Les opportunités créées par la facilitation et le développement des utilisations des données administratives à des fins de recherche et de production statistique

Stéphane Jugnot

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Céreq
WORKING PAPER
09
2021

Les opportunités créées par la facilitation et le développement des utilisations des données administratives à des fins de recherche et de production statistique

Stéphane JUGNOT
Céreq > Direction scientifique

DOCUMENT DE TRAVAIL

DONNÉE STATISTIQUE
PANEL
MÉTHODOLOGIE D'ENQUÊTE
FRANCE
ENQUÊTE
COMPARAISON

Sommaire

Introduction

Dans le cadre de ses missions, fixées par l'article R313-38 du code de l'Education, le Centre d'études et de recherches sur les qualifications (Céreq) réalise des études et des recherches sur les liens entre les qualifications, ses modes d'acquisition et l'emploi, ainsi que sur leurs effets sur les trajectoires professionnelles. Pour cela, il s'appuie sur des matériaux propres qu'il produit et sur l'exploitation de matériaux produits par d'autres, qu'ils soient qualitatifs ou quantitatifs, car la bonne connaissance des phénomènes sociaux en général, comme une bonne évaluation des politiques publiques en particulier, nécessitent de croiser les regards, les approches et les méthodes. L'économétrie seule ne dit rien.

Dans ce contexte, le Céreq est donc conduit à construire des dispositifs d'observation quantitatives, parmi lesquels des enquêtes statistiques. Il exploite aussi des enquêtes et des bases de données produites par d'autres organismes, parmi lesquels l'Institut national de la statistique et des études économiques (Insee) et les services statistiques ministériels concernés par ses domaines d'études et de recherches, en particulier : la Direction de l'animation de la recherche, des études et des statistiques (Dares), pour le ministère chargé de l'Emploi, du Travail et de la Formation professionnelle ; la Direction de l'évaluation, de la prospective et de la performance (Depp), pour le ministère chargé de l'Éducation nationale ; la sous-direction des systèmes d'information et des études statistiques (Sies), pour le ministère chargé de l'Enseignement supérieur et de la Recherche.

Depuis quelques années, l'évolution du cadre légal et réglementaire, notamment sous l'impulsion du Règlement général sur la protection des données (RGPD) et de la loi pour une République numérique, a fortement fait évoluer le paysage des possibles en matière de données quantitatives, alors que la demande de travaux d'évaluation des politiques publiques restait pressante. Cette évolution a ouvert des possibilités nouvelles d'utilisation massive des données individuelles d'origine administrative et, plus particulièrement, d'appariements de ces données entre elles pour produire de nouvelles bases d'informations destinées à la production d'indicateurs statistiques de suivi, de pilotage et de performance, mais aussi à alimenter des travaux d'évaluation et de recherche.

La mise en œuvre du dispositif « Inserjeunes », coproduit par la Depp et la Dares, à partir d'un rapprochement des informations administratives de scolarité et de la déclaration sociale nominative (DSN) en donne une illustration. Le site internet associé[1], mis en ligne au début de l'année 2021, propose ainsi aux jeunes de s'informer sur leur devenir possible à l'issue d'une formation professionnelle précise, par la voie scolaire ou par l'apprentissage, en leur proposant différents indicateurs d'insertion.

Autre illustration avec la signature, en décembre 2020, d'une convention entre le Centre d'accès sécurisé aux données (CASD), la Dares et Pôle emploi pour la production du dispositif « Force » sur les trajectoires d'emploi, de chômage et de formation des chercheurs d'emploi. Ce dispositif s'appuie sur un rapprochement de données administratives sur les demandeurs d'emploi, détenues par Pöle emploi, avec des informations issues de différentes bases de la Dares, elles-mêmes alimentées par des données administratives : la base régionalisée des stagiaires de la formation professionnelle (Brest), la base des jeunes en contact avec une mission locale (I-MILO) et la base des mouvements de main-d'œuvre produite à partir de la déclaration sociale nominative.

Ce paysage en pleine évolution a conduit le Céreq à initier une réflexion sur les opportunités nouvelles que ce nouveau contexte pouvait ouvrir pour ses travaux de recherche et de production statistique. Comme les investigations réalisées à cette occasion peuvent apporter des informations utiles aux chercheurs et aux acteurs concernés par ces évolutions en cours, il a semblé utile de proposer une synthèse de ces investigations et d'une partie de ces réflexions dans le présent document. Ce document de travail ne reflète que le point de vue de l'auteur. Il ne rend pas compte de la position du Céreq. Les informations présentées ont été arrêtées au 31 mai 2021.

La première partie revient sur l'évolution du cadre juridique depuis le début des années 1970 pour montrer en quoi il n'est devenu que tardivement favorable au développement du recours aux données administratives à des fins statistiques et de recherche en France, puisque la Commission nationale de

[1] https://www.inserjeunes.education.gouv.fr/diffusion/accueil

l'informatique et des libertés (Cnil) a vu le jour à la fin des années 1970 contre des projets d'appariement de telles données.

La deuxième partie aborde deux sujets structurants pour les domaines d'études du Céreq :

- d'une part, le déploiement d'un identifiant national des élèves et des étudiants unique et certifié, l'INE, grâce aux investissements réalisés notamment par la Depp et le Sies – ce déploiement devrait permettre, à terme, de mieux reconstituer l'ensemble des trajectoires scolaires à partir des données administratives disponibles ;
- d'autre part, le projet « Agora », piloté par la direction générale de l'emploi et de la formation professionnelle (DGEFP), pour faciliter les échanges d'information entre les différents acteurs de la formation professionnelle (prescripteurs, financeurs, formateurs). Ce projet, en cours de déploiement, s'appuie sur un silo de données partagées qu'alimentent les différents acteurs, qui est aussi accessible à France compétences dans le cadre de ses missions et à la Dares à des fins de production statistique, d'études et de recherches. Les chercheurs peuvent aussi y accéder s'ils sont sous convention avec la Dares.

La troisième partie présente rapidement les premiers panels sociodémographiques alimentés par des données administratives ou pseudo administratives : le panel « Tous salariés » et l'échantillon démographique permanent de l'Insee, ainsi que les fichiers historiques des demandeurs d'emploi de Pôle emploi. Ces panels sont aussi l'occasion d'illustrer certaines des difficultés que posent la production et l'utilisation des bases de données fondées sur des rapprochements de sources administratives. Cette partie est aussi l'occasion d'évoquer deux sources administratives centrales pour tout dispositif d'observation cherchant à suivre des trajectoires professionnelles : la déclaration annuelle de données sociales, maintenant remplacée par la déclaration sociale nominative, et les fichiers historiques de Pôle emploi.

La quatrième partie illustre la multiplication récente des projets d'appariement de données administratives en présentant rapidement plusieurs d'entre eux.

La dernière partie, en guise de conclusion, revient sur les perspectives et les enjeux de ce nouveau paysage en évolution.

1. Un contexte juridique devenu tardivement favorable au développement du recours aux données administratives à des fins statistiques et de recherche

« Il reste que tous les dispositifs techniques valent ce que valent les hommes qui les utilisent »[2]

1.1. Dès les années 1970, des grandes ambitions stoppées par le projet Safari

Longtemps, le développement des appariements de données administratives pour produire des statistiques publiques ou à des fins de recherche a été freiné par des raisons juridiques et des préoccupations éthiques externes qui renvoient aux origines de la loi informatique et libertés, au milieu des années 1970. À cette époque, les administrations commencent à s'informatiser[3] et dès le début, l'idée d'une interconnexion des fichiers sur la base du numéro Insee suscite des débats dans la société, mal anticipés par des administrations désireuses avant tout de profiter du progrès technique.

Encadré 1 • Extraits de l'article « Safari ou la chasse aux Français »

« En ordre dispersé, les départements ministériels tentent de développer à leur profit, à leur seul usage, l'informatique et son outil, l'ordinateur. Ce n'est pas tout à fait un hasard si, à l'époque ou le Journal officiel va publier un arrêté créant une "division informatique" au ministère de la Justice, celui de l'Intérieur met la dernière main à la mise en route d'un ordinateur puissant destiné à rassembler la masse énorme des renseignements grapillés sur tout le territoire ; pas un hasard non plus si le projet SAFARI (Système automatisé pour les fichiers administratifs et le répertoire des individus) destiné à définir chaque Français par un "identifiant", qui ne définisse que lui, maintenant terminé, est l'objet de convoitises ardentes ; le ministère de l'intérieur y souhaite jouer le premier rôle. En effet, une telle banque de données, soubassement opérationnel de toute autre collecte de renseignements, donnera à qui la possèdera, une puissance sans égale ».

« Ainsi se trouve d'évidence posé un problème fondamental, même s'il est rebattu : celui des rapports des libertés publiques et de l'informatique. Son importance exigerait qu'il en fît, au Parlement, publiquement débattu. Tel ne parait pas être, pourtant, la solution envisagée par le premier ministre [...]. »

« Le ministère de l'Intérieur a d'encore plus vastes ambitions. Détenteurs, déjà, du fichier national du remembrement, les services de M. Jacques Chirac font de grands efforts pour, affirme-t-on, s'en adjoindre d'autres : le cadastre, le fichier de la direction national des Impôts et, plus grave peut-être, celui du ministère du Travail. […] »

Par Philippe Boucher, *Le Monde,* 21 mars 1974.

[2] Propos du journaliste interviewant le rapporteur de la loi « Informatique et libertés » pour FR3, dans *Vendredi magazine* du 21 sept 1979 (vidéo de l'INA en ligne sur le site internet de la Cnil le 14 octobre 2020).

[3] L'assurance maladie lance son plan d'informatisation en 1969, l'assurance vieillesse en 1974. L'Insee finit l'informatisation du fichier électoral en 1975. À la même époque, des projets sont en cours au ministère de l'Intérieur, pour le fichier des permis de conduire et pour les recherches criminelles ; au ministère de la Justice, pour la population pénale et les jeunes relevant de l'éducation surveillée ; au ministère du Travail, sur les demandeurs d'emploi, les dossiers de naturalisation et les travailleurs étrangers, etc. Une opération pilote interministérielle a aussi commencé en 1973 pour proposer des fichiers administratifs harmonisés d'informations localisées, qui associe les ministères chargés de l'Économie, de l'Équipement et de l'Éducation.

En 1970, l'Insee initie un projet d'informatiser le répertoire des personnes physiques que René Carmille avait créé sous Vichy pour préparer une future mobilisation[4]. Dès la Libération, le numéro d'inscription au répertoire (Nir) est retenu comme identifiant pour gérer les droits des bénéficiaires de la Sécurité sociale nouvellement créée mais son origine alimentera les suspicions sur les déviances possibles de son utilisation par l'État. Car en juillet 1941, un recensement des personnes de treize à soixante-quatre ans résidant en zone libre a aussi été réalisé pour collecter leur adresse et des informations sur leur profession et leurs qualifications. Leur état civil complet a aussi été collecté dans le but de relier les informations collectées au Nir. Or ce recensement a aussi questionné sur l'appartenance à la « race juive », si bien qu'un doute a, un temps, existé sur l'utilisation des informations détenues par le service statistique national pour procéder au recensement des juifs. Certains se sont aussi demandé s'il avait pu, plus tard, aider à la gestion du service du travail obligatoire, instauré en 1943 pour fournir de la main-d'œuvre à l'Allemagne nazie.

Quand l'Insee lance son projet de « système automatisé pour les fichiers administratifs et le répertoire des individus », plus connu par son acronyme « Safari », il a pour objectif d'étendre progressivement l'utilisation du Nir à l'ensemble des administrations dans un souci de modernisation et d'efficacité, alors que l'informatisation des fichiers administratifs commence[5], mais dès cette époque, l'Insee imagine aussi des « *possibilités nouvelles* » de recours aux données administratives à des fins statistiques, non sans un certain angélisme sur la facilité à les utiliser :

« *L'administration dispose, en effet, d'informations très nombreuses, et l'une des voies les plus riches d'espoir pour la statistique consiste à mobiliser et à coordonner ces informations. C'est souvent par cette seule approche que pourront être obtenues à un coût raisonnable, les données sûres, détaillées, disponibles dans un grand détail géographique requises de plus en plus souvent par les nécessités de l'action. Actuellement, la grande faiblesse des informations administratives est d'être éparses. Or, pour les besoins de l'analyse, c'est disposer d'une information statistique supplémentaire très précieuse que de pouvoir rapprocher des informations qui ont en commun de concerner un même individu […]* »[6].

L'Insee avance notamment deux pistes de travail[7] : d'une part, le rapprochement des bulletins d'état civil et des bulletins des recensement successifs ; d'autre part, le suivi des élèves du secondaire et des étudiants, tant pour suivre les trajectoires scolaires que pour mettre, ensuite, en relation la carrière professionnelle des diplômés et leur formation initiale – cette préoccupation adéquationniste est alors importante. Elle donne naissance au Céreq en mars 1970. Tandis que la première piste se concrétise d'emblée par la mise en place de l'échantillon démographique permanent (EDP)[8], la deuxième piste sera une œuvre de longue haleine, toujours actuelle cinquante ans plus tard[9].

D'emblée, le projet Safari suscite des craintes. Un journaliste les évoque par exemple dans un reportage que l'ORTF consacre à l'informatisation du répertoire dans le journal télévisé de 20h le 23 juillet 1972. C'est en bottant en touche que le responsable de l'Insee lui répond en notant que la connexion entre fichiers pourrait aussi se faire à partir du patronyme des personnes – l'expérience montre qu'en réalité, ce n'est pas si facile. Le responsable de l'Insee poursuit avec ce truisme : il n'y a problème que si certains rapprochent des données qui n'ont pas à l'être... Dans le même reportage, un agent de l'institut enthousiaste espère voir le Nir figurer un jour sur la carte d'identité, car « *ce serait quand même plus*

[4] Pour cela, un service de la Démographie, rattaché au ministère des Finances, est créé en décembre 1940. Il procède jusqu'en aout 1941 à l'immatriculation de 55 millions de personnes nées en France à partir des actes de naissance. Le numéro attribué à chacun a 13 positions et commence par 1 pour les hommes et par 2 pour les femmes – les femmes entrent dans le champ de l'opération pour simuler une opération civile aux yeux des occupants. Ce répertoire s'appuie sur des cartes perforées pour permettre les traitements mécanographiques. À l'automne 1941, le service de la Démographique absorbe la Statistique générale de la France pour devenir le Service national statistique, qui devient l'Insee en 1946. Voir Lévy M.-L., « Le numéro Insee : de la mobilisation clandestine (1940) au projet Safari (1974) », *Dossier et recherche Ined*, n° 86, pp. 23-34, Ined, 2000.

[5] *« On entrevoit immédiatement les commodités qu'apporte un tel mode d'identification à la gestion moderne et notamment à la gestion automatisée […]. Grâce au développement vigoureux de l'informatique, de nombreux fichiers administratifs sont en voie d'automatisation. Les gestionnaires de ces fichiers sont affrontés au problème de la détermination d'un identifiant commode et fiable, et, bien entendu, il apparaît préférable à tous d'utiliser le numéro Insee […]. Ce numéro deviendrait ainsi, très rapidement, commun à toute l'administration (comme le prévoyaient d'ailleurs les créateurs du répertoire) et son usage, pourrait, sauf cas particulier, être rendu obligatoire »* (Desabie J., « L'Insee entreprend d'automatiser le répertoire des personnes », *Economie et statistique*, n°10, pp. 69-71, Insee, 1970).

[6] Jacques Désabie (1970). Voir note *supra*.

[7] Idem.

[8] L'EDP est présenté dans la partie 3.2.

[9] Ce point est développé dans la partie 2.1.

commode, plutôt que d'avoir dans son porte-monnaie un certain nombre de documents comportant chacun des numéros différents, ce qui n'apporte rien à personne »[10].

La controverse publique ne prend vraiment de l'ampleur que deux ans plus tard, après la publication d'un article de Philippe Boucher dans *Le Monde* du 21 mars 1974 intitulé : « *Safari* ou la chasse aux Français » (voir encadré 1). Dans cet article, le quotidien prête au ministère de l'Intérieur la volonté de jouer un rôle central dans l'articulation des fichiers administratifs. C'est cette velléité d'un ministère régalien en charge de la police et des préfets qui est visée par l'article, plus que l'Insee lui-même, sans doute perçu comme une administration de techniciens sans objectifs politiques.

Pour éteindre la polémique, le Premier Ministre interdit aux départements ministériels de procéder sans son autorisation à de nouvelles interconnexions de fichiers. Le gouvernement crée également une première « commission informatique et libertés ». Composée de personnalités qualifiées, elle est placée auprès du ministère de la Justice pour « *proposer au Gouvernement, dans un délai de six mois, des mesures tendant à garantir que le développement de l'informatique dans les secteurs public, semi-public et privé se réalisera dans le respect de la vie privée, des libertés individuelles et des libertés publiques* ». En juin 1975, le rapport de la commission, dit « rapport Tricot », est finalisé[11] (voir encadré 2). En juillet 1976, un projet de loi est présenté et, