaine de définition : valeur absente, valeur incohérente par rapport à une autre, etc. Ces types d'anomalies sont susceptibles d'évoluer dans le temps, si nécessaire, *via* une gestion de versions du domaine de définition.
13. Il s'agit du maître d'œuvre, du maître d'ouvrage ou de tout gestionnaire ayant une prise sur la base de données au moment de l'identification des éléments problématiques.

de la production d'un grand nombre d'anomalies systématiques ou jugées stratégiques : traitement inapproprié de certaines sources de données, émergence de situations nouvelles non encore prises en compte dans le domaine de définition de la base, interprétation inadéquate de la législation, concept mal documenté, erreurs de programmation, etc. Sur cette base, un diagnostic ainsi que des actions correctrices durables et structurelles peuvent être posés (correction de code formel dans les programmes, restructuration de processus, adaptation de l'interprétation d'une loi, clarification de la documentation, etc.).

Dans le domaine de la Sécurité sociale belge (en l'occurrence, à partir de la base de données DmfA mentionnée *supra*), les tests « grandeur nature » réalisés en synergie avec les développeurs, les spécialistes du domaine d'application ainsi que les expéditeurs de l'information furent concluants. Les apports de la méthode furent démontrés à maintes reprises dans les années deux mille (diminution structurelle du nombre d'anomalies de 50 % à 80 %, gain de temps grâce à une réduction du travail intellectuel fastidieux de vérification, meilleure interprétation de la loi, perception et redistribution financières plus rapides, etc.). Vu son caractère généralisable, la méthode fut actée dans la législation belge sous la forme d'un arrêté royal contraignant[14] en 2017 (Boydens, 2018). Cette législation s'inscrit dans le cadre de « baromètres de qualité » appliqués à la Sécurité sociale. Si les effets de la méthode sont durables et structurels, celle-ci doit être appliquée de manière récurrente afin de prendre en compte d'éventuels nouveaux phénomènes, ce qui demande toutefois un effort dégressif dans le temps. Elle repose sur une organisation, des procédures et un système d'information rigoureusement documentés (Boydens, 2010)[15].

Le mécanisme d'ATMS qui a soutenu les opérations de *back tracking* fut à ce jour déployé à grande échelle *via* un Système de gestion de base de données (SGBD) hiérarchique et du code externalisé spécifique associé à un moteur de règles générique. Désormais, un nouveau développement d'ATMS générique est applicable à tout SGBD relationnel et aux technologies actuelles.

L'ATMS RELATIONNEL : UNE DYNAMIQUE ENTRE DONNÉES EN PRODUCTION ET GESTION DES ANOMALIES

Le *Centre de Compétence en Qualité de Données* de Smals[16], qui repose sur une synergie entre la section *Databases* et la section Recherche, a entrepris la mise en place d'un prototype appliqué aux SGBD relationnels et aux standards associés. Le prototype repose sur la version *open data* complète de la Banque Carrefour des Entreprises, correspondant belge du répertoire des entreprises Sirène en France.

Dans ce nouveau modèle, la base de données principale et l'ATMS sont séparés, ce qui nécessite de convenir du routage des données entre les deux systèmes et de définir les principes permettant de stocker les anomalies et transactions associées.

14. Voir références juridiques en fin d'article.
15. Concrètement, au-delà d'un certain seuil d'anomalies fixé par l'administration, les fournisseurs de déclarations sociales sont contraints d'en diminuer le nombre dans un délai donné, en participant à une opération de *back tracking*.
16. Smals est une société informatique créée en 1939 prestataire de services pour l'administration fédérale et régionale belge.

BASE DE DONNÉES PRINCIPALE ET ATMS : LES DEUX HÉMISPHÈRES DU MONDE REPRÉSENTÉ

L'un des traits majeurs de la spécification de l'ATMS relationnel est la séparation induite entre la base de données principale, d'une part, et la base de données des anomalies et de leurs traitements, d'autre part. La ***figure 4*** illustre cette séparation, qui repose sur les deux principes suivants :

- la base de données principale représente les concepts du domaine d'activité qu'elle sert. À ce titre, elle est modélisée en vue de traiter des données qui respectent strictement le domaine de définition, en vertu de l'hypothèse du monde clos (voir *supra*) ;
- il est possible de factoriser l'enregistrement et le traitement des données en anomalie dans un système dédié distinct de la base de données principale, à condition que celui-ci soit conçu de façon extensible.

Au-delà de la satisfaction d'hypothèses de conception théoriques, la séparation entre base de données principale et ATMS ouvre la porte à un certain nombre d'avantages non négligeables[17] :

- la conception de la base de données principale est simplifiée, puisqu'elle s'affranchit de la prise en compte des anomalies ou de leur traitement ;
- le contenu de la base de données principale reste en permanence conforme au domaine de définition, sans en empêcher l'évolution ;
- la gestion des anomalies peut faire l'objet de processus et outils standardisés, réutilisables à l'échelle d'une équipe, d'un projet ou d'une organisation entière ;
- l'existence d'un ATMS dédié encourage une prise en compte accrue du domaine de définition au plus tôt dans la conception du système informatique, qui devra se charger de la détection des anomalies et du routage des données en conséquence.

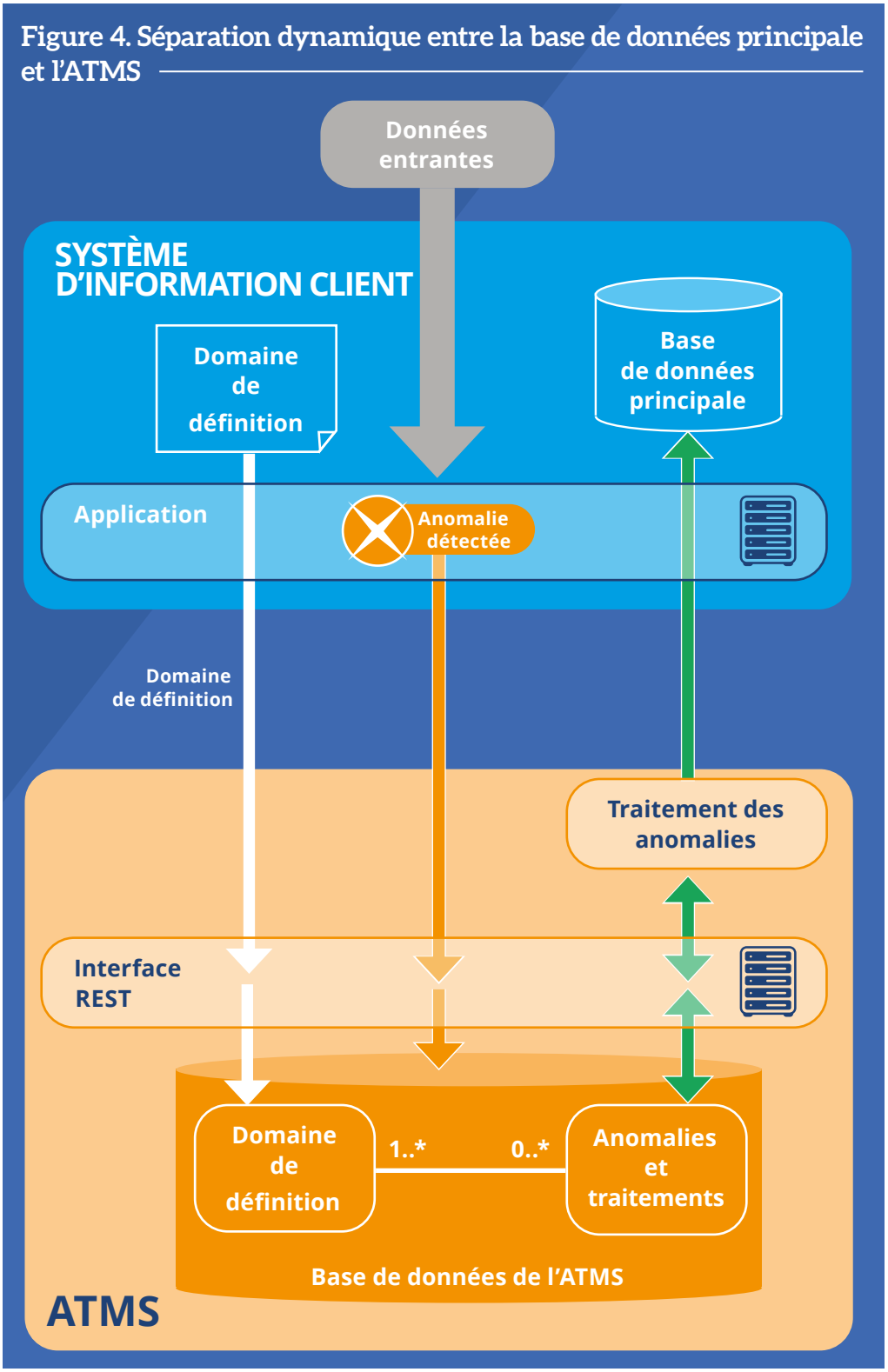
ROUTAGE DES DONNÉES

Lorsque des données entrent dans le système d'information, leur conformité au domaine de définition (voir *supra*) est vérifiée. Si des anomalies non bloquantes sont détectées, les données correspondantes sont envoyées dans l'ATMS ***(figure 3)***. Ce n'est qu'après avoir été sujettes à vérification et traitement intellectuel, historisé automatiquement, que les données seront intégrées à la base de données principale[18]. Dans un cas de validation, tel qu'évoqué précédemment, il se peut que le domaine de définition doive être adapté avant que les enregistrements ne puissent être acceptés dans la base de données principale.

Ce routage automatique n'exclut aucunement la possibilité qu'un agent déclenche manuellement une anomalie sur des données conformes au domaine de définition et déjà présentes dans la base de données principale. Ceci pourrait se justifier, par exemple, par une inspection de terrain qui révèle qu'un enregistrement est frauduleux ou obsolète.

17. En corollaire, la resynchronisation de l'état des deux bases de données en cas de catastrophe doit être pensée en amont. Diverses options plus ou moins sophistiquées sont possibles, allant de l'arrêt pur et simple de l'un des deux systèmes si l'autre ne répond plus, au recours à une file de messages au sein d'un système tiers dédié.
18. L'état complet du système d'information se compose donc de l'union – indifféremment inclusive ou exclusive, l'intersection n'étant de toute façon pas acceptée – entre les données valides et les anomalies non traitées, distribuées respectivement dans la base de données principale d'une part et dans l'ATMS d'autre part.

Figure 4. Séparation dynamique entre la base de données principale et l'ATMS



La circulation de données entre la base principale et l'ATMS doit être fluide et standardisée ; elle est donc implémentée dans des procédures automatisées, dont le principe a déjà été illustré une première fois par la ***figure 3***. La nature bidirectionnelle de cette circulation requiert de pouvoir préserver ou reconstruire l'état original de l'information à partir de l'ATMS. À cette fin, deux parties constitutives permettent de stocker respectivement le domaine de définition ainsi que les anomalies et métadonnées associées ***(figure 4)*** ; nous ne détaillerons pas ce fonctionnement ici mais plus de références sont disponibles en ligne (Boydens, Hamiti et Van Eeckhout, 2020).

IMPLÉMENTATION DU PROTOTYPE ET PERSPECTIVES

Le prototype implémente intégralement la base de données de l'ATMS. Celui-ci a été développé et testé de façon itérative en l'exposant, très tôt, à un système d'information simulé par une application rudimentaire et une base de données principale réaliste, les *open data* du Répertoire des entreprises belges ***(encadré 2)***. Dans ce cadre, quatre scénarios d'utilisation ont été implémentés à titre d'exemple :

- la correction d'une valeur simple en anomalie ;
- la correction d'une anomalie déclenchée par l'incompatibilité de données entrantes avec un enregistrement déjà existant dans la base de données principale ;
- la validation d'anomalies par lots : le traitement d'une anomalie (ici la validation) est propagé automatiquement à une série d'autres anomalies désignées comme similaires par l'agent ;

Encadré 2. Quelques détails techniques sur l'ATMS

Le prototype décrit dans cet article a été développé sur le SGBD PostgreSQL ; il est néanmoins transposable à n'importe quel système permettant la manipulation de données au format JSON. Cette notation « sans schéma » permet de stocker les anomalies les plus variées au sein d'une table relationnelle classique, ainsi que de les échanger sous forme de messages entre le système d'information principal et l'ATMS. Les traitements produisant et consommant ces messages JSON sont implémentés de part et d'autre sous la forme de procédures stockées accessibles en requêtant directement les bases de données ; en pratique, dans un projet de production, cette logique serait typiquement exposée *via* des interfaces REST.

Dans le cadre de ce prototype, le volume de la base de données principale est d'environ 4,4 gibibits (GiB) pour un total de 30,2 millions d'enregistrements répartis essentiellement sur 9 tables. L'exécution des trois premiers scénarios de la façon la plus exigeante possible (exécution en boucle de 100 000 itérations sans aucune optimisation explicite*) a permis d'observer :

- un temps d'exécution constant, souvent de l'ordre du millième de seconde ou moins, pour la plupart des opérations de calcul et d'insertion légère (par exemple, générer un message de création d'anomalie complet, allouer une anomalie pour traitement par un agent, marquer une anomalie comme traitée) ;
- une croissance linéaire du temps d'exécution pour les opérations d'écriture plus importantes (enregistrement de la version corrigée d'une anomalie) ;
- une croissance linéaire de la consommation d'espace de stockage.

** Par exemple, PostgreSQL permet de déclencher explicitement des opérations comme « VACUUM ANALYZE » afin d'optimiser le stockage et les plans d'exécution des requêtes. D'autres SGBD relationnels offrent des fonctionnalités analogues.*

- enfin, la création de plusieurs vues permettant de suivre les anomalies et leurs traitements dans le temps à un niveau plus ou moins agrégé[19]. Ce quatrième scénario est fondamental pour orienter les opérations de *back tracking* telles que décrites précédemment.

Le prototype ayant rencontré un succès technique et suscité l'intérêt de la maîtrise d'ouvrage, un projet pilote est en cours. Cette initiative pourra ultérieurement être appliquée grandeur nature à des bases de données de l'administration belge. Elle est par ailleurs généralisable à tout système d'information. Plusieurs cas de figure sont possibles pour la mise en œuvre d'un ATMS. Idéalement, celui-ci est envisagé dès la conception du système d'information, en lien avec la base de données principale et les outils de qualité de données ***(encadré 1)***, si l'on en dispose. Un projet de réingénierie représente également un moment opportun pour l'intégration d'un ATMS à un système existant[20].

Les apports récurrents du *back tracking* évoqués plus haut, méthode nécessairement supportée par un ATMS, appliqué à la base de données DmfA constituent un précédent encourageant fortement la généralisation de ce prototype adapté à des technologies récentes : diminution structurelle du nombre d'anomalies de 50 % à 80 %, gain de temps grâce à une réduction du travail intellectuel fastidieux de vérification, meilleure interprétation de la norme, perception et redistribution financières plus rapides, de manière plus générale, amélioration de la qualité de toute base empirique, mise en place d'un partenariat entre les gestionnaires de la base et les fournisseurs de l'information, etc. Comme nous l'avons vu, le service peut bénéficier tant aux gestionnaires des bases de données administratives dont la qualité est améliorée qu'aux statisticiens souhaitant les exploiter à d'autres fins.

19. Types d'anomalies les plus fréquents, anomalies validées par qui et quand, anomalies non traitées, etc.
20. Si le système original dispose déjà d'une forme de gestion des anomalies, la transposition de ce contenu à l'ATMS peut cependant représenter un certain effort en fonction du degré auquel les anomalies et leurs métadonnées de traitement sont fondues et éclatées dans le système.

BIBLIOGRAPHIE

AGENCE BELGA, 2018. Des lacunes dans la base de données belge sur les terroristes. In : *La Libre Belgique*. [en ligne]. 1er mars 2018. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.lalibre.be/belgique/des-lacunes-dans-la-base-de-donnees-belge-sur-les-terroristes-5a97a5f4cd700399f72087da*.

ARON, Raymond, 1969. *La philosophie critique de l'histoire.* 1969. Édition Librairie philosophique J. Vrin. Collection Points – Sciences humaines. ISBN 2560848158182.

BADE, David, 2011. It's about Time!: Temporal Aspects of Metadata Management in the Work of Isabelle Boydens. In : *Cataloging & Classification Quarterly (The International Observer).* 16 mai 2011. Volume 49, n° 4, pp. 328-338.

BATINI, Carlo et SCANNAPIECO, Monica, 2016. *Data and Information Quality. Dimensions, Principles and Techniques*. Springer, New York. ISBN 978-3-319-24106-7.

BENS, Arno et SCHUKRAFT, Stefan, 2019. Modernisation des registres administratifs en Allemagne – Développements actuels et enjeux pour la statistique publique. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 10-20. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168390/courstat-2-3.pdf*.

BOYDENS, Isabelle, 1999. *Informatique, normes et temps*. Bruylant, Bruxelles. ISBN 2-8027-1268-3.

BOYDENS, Isabelle, 2010. Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium. In : *Practical Studies in E-Government. Best Practices from Around the World.* 17 novembre 2010. Springer. New York. pp. 113-130 (chapitre 7). ISBN 978-1489981899.

BOYDENS, Isabelle, 2012. L'océan des données et le canal des normes. In : *La normalisation : principes, histoire, évolutions et perspectives.* [en ligne]. Juillet 2012. Annales des Mines, Responsabilité et Environnement. Édition FFE. Paris. N° 2012/3 (67), pp. 22-29. [Consulté le 31 mai 2021]. Disponible à l'adresse : *http://www.ulb.ac.be/cours/iboydens/annales.pdf*.

BOYDENS, Isabelle, 2018. *Data Quality & « back tracking » : depuis les premières expérimentations à la parution d'un Arrêté Royal.* [en ligne]. 14 mai 2018. Smals Research. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.smalsresearch.be/data-quality-back-tracking-depuis-les-premieres-experimentations-a-la-parution-dun-arrete-royal/*.

BOYDENS, Isabelle, 2021. *Qualité de l'information et des documents numériques*. Cours dispensé à l'Université libre de Bruxelles, Master en sciences et technologies de l'information et de la communication.

BOYDENS, Isabelle, HAMITI, Gani et VAN EECKHOUT, Rudy, 2020. *Data Quality: "Anomalies & Transactions Management System" (ATMS), prototype and "work in progress"*. [en ligne]. 8 décembre 2020. Smals Research. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.smalsresearch.be/data-quality-anomalies-transactions-management-system-atms-prototype-work-in-progress/*.

BRAUDEL, Fernand, 1949. *La Méditerranée et le monde méditerranéen à l'époque de Philippe II.* Armand Colin, Paris.

BYRNES, Nanette, 2016. Why we should expect algorithms to be biased. In : *MIT Technology Review.* [en ligne]. 24 juin 2016. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.technologyreview.com/2016/06/24/159118/why-we-should-expect-algorithms-to-be-biased/*.

CHAPUIS, Nicolas, 2018. Des fichiers de police mal organisés et trop complexes. In : *Le Monde.* 17 octobre 2018.

DIERICKX, Laurence, 2019. *Why News Automation Fails*. [en ligne]. Février 2019. Computation & Journalism Symposium, Miami, USA. [Consulté le 31 mai 2021]. Disponible à l'adresse : *http://mastic.ulb.ac.be/wp-content/uploads/2019/02/Why_news_automation_fails.pdf*.

ELIAS, Norbert, 1986. *Du temps.* Édition Fayard. Paris. ISBN 978-2818503454.

HAINAUT, Jean-Luc, 2018. *Bases de données – Concepts, utilisation et développement.* Octobre 2018. Édition Dunod, Paris, collection InfoSup. 4e édition. ISBN 978-2100790685.

HAMITI, Gani, 2019. *Data Quality Tools: concepts and practical lessons from a vast operational environment*. [en ligne]. 13 mars 2019. Université libre de Bruxelles. Cours-conférence. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://mastic.ulb.ac.be/wp-content/uploads/2019/03/Data_Quality_Tools_ULB_2019.pdf*.

HAND, David J., 2018. Statistical challenges of administrative and transaction data. In : *Journal of the Royal Statistical Society.* [en ligne]. Series A, 181, Part 3, pp. 555-605. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://spiral.imperial.ac.uk/bitstream/10044/1/61527/2/Statistical%20challenges%20of%20administrative%20and%20transaction%20data%20FINAL%20version.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 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.

MADNICK, Stuart E., WANG, Richard Y., LEE, Yang W. et ZHU, Hongwei, 2009. Overview and Framework for Data and Information Quality Research. In : *Journal of Data and Information Quality.* [en ligne]. 1er juin 2009. Volume 1, n° 1, pp 1–22. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://dl.acm.org/doi/10.1145/1515693.1516680*.

RADIO, Eric, 2014. Information Continuity: A Temporal Approach to Assessing Metadata and Organizational Quality in an Institutional Repository. In : *Metadata and Semantics Research*. 27-29 novembre 2014. 8th Research Conference, MTSR 2014, Karlsruhe. Springer, Cham. Communications in Computer and Information Science, vol 478. ISBN 978-3-319-13673-8.

REDMAN, Thomas C., 1996. *Data Quality for the Information Age*. Artech House Computer Science Library. ISBN 978-0890068830.

RENNE, Catherine, 2018. Bien comprendre la déclaration sociale nominative pour mieux mesurer. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 35-44. [Consulté le 31 mai 2021]. *https://www.insee.fr/fr/statistiques/fichier/3647029/courstat-1-7.pdf*.

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 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647013/courstat-1-5.pdf*.

RIVIÈRE, Pascal, 2020. Qu'est-ce qu'une donnée ? Impact des données externes sur la statistique publique. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. N° N5. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008707/courstat-5-8.pdf*.

SRIVASTAVA, Divesh, SCANNAPIECO, Monica et REDMAN, Thomas C., 2019. Ensuring High-Quality Private Data for Responsible Data Science: Vision and Challenges. In : *Journal of Information and Data Quality.* 4 janvier 2019. Volume 1, n° 11, pp. 1-9.

VAN DER VLIST, Eric, 2011. *Relax NG*. Mai 2011. Édition O'Reilly Media. ISBN: 0596004214.

FONDEMENTS JURIDIQUES

Arrêté royal du 2 février 2017 modifiant le chapitre IV de l'arrêté royal du 28 novembre 1969 pris en exécution de la loi du 27 juin 1969 révisant l'arrêté-loi du 28 décembre 1944 concernant la sécurité sociale des travailleurs. In : Moniteur belge. [en ligne]. [Consulté le 31 mai 2021]. Disponible à l'adresse : *http://www.ejustice.just.fgov.be/mopdf/2017/02/20_2.pdf#Page113*.

LE CONSEIL NATIONAL DE L'INFORMATION STATISTIQUE

LA QUALITÉ DES STATISTIQUES PUBLIQUES PASSE AUSSI PAR LA CONCERTATION

Isabelle Anxionnaz et Françoise Maurel***

La mission principale du Cnis est d'assurer la concertation entre producteurs et utilisateurs de statistique publique. Instance représentative de la société civile, transverse au Service statistique public (SSP), légitime et pérenne, le Cnis est le lieu d'échanges sur l'offre et la demande de statistiques, et le vecteur qui porte les projets du SSP à la connaissance de tous. La transparence du débat, sur des sujets ponctuels comme sur les orientations générales, permet au SSP de partager une vision prospective avec les utilisateurs et de s'assurer de la pertinence de ses productions. Les orientations de moyen terme du Cnis débouchent sur une feuille de route pluriannuelle, qui est ainsi la résultante des besoins des utilisateurs et des avancées des producteurs. Sur les sujets émergents présentant un intérêt pour le débat public, les expertises croisées au sein des groupes de travail donnent lieu à des recommandations adressées aux producteurs. Le Cnis contribue ainsi à augmenter la confiance dans les statistiques et leur acceptabilité sociale. Son organisation rodée ne l'empêche pas de s'adapter en continu aux évolutions des usages comme à celles de la production statistique. Il s'agit moins aujourd'hui d'explorer des thématiques vierges que de pousser à développer de nouvelles méthodes de production, à les rationaliser ou à mieux articuler les sources existantes.

The CNIS' main mission is to ensure consultation between producers and users of official statistics. As a representative of civil society, it is a legitimate and permanent body that cuts across the official statistical service. CNIS is both a forum for discussion on the supply and demand for statistics and a vehicle for informing all users of official statistics projects. As a forum for transparent debate on specific topics as well as on general orientations, it enables Official statistics to share a prospective vision with users and to ensure the relevance of its productions. The five-year roadmap, made by the midterm orientations formulated by Cnis, is henceforth at the crossroad between the needs expressed by users and progress made by producers. For emerging topics of particular interest in the public debate, the shared expertise within working groups leads to recommendations geared towards producers. CNIS thus contributes to enhance trust in the official statistics and their social acceptability. Its well-established organisation does not prevent it from continuously adapting to changes in usage and in statistical production. Nowadays the issue is less to explore new thematic fields than to promote the development of new production methods, their rationalisation or better articulation between existing sources.

* Secrétariat général du Cnis, Insee, *isabelle.anxionnaz@insee.fr*
** Secrétariat général du Cnis, Insee, *francoise.maurel@insee.fr*

ASSURER LA CONCERTATION ENTRE PRODUCTEURS ET UTILISATEURS…

Dans le triptyque de la gouvernance de la statistique publique, tel qu'il a été instauré par la loi de Modernisation de l'économie du 4 août 2009[1], le Conseil national de l'information statistique se place aux côtés du Service statistique public et de l'Autorité de la statistique publique (Bureau, 2020). Situé en amont du dispositif, sa mission principale est d'assurer la concertation entre les producteurs et les utilisateurs de la statistique publique. Instance représentative de la société civile par le large éventail des organismes qui y contribuent, le Cnis est à la fois le lieu de rencontre de l'offre et de la demande de statistiques, et le vecteur qui porte les projets de la statistique publique à la connaissance de tous les utilisateurs, pour s'assurer qu'ils répondent à leurs besoins (Duport, 2009).

… DANS UN CADRE COMMUN À L'ENSEMBLE DE LA STATISTIQUE PUBLIQUE

Les statistiques publiques couvrent un champ très large et ont vocation à répondre aux besoins de tous les publics, par nature extrêmement variés. Entre le grand public qui souhaite connaître le chiffre de référence de l'inflation ou le chercheur en sciences sociales qui veut exploiter les données individuelles de l'enquête cadre de vie et sécurité, par exemple, figure un continuum d'utilisateurs plus ou moins experts, qui s'appuient sur les résultats statistiques pour des usages professionnels ou privés et ce, dans tous les domaines. Face à la richesse de l'offre statistique et à la diversité des usages et des utilisateurs, comment assurer le contact entre les deux mondes des producteurs et des utilisateurs de statistique de manière cohérente et homogène ?

> *« Comment assurer le contact entre les deux mondes des producteurs et des utilisateurs de statistique de manière cohérente et homogène ? »*

Chaque producteur statistique connaît peu ou prou ses grandes catégories d'utilisateurs et peut évidemment organiser des échanges avec un échantillon d'entre eux pour faire émerger de nouveaux besoins ou pour évaluer la pertinence des statistiques existantes. Il le fait d'ailleurs régulièrement : ainsi, les statisticiens d'entreprises rencontrent périodiquement les organisations professionnelles pour les informer et pour solliciter leur avis sur les nouveaux travaux ; les services statistiques ministériels recueillent la demande des directions thématiques de leur ministère ; l'Insee organise des conférences de presse fréquentes et met au besoin les journalistes en relation directe avec ses experts ; à l'occasion des nouveaux projets d'enquête, les concepteurs constituent des comités de concertation, associant notamment des chercheurs. La demande des instances européennes est quant à elle traitée au niveau du Système statistique européen et intégrée à la source aux projets statistiques nationaux.

Mais avec une concertation totalement décentralisée et au cas par cas, se pose une question de cohérence et d'homogénéité de traitement de tous les utilisateurs, à laquelle s'ajoute la difficulté de rendre compte de l'ensemble de ces échanges, sans un cadre commun.

1. Voir les références juridiques en fin d'article.

Le cadre qui assure cette transversalité est, en France, le Conseil national de l'information statistique (Cnis) : il permet, à l'échelle de l'ensemble de la Statistique publique, de faire le lien entre tous les producteurs thématiques et toutes les catégories d'utilisateurs.

La mission de concertation du Cnis est inscrite dans la loi statistique de 1951, elle en fait un lieu légitime pour informer tous les utilisateurs sur les projets statistiques des producteurs, mais aussi pour recueillir les besoins et organiser le débat. Le Cnis contribue également par sa position transverse à la coordination du Service statistique public (SSP) : la concertation assurée par le Cnis débouche ainsi sur des « *propositions pour l'élaboration du programme de travaux statistiques* ».

DÉFINIR ET PARTAGER UNE VISION DE MOYEN TERME…

Bien que très stables par nature, les productions statistiques évoluent en continu, en particulier en fonction de la demande exprimée au sein du Cnis. Ainsi les grandes tendances de transformation des statistiques et la plupart des innovations majeures sont accompagnées dès l'origine par des travaux de concertation entre producteurs et utilisateurs. Un exemple transverse illustre ce point, celui des sources administratives.

En 2004, les travaux du Cnis dressaient déjà un bilan très favorable de l'utilisation des sources administratives pour la production des statistiques, notamment pour répondre au besoin croissant de données localisées résultant de la décentralisation, mais ces usages étaient encore peu nombreux. Dans sa vision prospective (Cnis, 2004), le Cnis encourageait alors la poursuite des efforts des producteurs : « *Le Conseil souligne l'importance qu'il attache à l'utilisation des sources administratives à des fins statistiques afin d'alléger la charge de collecte tout en améliorant la réponse aux besoins d'information au niveau local* ».

En 2009, au vu des avancées obtenues depuis 2004, le Cnis précisait les attentes des utilisateurs sous l'angle de la complémentarité avec les enquêtes : « *Le Conseil soutient par ailleurs les travaux méthodologiques entrepris pour coupler les données administratives et les données d'enquêtes afin de suivre les parcours individuels qui permettent de mieux rendre compte de la diversité des situations notamment en matière économique et sociale…*» (Cnis, 2009).

Les dernières orientations de moyen terme du Cnis (2019-2023) signalent la plus-value de nouveaux dispositifs agrégeant plusieurs sources administratives : « *Le Conseil demande à l'ensemble des producteurs de la statistique publique de développer les appariements entre sources de données afin d'enrichir l'analyse des liens entre différents thèmes, en veillant au strict respect de la confidentialité lorsque les appariements reposent sur des informations identifiantes* » (Cnis, 2019).

Au-delà de cet exemple, la vision partagée des orientations de moyen terme dégagée par le Cnis est ainsi la résultante des (nouveaux) besoins exprimés par les utilisateurs et des avancées de la part des producteurs qui les rendent suffisamment réalistes. (Geoffard, 2019) rappelle ainsi le processus de création de l'identifiant national élève dans les systèmes d'information statistiques dans le domaine de l'éducation nationale et le rôle déterminant que le Cnis a joué dans celui-ci. Les avancées en matière d'information sur les groupes d'entreprises trouvent aussi leur origine dans un groupe de travail du Cnis exprimant des besoins importants de connaissance (Salustro, Ménard et Depoutot, 2008). Dans un autre domaine, la demande pressante du Cnis de la fin du siècle dernier pour disposer de plus de données localisées est aujourd'hui bien mieux satisfaite grâce à l'usage de sources exhaustives et aux progrès en matière de géoréférencement qui permettent de diffuser de nombreux indicateurs à un niveau géographique fin, tout en respectant le secret statistique.

… POUR MIEUX SUIVRE LE PROGRAMME DE TRAVAIL DES PRODUCTEURS

L'ensemble des orientations transversales ou thématiques du Cnis, issues de la concertation au cours de la période précédente, constitue une feuille de route pluriannuelle pour les producteurs de statistiques publiques, qui s'efforcent de la faire avancer pour ce qui les concerne, au cours de la période (Afsa-Essafi, 2019). Le service statistique public dispose à cet égard d'une certaine marge de manœuvre, les priorités exprimées par le Cnis, de nature générale ou méthodologique, étant rarement contraignantes.

> *« Une feuille de route pluriannuelle pour les producteurs de statistiques publiques. »*

Les avis de moyen terme servent en particulier de cadre pour la présentation aux utilisateurs des programmes annuels de travail et des bilans d'activité des producteurs. Les utilisateurs sont ainsi informés régulièrement de la manière dont les producteurs répondent aux recommandations du Cnis.

UN EXAMEN EN AMONT DE CHAQUE NOUVELLE ENQUÊTE …

Parallèlement au recueil de l'ensemble des besoins prospectifs et à la discussion des programmes de travail des producteurs de statistiques publiques, chaque nouvelle enquête de statistique publique doit recueillir un avis du Cnis, dit « avis d'opportunité », en amont de sa mise en œuvre.

Cette première étape en vue d'obtenir l'inscription au programme d'enquête officiel de la statistique publique, vise à s'assurer de la pertinence du projet au regard de ses finalités, de sa place dans le système d'information et des conditions prévues de sa diffusion. Les utilisateurs peuvent réagir sur l'opportunité de chaque opération. Si l'avis d'opportunité est favorable, le projet fait l'objet dans une deuxième étape d'un examen en conformité par le **Comité du label de la statistique publique** (Christine et Roth, 2020) afin de déterminer s'il satisfait aux standards de qualité requis. Au total, de 30 à 40 avis d'opportunité d'enquêtes ont été délivrés annuellement par le Cnis depuis 2017.

L'examen en opportunité d'une nouvelle opération statistique par le Cnis est un jalon important avant de passer à l'instruction technique des projets, qui met les producteurs en situation d'exposer le bien-fondé de leur projet. Si l'opportunité d'ensemble de chaque projet fait rarement débat, des discussions peuvent avoir lieu sur certains aspects du questionnement ou du protocole proposés.

C'est aussi l'occasion, lorsque le projet répond à une demande européenne, de préciser les objectifs poursuivis à ce niveau et leur cohérence avec les besoins des utilisateurs nationaux. Des questions de coordination avec d'autres enquêtes ou d'autres producteurs de la statistique publique peuvent également apparaître. Si la santé des jeunes en milieu scolaire est à l'évidence un sujet partagé entre les services statistiques ministériels Drees[2] et Depp[3], il arrive aussi qu'un autre producteur souhaite collecter une enquête sur un sujet connexe.

2. Service statistique du ministère de la Santé.
3. Service statistique du ministère de l'Éducation.

Ainsi en 2020, l'Observatoire français des drogues et des toxicomanies (OFDT) a présenté en opportunité au Cnis le projet de réédition de l'enquête nationale en collèges et en lycées chez les adolescents sur la santé et les substances (EnCLASS), projet qui devait se positionner par rapport aux opérations existantes des deux SSM.

Enfin, comme l'illustre ce dernier exemple, le périmètre du Cnis est, de fait, un peu plus large que les productions du seul Service statistique public : il porte sur toutes les statistiques publiques qui sont à ce standard de qualité, y compris sur une base volontaire. En effet, plusieurs administrations ou centres de recherche ont pris l'habitude de viser le niveau de qualité de la Statistique publique et de présenter publiquement leurs enquêtes, selon les procédures usuelles d'opportunité et de conformité du SSP. L'Ined[4], le Céreq[5], l'OFDT mènent régulièrement des enquêtes labellisées Statistique publique. D'autres administrations présentent systématiquement leurs programmes de travail statistiques au Cnis, notamment les opérateurs de sécurité sociale, ce qui permet aux utilisateurs d'avoir une vision plus complète de l'offre de statistiques publiques.

… AINSI QUE POUR L'ACCÈS DES STATISTICIENS PUBLICS AUX SOURCES ADMINISTRATIVES…

De manière similaire aux nouvelles enquêtes, la loi de 1951 donne au Cnis un rôle important en amont de l'usage des sources administratives par le SSP. Le Cnis est en effet chargé (article 7bis) d'émettre un avis préalable à chaque demande d'accès à des sources administratives, effectuée par l'Insee ou un service statistique ministériel à des fins de statistiques publiques. La transmission est ensuite rendue obligatoire sur décision du Ministre de l'Économie. Historiquement la transmission au SSP était seulement facultative et ne requérait pas d'avis du Cnis[6]. Celui-ci constate désormais la finalité statistique et l'opportunité de la demande, en amont, ce qui permet d'expliciter le caractère obligatoire. Pour autant, les services producteurs peuvent toujours contractualiser bilatéralement la transmission des données avec les administrations qui les détiennent, sans recourir à une décision ministérielle. Néanmoins, l'utilisation de l'article 7bis et le passage par le Cnis ont deux avantages : d'une part, les données finales deviennent des données statistiques, couvertes par le secret statistique, d'autre part, le public est informé de ces projets de nouvelles sources en amont. La publicité donnée au sein du Cnis à ce mode de construction des statistiques est le pendant de chaque nouveau projet d'enquête. Les utilisateurs insistent souvent sur le besoin pour les chercheurs d'accéder aux micro-données à l'issue de l'opération statistique.

Parfois sensibles car portant sur des données personnelles exhaustives, les demandes d'avis préalables à l'accès du SSP aux bases de données administratives sont en nette augmentation, à l'instar de l'usage de ces sources : 16 avis ont été émis par le Cnis en 2020 au titre de l'article 7 bis de la loi de 1951, après 19 en 2019 et seulement 8 en 2017. En conséquence, au-delà de leur présentation opération par opération, le Cnis a choisi d'augmenter la visibilité des nouvelles utilisations de données administratives à des fins statistiques en créant une rubrique sur son site qui rassemble les « avis 7 bis » donnés.

4. Institut national des études démographiques.
5. Centre d'études et de recherche sur les qualifications.
6. En 2004, dans le cadre d'ordonnances visant à la simplification administrative, la loi a constaté l'importance pour les statisticiens de ces accès, notamment pour réduire la charge d'enquêtes et a introduit ce rôle majeur du Cnis.

… ET DEPUIS QUELQUES ANNÉES, L'ACCÈS AUX DONNÉES PRIVÉES

S'agissant de l'accès du SSP aux données privées, le Cnis est chargé, depuis 2016[7], d'émettre un avis préalable à « *la transmission par voie électronique des informations contenues dans des bases de données détenues par des personnes morales de droit privé* » au SSP.

Cet avis du Cnis ne se substitue pas à la concertation du producteur, porteur du projet, avec les organismes fournissant leurs bases de données, mais il intervient à l'issue de celle-ci pour consulter les utilisateurs de statistiques. L'usage des données privées et, plus généralement, des nouvelles sources numériques est en effet une évolution naturelle, à laquelle la Statistique publique avait en 2016 consacré un groupe de travail (Bozio *et alii*, 2017) et dont une « Rencontre Cnis » a mis en avant en 2018 l'intérêt et les limites (Elbaum, 2018).

Récemment, l'usage par le service statistique public des sources *big data* s'est accéléré à l'occasion de la crise sanitaire de 2020, l'avantage de fraîcheur et de fréquence que celles-ci apportent étant décisif par rapport aux sources traditionnelles (Cnis, 2021b).

En pratique, toutefois, le Service statistique public n'a eu recours à cette disposition législative que pour l'utilisation, par l'Insee, de données de caisse de la grande distribution en vue du calcul d'une partie de l'indice des prix à la consommation[8]. Dans son avis sur cette transmission, le Cnis suggérait notamment que la précision accrue des nouvelles productions soit mise à profit pour augmenter le détail des statistiques diffusées. Un nouvel usage des mêmes données de caisse pourrait voir le jour de manière imminente pour améliorer les indicateurs d'activité de court terme dans le secteur du commerce.

UNE EXPERTISE PARTAGÉE SUR DES SUJETS D'INTÉRÊT POUR LE DÉBAT PUBLIC

Si les productions du Cnis évoquées jusqu'ici s'expriment sous forme d'avis ou de recommandations de portée générale, elles ne constituent pas cependant son seul mode d'expression. En effet, le Cnis est aussi le lieu où se développent des expertises partagées entre producteurs et utilisateurs sur des sujets émergents, lorsqu'ils présentent un intérêt particulier pour le débat public.

Les groupes de travail du Cnis sont présidés par une personnalité extérieure au SSP et rassemblent des experts du domaine, quelle que soit leur affiliation. Leur rapport final, qui établit un consensus entre producteurs et utilisateurs sur les besoins statistiques et la manière d'y répondre, fait en général référence, parfois même en dehors de la sphère statistique (par exemple, *La mesure du travail dissimulé* (Gubian, Hagneré et Mahieu, 2017) a eu des suites au sein du Haut Conseil du financement de la protection sociale).

7. Avec la création d'un nouvel article 3 bis dans la loi de 1951.
8. Pour plus d'information, voir (Leclair, 2019).

La liste des rapports de groupes de travail du Cnis illustre la variété et la richesse de ces travaux, qu'il s'agisse :

- de faire l'état des lieux des besoins au regard du système d'information statistique existant dans un domaine thématique large (*L'information statistique sur le logement et la construction* (Vorms, Jacquot et Lhéritier, 2010)) ;
- de définir les indicateurs à privilégier pour le débat public sur les sujets économiques et sociaux : (*La déclinaison française des indicateurs de suivi des objectifs de développement durable* (Brunetière *et alii*, 2018)) ;
- d'investir des domaines émergents pour les statistiques (*La diversité des formes d'emploi* (Gazier, Minni et Picart, 2016)) ;
- ou encore de réaliser une concertation approfondie des experts, utilisateurs et parties prenantes en amont de la mise en place d'une nouvelle nomenclature afin de s'assurer qu'elle répond aux différents usages statistiques (*Rénovation de la nomenclature des professions et catégories socio-professionnelles* (Amossé, Chardon et Eidelman, 2019)), pour n'en citer que quelques-uns.

Ces dernières années, les groupes de travail du Cnis ont aussi été souvent orientés vers la diffusion des informations et leur accessibilité (*Accès des chercheurs aux données administratives* (Bozio *et alii*, 2017)). Les progrès importants réalisés dans l'accès des chercheurs aux données sont en partie le fruit de cette pression discrète de la demande, exprimée aussi dans les orientations générales des précédents moyens termes.

> *« Dans le domaine considéré, les recommandations des groupes de travail constituent une feuille de route précise adressée aux producteurs pour améliorer la connaissance. »*

Dans le domaine considéré, les recommandations des groupes de travail constituent une feuille de route précise adressée aux producteurs pour améliorer la connaissance, leur suivi est assuré au sein du Cnis. Ainsi, le groupe de travail sur l'*Observation des ruptures familiales* (Thélot, Bourreau-Dubois et Chambaz, 2016), a été constitué pour apporter des réponses au constat d'insatisfaction de certaines instances en charge de la politique familiale, concernant la prise en compte dans les statistiques de la situation des enfants de foyers éclatés. Le groupe a produit trente recommandations adressées à l'Insee et aux services statistiques ministériels concernés. Lors du suivi de sa mise en œuvre réalisé en 2018, le Cnis a pu constater les avancées importantes déjà réalisées à la suite du rapport, avec notamment une prise en compte des recommandations dans le questionnaire de l'enquête annuelle de recensement et de l'enquête Emploi de l'Insee, ainsi par différents travaux d'études de la Drees, la Depp ou l'Insee (Buisson et Reynaud, 2019).

DES RÉFLEXIONS TRANSVERSALES SUR LES SUJETS À ENJEUX

Les colloques et séminaires organisés par le Cnis sont l'occasion de poser un diagnostic partagé sur des thèmes majeurs et d'identifier les principaux enjeux pour les statistiques. Le colloque de 2018 a permis par exemple de mettre en avant les enjeux transverses (comptabilité nationale, statistiques d'entreprises ou sociales) de la transformation numérique de l'économie pour les statistiques, avec des exemples concrets d'innovations des producteurs de la statistique publique, des bonnes pratiques des instituts statistiques étrangers mais aussi l'approche *big data* de certains utilisateurs, l'ensemble constituant

une incitation collective à poursuivre ces travaux de manière coordonnée (Tagnani, 2018). Les travaux peuvent aussi être plus ciblés, comme ceux du séminaire sur le recensement de 2020, qui a permis de poser les bases des prochaines évolutions du questionnaire ou de sa diffusion.

UN LIEU D'INFORMATION ET DE TRANSPARENCE TOURNÉ VERS LES UTILISATEURS

Le rôle consultatif du Cnis se traduit par des avis adressés aux producteurs statistiques, qui ont vocation à être publics pour en renforcer l'efficacité. Le Cnis a aussi toujours pratiqué des réunions ouvertes pour ses commissions thématiques, afin d'enrichir ses travaux de toutes les contributions et de diffuser l'information rassemblée. Sans que ce rôle lui soit explicitement confié par les textes, le Cnis pratique de longue date une politique de transparence complète, gage à la fois d'efficacité de la concertation et de qualité de l'information elle-même. Aujourd'hui, la transparence assurée par le Cnis *via* son site internet va bien au-delà de sa production propre : programmes de travail annuels du SSP et bilan de ceux-ci, description des projets d'enquêtes ou d'opérations statistiques soumis à un avis ponctuel du Cnis (cf. *supra*), mais aussi dossiers préparatoires, supports de présentation de séance et comptes-rendus de réunions Cnis sont diffusés en ligne de manière organisée. Cette transparence s'applique aussi, au moins en partie, aux réunions « fermées » du Cnis, i.e. de la Commission nationale d'évaluation du recensement (Cnerp), du Bureau ou du Conseil (voir *infra* pour la présentation de ces instances). Dans un souci de pédagogie et d'accessibilité, le Cnis diffuse aussi des synthèses des réunions et publie des 4-pages (les « Chroniques » du Cnis) et il a modernisé fortement son site *web* en 2017 pour mieux orienter les utilisateurs. Celui-ci permet de restituer les débats au public et de laisser une trace permettant à tous de suivre l'avancement des avis rendus.

> *« Sans que ce rôle lui soit explicitement confié par les textes, le Cnis pratique de longue date une politique de transparence complète, gage à la fois d'efficacité de la concertation et de qualité de l'information elle-même. »*

Le Cnis référence aussi exhaustivement toutes les enquêtes de statistique publique vues en opportunité et rassemble de manière standardisée un certain nombre de métadonnées sur chacune, telles que définies en amont de l'opération[9], ainsi que l'avis d'opportunité donné par le Cnis et l'avis de conformité du Comité du label de la statistique publique. Il s'efforce depuis peu de rendre visible aussi les demandes d'accès aux sources administratives du SSP (cf. *supra*). Par l'information qu'il fournit sur l'ensemble des projets de statistiques publiques, le Cnis contribue donc au principe de transparence de celles-ci, selon le code des bonnes pratiques de la statistique européenne. Par ailleurs, cette transparence ne s'adresse pas uniquement aux utilisateurs des données statistiques en aval, mais aussi aux parties prenantes du processus statistique que sont les personnes ou organismes enquêtés, contribuant ainsi à augmenter leur confiance.

Enfin le Cnis rend compte annuellement de ses activités, sous différentes formes, notamment auprès de **l'Autorité de la statistique publique** (un bilan d'activité du Cnis est annexé au rapport de l'ASP, qui est diffusé très largement).

9. Le secrétariat général du Cnis assure l'harmonisation de ces métadonnées avec celles du référentiel de métadonnées statistiques tenu par l'Insee (RMéS, voir (Bonnans, 2019)).

UNE ORGANISATION EFFICACE ET RODÉE

La qualité de la concertation et du recueil des besoins s'appuie sur une organisation efficace et rodée, qui repose sur le principe d'une représentation et d'une expression la plus large possible des utilisateurs de la statistique publique. Elle s'articule autour de **commissions thématiques**, lieux de rencontre ouverts à tous, dont les travaux sont analysés et consolidés au sein du **Bureau** et validés par le **Conseil** dans son ensemble, réuni annuellement en assemblée plénière ***(figure 1).***

Cette organisation pyramidale, qui permet une consolidation progressive des travaux du Cnis, assure au final l'intégration des problématiques émergentes ou faisant débat, dès leur expression par les utilisateurs au niveau le plus large, jusqu'à leur traduction opérationnelle dans les programmes de travail de la statistique publique.

SEPT COMMISSIONS THÉMATIQUES COMME BASE DE LA CONCERTATION…

C'est au sein de sept commissions thématiques, ouvertes à tous et qui se réunissent chacune deux fois par an, que s'opère la base de la concertation et du recueil des besoins, autour de :

- trois thèmes à dimension sociale, (« Démographie et questions sociales », « Emploi, qualification et revenus du travail », « Services publics et services aux publics ») ;
- deux thèmes économiques (« Entreprises et stratégie de marché », « Système financier et financement de l'économie ») ;
- et deux thèmes transversaux, « Environnement et développement durable » et « Territoires ».

Pour assurer un équilibre dans la concertation entre les utilisateurs et les producteurs de statistiques publiques, chaque commission est présidée par une personnalité extérieure au Service statistique public et faisant référence sur le thème de la commission, assistée de deux rapporteurs, membres du SSP et exerçant des responsabilités dans le domaine.

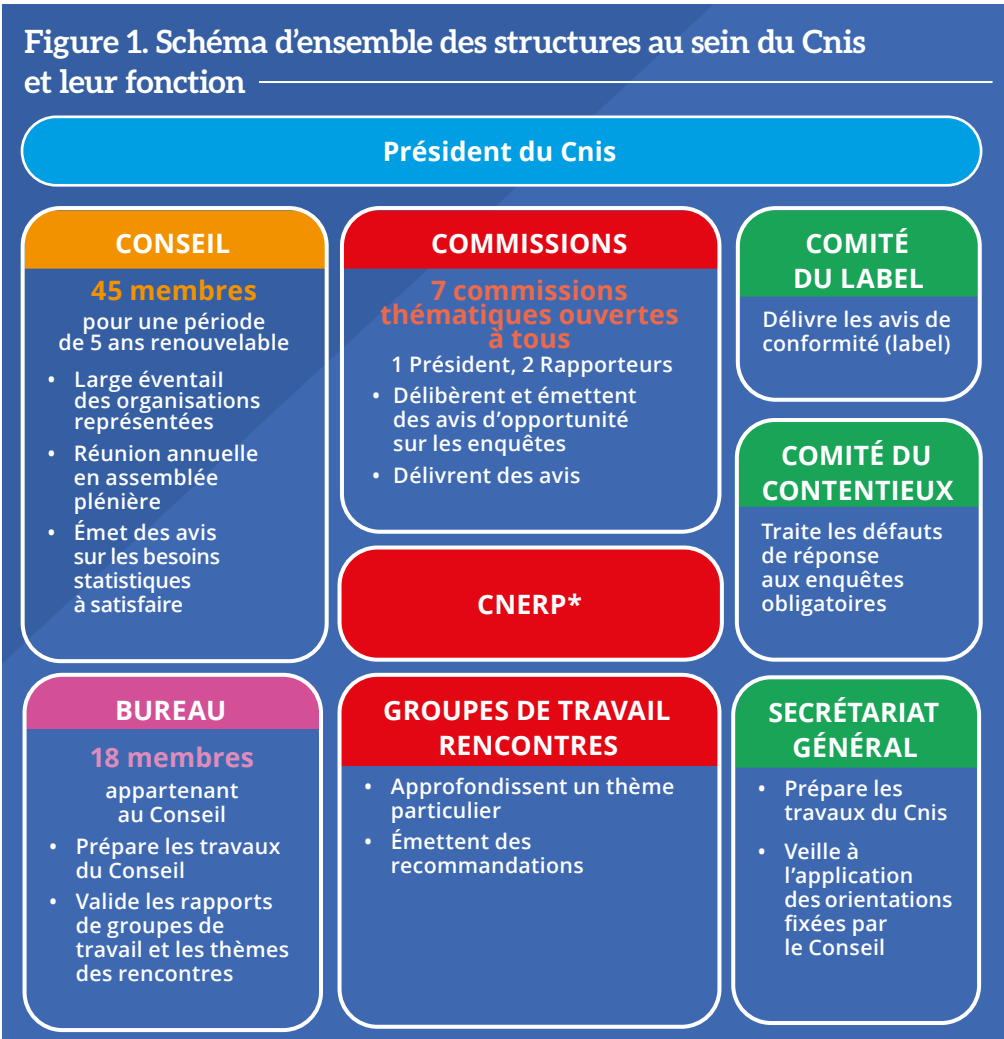
Le dispositif des commissions thématiques est complété par une huitième commission spécifique, la Commission nationale d'évaluation du recensement de la population (Cnerp), qui traite des modalités de collecte du recensement de la population.

Les réunions des commissions thématiques sont structurées à partir d'un ordre du jour assez stable qui comprend, outre l'examen des demandes d'avis d'opportunité et d'accès aux demandes administratives dans le cadre de l'article 7bis, un sujet central qui se conclut par un avis de la commission et, en fonction de l'actualité, un (ou des) point (s) d'information sur l'avancement des travaux des producteurs pouvant également, le cas échéant, donner lieu à un avis de la commission.

Le sujet central est très souvent organisé en trois temps. Les producteurs de la statistique publique dressent un état des lieux sur les statistiques disponibles et les dispositifs à partir desquels elles sont élaborées. Des utilisateurs (associations ou administrations) présentent des usages de ces données en faisant ressortir les améliorations souhaitées ou les manques éventuels. Enfin, le point de vue des chercheurs met ces données en perspective et constitue parfois une introduction au débat. À l'issue des discussions, un avis reprend les termes principaux des échanges et consigne les demandes d'amélioration.

À titre d'exemple, le sujet principal à l'ordre du jour de la commission Emploi du 5 novembre 2020 qui s'intitulait « *Améliorer la connaissance des tensions sur le marché du travail* » comportait ainsi ces trois volets. L'état des lieux a été dressé par les producteurs à partir de quatre interventions : introduction Insee/Dares sur les sources statistiques disponibles pour appréhender les tensions sur le marché du travail, les nouveaux indicateurs créés (Dares), l'apport de l'enquête besoin de main-d'œuvre dans la compréhension des tensions (Pôle Emploi), le lien entre chômage et pénurie de main-d'œuvre (Insee). Le point de vue des utilisateurs a été apporté par France Stratégie à partir d'une présentation de l'exercice « *Prospective des métiers* » et par une organisation professionnelle[10]. Un chercheur a complété les approches et introduit les échanges.

Figure 1. Schéma d'ensemble des structures au sein du Cnis et leur fonction



10. L'Union des Industries et Métiers de la Métallurgie.

Cette réunion a regroupé 70 participants, dont 38 n'appartenant pas au SSP, et parmi eux des associations et organisations professionnelles[11], des administrations nationales[12], des administrations régionales[13], des organisations syndicales[14], des organismes de recherche[15].

Plus généralement, les commissions regroupent chacune de 40 à 100 participants, invités de manière ciblée, à partir d'un fichier de 4 000 contacts. En 2019 et 2020 (au deuxième semestre uniquement), les commissions ont rassemblé 1 066 participants dont 52 % n'appartenant pas au SSP. Cette participation n'a pas été affectée par les réunions en distanciel mises en place en raison de la crise sanitaire de 2020 ***(encadré 1)***.

… ET QUI MÉRITENT PARFOIS D'ÊTRE ÉLARGIES

Le partitionnement des sujets d'intérêt, qui résulte de cette organisation thématique, nécessite parfois d'être aménagé, notamment pour aborder les interactions entre des thèmes relevant de deux commissions différentes. C'est alors au sein d'une inter-commission que la concertation prend corps.

Ainsi, en avril 2017, une inter-commission *Services publics et services aux publics / Environnement et développement durable*, après avoir dressé un panorama des statistiques disponibles et leur exploitation en matière de santé-environnement, a mis en évidence la difficulté à relier les données environnementales aux pathologies existantes, d'une part du fait de l'insuffisance de données pour quantifier ou caractériser la présence et la teneur de certaines sources de nuisance ou de pollution au niveau territorial approprié et d'autre part, à cause de la difficulté à quantifier l'impact direct des facteurs environnementaux sur l'état de santé de la population. D'une manière analogue, en mai 2018, les sources statistiques sur l'emploi des personnes handicapées ont été traitées lors d'une inter-commission *Emploi, qualification et revenus du travail / Services publics et services aux publics*.

> *« Certains des thèmes abordés lors des commissions thématiques permettent en outre d'élargir la concertation au-delà des commissions en créant des synergies avec d'autres instances de concertation et d'animation du débat public, et notamment les hauts conseils. »*

Certains des thèmes abordés lors des commissions thématiques permettent en outre d'élargir la concertation au-delà des commissions en créant des synergies avec d'autres instances de concertation et d'animation du débat public, et notamment les hauts conseils. Pour traiter de l'enfant dans la statistique publique, la commission *Démographie et questions sociales* a accueilli la présidente du Haut conseil de la famille, de l'enfance et de l'âge (HCFEA) parallèlement à une présentation des sources de statistiques publiques relatives à l'enfant au sein de ce Haut Conseil. Un avis de chaque instance a été émis, allant dans le même sens d'une définition de l'enfant lisible et partagée, de manière à en faciliter l'approche globale.

11. Fédération nationale des syndicats d'exploitants agricoles, Solidarités nouvelles face au chômage, Fédération des particuliers employeurs de France, Agences d'urbanismes, Carif-Oref (Centre animation ressources d'information sur la formation / Observatoire régional emploi formation).
12. Agence Nationale de la Cohésion des Territoires.
13. Région Grand-Est.
14. CGT, CGT-FO.
15. Institut de recherches économiques et sociales, Centre d'études et de recherches sur les qualifications, Université Paris Sorbonne.

DES TRAVAUX CONSOLIDÉS PAR LE BUREAU ET LE CONSEIL

L'activité des commissions thématiques et notamment les avis émis à l'issue de chaque séance sont examinés par le **Bureau du Cnis**, qui les complète par des orientations générales mettant en exergue les attentes de portée transversale pour l'année à venir, préparant ainsi les délibérations de l'Assemblée plénière. Autour du président du Cnis, le bureau comprend 18 membres appartenant au Conseil et notamment le Directeur général de l'Insee, un représentant de la Banque de France, un représentant de France Stratégie, un représentant de chacune des organisations syndicales, professionnelles et consulaires représentées au Cnis, un représentant des collectivités locales, un représentant des chercheurs, un représentant de la Fédération française de l'assurance et une personnalité qualifiée.

C'est également au sein du bureau, qui se réunit quatre fois par an, que sont validées les propositions de groupes de travail formulées par les commissions au cours de leurs réunions, ainsi que les thèmes des colloques et séminaires et que sont présentés et examinés les rapports des groupes de travail.

Enfin, au terme d'une année d'échanges, de concertation et de recueil de besoins, l'ensemble des avis émis par les commissions et des orientations formulées par le bureau est validé par **le Conseil**, réuni en **Assemblée plénière**. Renouvelé en 2019, ses 45 membres, nommés par arrêté ministériel pour une période de cinq ans renouvelable, forment un large éventail de la société civile par les organisations qui y sont représentées : élus nationaux et locaux, syndicats professionnels, syndicats de salariés, associations, organismes publics, universitaires et personnalités qualifiées.

La mise en œuvre de cette organisation est assurée par un Secrétariat général mis à disposition par l'Insee, qui prépare les commissions et les travaux du Bureau et du Conseil, accompagne les groupes de travail, organise les colloques et les séminaires et veille à l'application des orientations fixées par le Conseil dans le cadre du Moyen terme.

S'ADAPTER EN CONTINU AUX ÉVOLUTIONS DE SON ENVIRONNEMENT...

Si l'organisation du Cnis a assez peu changé dans son principe général et ses structures depuis plusieurs décennies, cette stabilité n'est pas synonyme d'immobilité. Historiquement, le Cnis a toujours accompagné l'évolution de son environnement, qu'il s'agisse des usages ou du cadre général de la production des statistiques publiques ***(encadré 2)***.

Tout d'abord, le rôle central de coordination des producteurs joué par le Cnis à sa création s'est progressivement effacé au profit de la concertation avec la société civile, corrélativement à la montée en puissance de la coordination et des collaborations internes au SSP.

La concertation elle-même a évolué en s'orientant désormais principalement vers les utilisateurs, relativement à d'autres parties prenantes des statistiques. La maîtrise de la charge d'enquête, qui a donné lieu à de nombreux débats au début des années 2000, est ainsi passée au second plan, le volume d'enquête ne croissant plus aujourd'hui. Les organisations professionnelles, qui étaient au sein du Cnis à la fois porte-paroles des entreprises, des intermédiaires de collecte de nombreuses enquêtes, et des utilisateurs, représentent aujourd'hui davantage des utilisateurs.

Encadré 1. Les sujets traités par les commissions dans le moyen terme en cours

DOMAINE SOCIAL

La ***Commission Services publics, services aux publics*** s'est intéressée à la mesure du non recours aux droits sociaux, à l'ouverture des données pénales, à l'appariement des données, aux avancées du suivi statistique des retraites et aux méthodes statistiques du suivi de la délinquance.

La ***Commission Démographie et questions sociales*** a traité de l'enfant dans la statistique publique, des nouvelles sources de données et des utilisations innovantes dans le domaine de la santé et de la consommation (*Health Data Hub**, SNDS**, Données de caisse***), du suivi statistique du grand âge et des méthodes d'observation de la grande pauvreté.

Les travaux de la ***Commission Emploi, qualification et revenus du travail*** ont porté sur le sentiment d'insécurité sur le marché du travail, la rénovation de la nomenclature des Professions et catégories sociales (PCS 2020), les tensions du marché du travail, le suivi des trajectoires professionnelles des indépendants.

DOMAINE ÉCONOMIQUE ET FINANCIER

La ***Commission Entreprises et stratégies de marché*** a abordé le dispositif de suivi de l'activité internationale des groupes, les conclusions du groupe de travail sur la diffusion des statistiques structurelles d'entreprises, la consommation d'énergie des entreprises, la mesure de l'empreinte carbone du système productif et la mesure de la R&D et de l'innovation dans les entreprises.

Les thèmes traités par la ***Commission Système Financier et financement de l'économie*** ont été la place des sociétés de gestion dans le financement de l'économie française, les marchés financiers et la transition écologique, l'accès aux données granulaires financières, les enseignements de la crise sanitaire en matière de statistique publique dans le domaine financier.

DOMAINES TRANSVERSES

Les deux commissions transversales ont respectivement axé leurs travaux, pour la ***Commission Environnement et développement durable*** sur le gaspillage alimentaire, les statistiques relatives aux émissions de gaz à effet de serre et l'évaluation du coût des catastrophes naturelles, l'utilisation des données individuelles pour la transition écologique…

… et pour la ***Commission Territoires*** sur la mesure de l'accès aux équipements et aux services, la délimitation des territoires, les données et l'utilisation du prix du foncier et de l'immobilier, l'information disponible sur l'Outre-mer.

Exemples tirés des sujets traités entre 2019 et 2021, en lien avec le moyen terme 2019-2023, voir (Cnis, 2019).
** Health Data Hub (HDH) : plateforme d'accès aux données de santé, pour favoriser la recherche.*
*** SNDS : système national des données de santé.*
**** Données servant au calcul de l'indice des prix à la consommation.*

Les utilisateurs eux-mêmes ont évolué vers une expertise croissante de l'exploitation de données, ce qui a conduit le Cnis à chercher à faciliter l'accès des chercheurs aux données statistiques individuelles.

La place croissante de la demande européenne dans les programmes statistiques, fait marquant des trois dernières décennies, a amené le Cnis à informer régulièrement les utilisateurs nationaux des projets européens et, depuis l'instauration de règlements-cadres par grand domaine[16], à prévoir la rédaction d'avis du Cnis suffisamment en amont de l'élaboration des projets d'opérations européennes afin qu'ils puissent être transmis aux instances européennes.

« Les utilisateurs eux-mêmes ont évolué vers une expertise croissante de l'exploitation de données, ce qui a conduit le Cnis à chercher à faciliter l'accès des chercheurs aux données statistiques individuelles. »

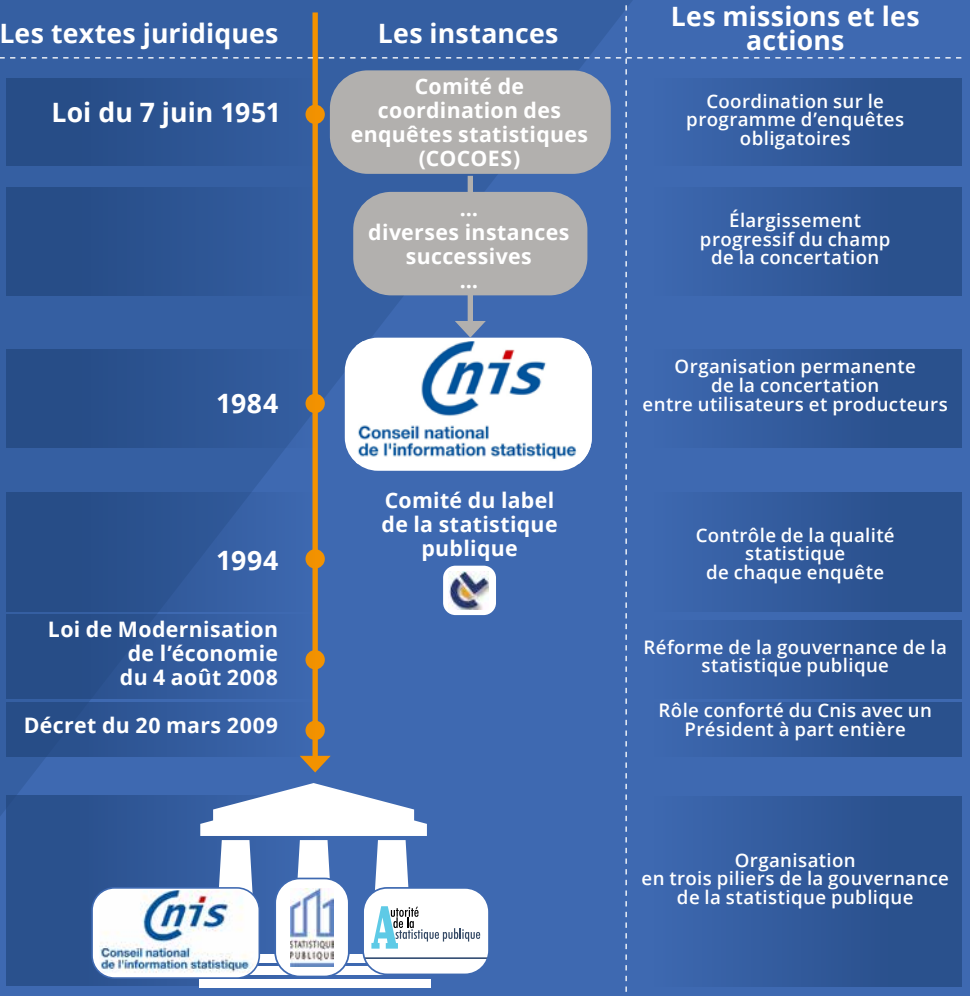
Face à l'évolution des modes de construction des statistiques publiques, en particulier la croissance de l'utilisation des données administratives ainsi que d'autres modes de collecte alternatifs aux enquêtes (données privées, *web scraping*, etc.), le Cnis a élargi de fait le périmètre de ses échanges à l'ensemble des systèmes d'information statistiques, quel que soit leur mode d'élaboration, afin que les utilisateurs en aient une vue aussi complète que possible. Plus généralement, face à la multiplication des productions de données extérieures au Service statistique public (d'associations, de plateformes, destinées à du rapportage non statistique, etc.), ces producteurs peuvent être invités à présenter leurs statistiques, notamment sur des domaines non couverts par la Statistique publique afin d'examiner les voies de progrès réciproque. Dans cet esprit, il a aussi été décidé que le Cnis jouerait un rôle dans l'homologation des « Statistiques d'intérêt général »[17] issues de producteurs extérieurs à la statistique publique (Cnis, 2020b). Enfin, le Cnis sera amené également à documenter les usages du nouveau « code statistique non signifiant » par le SSP, de nature à faciliter les appariements de sources et il pourra à terme émettre des avis pour encadrer les bonnes pratiques en ce domaine.

Le Cnis s'est aussi positionné vis-à-vis des grandes tendances technologiques, dès lors qu'elles avaient des enjeux pour les parties prenantes des statistiques. L'usage des données administratives a ainsi été recommandé pour réduire la charge d'enquête mais aussi dans la perspective que les utilisateurs bénéficient du niveau de détail élevé qu'elles permettent de diffuser. Le Cnis a aussi préconisé en son temps le passage à des enquêtes en ligne pour rendre la collecte plus rapide et plus efficace, ainsi que la généralisation de la diffusion sur internet pour faciliter l'accès des utilisateurs aux travaux statistiques.

Sur la période récente, marquée par la crise sanitaire de la Covid-19, la réactivité des travaux du Cnis a accompagné celle de la statistique publique. En 2020, le Cnis a mis en place de nouveaux modes de consultation « virtuels » sur les avis d'opportunité et sur certains aspects du programme de travail à venir, notamment pour les projets du SSP en rapport avec la crise, sans sacrifier la qualité de la concertation. Ceux-ci ont vocation à être pérennisés, au moins en partie ***(encadré 2)***.

16. Pour plus de détail sur les règlements-cadres européens voir (Cases, 2019), (Colin, 2019) et (Piffeteau, 2019).
17. Ce projet fait suite au rapport éponyme de l'Inspection générale de l'Insee (2019).

Encadré 2. Quelques repères historiques de l'évolution du Cnis



Historiquement constitué en 1951, le Cnis s'est transformé progressivement en une instance de concertation permanente et transverse, représentant l'ensemble de la société civile et couvrant l'ensemble des services producteurs, consultée systématiquement sur les projets d'enquêtes.

Le dernier changement important dans les missions et l'organisation du Cnis s'est effectué à l'occasion de la refonte de la gouvernance statistique de 2009, qui a vu la création de l'ASP, autorité indépendante destinée à contrôler le respect de l'indépendance professionnelle de la statistique publique, ainsi que les autres principes du code de bonnes pratiques.

À cette occasion, la gouvernance du Cnis a été clarifiée, notamment :

- en la resserrant autour d'un nombre de membres plus faible ;
- en réduisant le nombre de commissions thématiques de 13 à 7 ;
- et en nommant un président de plein exercice, alors que le ministre de l'Économie était autrefois le président en titre du Conseil, dont l'animation était confiée au Vice-président.

L'articulation entre le Cnis, le Comité du label et le Comité du secret statistique a également été revue à cette occasion.

… ET CONTRIBUER À LA CONFIANCE DU PUBLIC

Instance transversale à tous les domaines statistiques et à tous les types d'utilisateurs, le Cnis assure la bonne information de ces derniers et leur permet d'exprimer des besoins en amont ou de réagir sur les projets statistiques. Il est l'interface légitime et pérenne d'une concertation transparente à l'échelle de toute la statistique publique, entre les producteurs et les utilisateurs de statistique. Lieu d'un débat dans la transparence sur des sujets ponctuels comme sur les orientations statistiques générales, il permet à la statistique publique de s'assurer de la pertinence de ses productions. Il contribue ainsi à augmenter la confiance dans les statistiques et leur acceptabilité sociale. Les débats du Cnis ont également des vertus de pédagogie, et parfois même de modération du débat public sur des questions « sensibles », comme celles de l'emploi, des inégalités, des discriminations, etc.

Le système d'information statistique étant largement arrivé à maturité, il s'agit moins aujourd'hui pour le Cnis d'explorer des domaines thématiques vierges que de pousser au développement de nouvelles méthodes de production, à leur rationalisation ou à une meilleure articulation entre sources existantes.

Le Cnis souhaite ainsi pouvoir donner plus de visibilité aux statistiques fondées sur des données administratives, pour en augmenter la transparence au même niveau que les sources fondées sur des enquêtes, notamment vis-à-vis de l'utilisation des données personnelles. À terme, il devra également pouvoir offrir une vision aussi exhaustive que possible de l'ensemble des sources de données du SSP, en cartographiant les systèmes d'information. Enfin, pour améliorer la concertation au cours des années à venir, le Cnis devra s'articuler efficacement avec les retours d'utilisateurs de plus en plus nombreux, recueillis directement par les producteurs ou les diffuseurs ou par d'autres composantes de la Statistique publique (le Comité du secret statistique par exemple pour les chercheurs), pour en rendre compte dans ses travaux.

BIBLIOGRAPHIE

AFSA-ESSAFI, Cédric, 2019. *2023, nouvel horizon du Cnis*. [en ligne]. Mars 2019. Chroniques du Cnis n° 17. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/03/Chroniques17_nouvel_horizon_Cnis_2023.pdf*.

AMOSSÉ, Thomas, CHARDON, Olivier et EIDELMAN, Alexis, 2019. *La rénovation de la nomenclature socio-professionnelle (2018-2019)*. [en ligne]. Décembre 2019. Cnis, rapport de groupe de travail, n° 156. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/01/Rapport-n%C2%B0-156.pdf*.

ARCHAMBAULT, Édith, ACCARDO, Jérôme et LAOUISSET, Brahim, 2010. *Connaissance des associations*. Décembre 2010. [en ligne]. Cnis, rapport de groupe de travail, n° 122. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2010_122_connaissance_associations.pdf*.

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 26 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf*.

BOZIO, Antoine, GEOFFARD, Pierre-Yves, BREUIL, Pascale, PERRIÈRE, Manon et MALVERTI, Clément, 2017. *L'accès des chercheurs aux données administratives. État des lieux et propositions d'actions.* [en ligne]. Mars 2017. Cnis, rapport de groupe de travail, n° 147. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2019/07/rapportcnis147completweb.pdf*.

BRUNETIÈRE, Jean-René, MESQUI, Bérengère, MORARD, Valéry, MOREAU, Delphine, EGHBAL-TÉHÉRANI, Sylvie et VEY, Frédéric, 2018. *La déclinaison française des indicateurs de suivi des objectifs de développement durable.* [en ligne]. Juin 2018. Cnis, rapport de groupe de travail, n° 150. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2019/03/Rapport_Cnis_N%C2%B0150_GT_iODD.pdf*.

BUISSON, Guillemette et REYNAUD, Émilie, 2019. *Nous nous sommes tant aimés : les ruptures familiales et la statistique. Bilan à trois ans du suivi des recommandations du groupe de travail, du Cnis.* [en ligne]. Septembre 2019. Chroniques du Cnis, n° 20. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/03/Chronique-20_ruptures_familiales.pdf*.

BUREAU, Dominique, 2020. L'Autorité de la statistique publique. Dix ans d'activité, pour une statistique indépendante et de qualité. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 21-38. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008710/courstat-5.pdf*.

CASES, Chantal, 2019. IESS : l'Europe harmonise ses statistiques sociales pour mieux éclairer les politiques. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 125-139. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254233/courstat-3-10.pdf*.

CHRISTINE, Marc et ROTH, Nicole, 2020. Le Comité du Label. Un acteur de la gouvernance au service de la qualité des statistiques publiques. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 39-52. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008698/courstat-5-4.pdf*.

CNIS, 2004. *Avis du Conseil national de l'information statistique sur le programme statistique à moyen terme 2004-2008 et sur sa première année d'exécution. Assemblée plénière du 18 décembre 2003.* [en ligne]. Février 2004. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/11/RAP_2004_84_moyen_terme_2004_2008.pdf*.

CNIS, 2009. *Avis moyen terme 2009-2013 et avis 2009 première année d'exécution du Conseil national de l'information statistique. Assemblée plénière du 23 janvier 2009.* [en ligne]. 12 mars 2009. N° 86/D130.[Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/11/RAP_2009_115_avis_moyen_terme_2009_2013.pdf*.

CNIS, 2019. *Avis du moyen terme 2019-2023 du Cnis adoptés par l'Assemblée plénière du 31 janvier 2019.* [en ligne]. Avis n° 154. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2019/05/Rapport_Cnis_n%C2%B0154.pdf*.

CNIS, 2020a. *Avis du Conseil national de l'information statistique sur les programmes statistiques 2020. Assemblée plénière du 4 février 2020.* [en ligne]. Avis n° 158. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/08/Avis-sur-les-programmes-statistiques-2020.pdf*.

CNIS, 2020b. Bureau – 2020 – 2e réunion. In : *site du Cnis*. [en ligne]. 18 juin 2020. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/evenements/bureau-2020-2e-reunion/?category=994*.

CNIS, 2020c. Bureau – 2020 – 4e réunion. In : *site du Cnis*. [en ligne]. 9 décembre 2020. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/evenements/bureau-2020-4e-reunion/?category=994*.

CNIS, 2021a. Rubrique « Qui sommes-nous ? ». In : *site du Cnis.* [en ligne]. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/cnis/*.

CNIS, 2021b. *Avis du Conseil national de l'information statistique sur les programmes statistiques 2021. Assemblée plénière du 27 janvier 2021.* [en ligne]. Février 2021. Avis n° 159. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2021/01/Avis_Programmes_Statistiques_2021_Adopt%C3%A9s.pdf*.

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 26 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254231/courstat-3-9.pdf*.

DE FOUCAULD, Jean-Baptiste, CÉZARD, Michel et REYNAUD, Marie, 2008. *Emploi, chômage, précarité. Mieux mesurer pour mieux débattre et mieux agir.* [en ligne]. Septembre 2008. Cnis, rapport de groupe de travail, n° 108. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2008_108_emploi_chomage_precarite.pdf*.

DUCHÂTEAU, Alain et COLIN, Christel, 2019. *La nouvelle diffusion des statistiques structurelles d'entreprises.* [en ligne]. Octobre 2019. Cnis, rapport de groupe de travail, n° 157. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/01/Rapport-n%C2%B0-157.pdf*.

DUPORT, Jean-Pierre, 2009. Le Conseil national de l'information statistique, In : *Courrier des statistiques*. [en ligne]. Septembre-décembre 2009. n° 128, pp. 9-13. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/122489/1/cs128.pdf*.

ELBAUM, Mireille, 2018. *Les enjeux des nouvelles sources de données.* [en ligne]. Septembre 2018. Chroniques du Cnis, n° 16. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/03/Chronique-16_enjeux_nouvelles_sources_donn%C3%A9es.pdf.*

FREYSSINET, Jacques, CHEVALIER, Pascal et DOLLÉ, Michel, 2007. *Niveaux de vie et inégalités sociales.* [en ligne]. Mars 2007. Cnis, rapport de groupe de travail, n° 103. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2007_103_niveaux_de_vie_inegalites_sociales.pdf*.

GAZIER, Bernard, MINNI, Claude et PICART, Claude, 2016. *La diversité des formes d'emploi.* [en ligne]. Juillet 2016. Cnis, rapport de groupe de travail, n° 142. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2016_142_diversite_forme-demploi.pdf*.

GEOFFARD, Pierre-Yves, 2019. *Dix ans de compagnonnage avec le Cnis : quelques leçons.* [en ligne]. Avril 2019. Chroniques du Cnis, n°18. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/03/Chroniques-18_dix_ans_compagnonnage_Cnis.pdf*.

GREGOIR, Stephan et DUPONT, Françoise, 2016. *La réutilisation par le système statistique public des informations des entreprises.* [en ligne]. Mars 2016. Insee – Cnis, rapport de groupe de travail, n° 143. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2016_143_reutilisation_syst_stat_information_ets.pdf*.

GUBIAN, Alain, HAGNERÉ Cyrille et MAHIEU, Ronan, 2017. *La mesure du travail dissimulé et ses impacts pour les finances publiques*. [en ligne]. Juin 2017. Cnis, rapport de groupe de travail, n° 145. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/11/Rapport_145web.pdf*.

LECLAIR, Marie, 2019. Utiliser les données de caisses pour le calcul de l'indice des prix à la consommation. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 61-75. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254225/courstat-3-6.pdf*.

PIFFETEAU, Hervé, 2019. Un nouveau triptyque juridique pour les statistiques européennes. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 93-109. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254229/courstat-3-8.pdf.*

SALUSTRO, Édouard, MÉNARD, Claude et DEPOUTOT, Raoul, 2008. *Statistiques Structurelles fondées sur les groupes d'entreprises et leurs sous-groupes*. [en ligne]. Janvier 2008. Cnis, rapport de groupe de travail. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2008_107_statistiques_structurelles.pdf*.

TAGNANI, Stéphane, 2018. *L'économie numérique : enjeux pour la statistique publique. Synthèse du Colloque du 7 mars 2018.* [en ligne]. Juillet 2018. Chroniques du Cnis, n° 15 [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/03/Chronique15_%C3%A9conomie_num%C3%A9rique_enjeux_statistique_publique.pdf*.

THÉLOT, Claude, BOURREAU-DUBOIS, Cécile, CHAMBAZ, Christine, 2016. *Les ruptures familiales et leurs conséquences: 30 recommandations pour en améliorer la connaissance*. [en ligne]. Mars 2016. Cnis, rapport de groupe de travail. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAPPORT-RUPTURES-FAMILIALES-_-nouvelle-version-29mai2017.pdf*.

VORMS, Bernard, JACQUOT, Alain et LHÉRITIER, Jean-Louis, 2010. *L'information statistique sur le logement et la construction*. [en ligne]. 16 mars 2010. Cnis, rapport de groupe de travail. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2010_121_logement_construction.pdf*.

FONDEMENTS JURIDIQUES

Décret n° 2009-318 du 20 mars 2009 relatif au Conseil national de l'information statistique, au comité du secret statistique et au comité du label de la statistique publique. In : *site de Légifrance*. [en ligne]. Mise à jour du 2 janvier 2021. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000020428769/*.

Loi n° 51-711 du 7 juin 1951 sur l'Obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance*. [en ligne]. Mise à jour du 25 mars 2019. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573/.*

Loi n° 2008-776 du 4 août 2008 de Modernisation de l'économie. In : *site de Légifrance*. [en ligne]. Mise à jour du 5 juillet 2019. [Consulté le 26 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000019283050/*.

Présentation du numéro N6

Dans cette sixième édition, le Courrier des statistiques explore quatre sources, deux méthodes, une institution, tout en veillant à rester ouvert sur l'extérieur, en France comme à l'étranger.

Avec la refonte de 2021, l'enquête Emploi modernise ses modes de collecte et s'harmonise avec les exigences européennes. Fidéli, fichier démographique sur les logements et les individus, est devenu incontournable, notamment comme pivot des études sociales. L'échantillon démographique permanent, aux possibilités étendues, apporte une profondeur temporelle aux analyses de trajectoires individuelles. Enfin, le RGCU, gigantesque base de données sur les carrières professionnelles, conçue par la Cnav, promet de devenir une source précieuse pour les chercheurs.

Mais comment apparier des fichiers, sans identifiant commun ? La Depp nous présente sa méthode, à travers son système d'information sur l'insertion des jeunes. En amont, comment améliorer les bases de données administratives ? À cette fin, la Belgique a institutionnalisé et mis en œuvre une démarche, privilégiant des méthodes préventives, fondées sur l'analyse des anomalies.

Le numéro se conclut en illustrant comment le Cnis organise la concertation entre utilisateurs et producteurs de statistiques publiques, pour garantir la pertinence des productions et les améliorer.

ISSN 2107-0903
ISBN 978-2-11-162333-0

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE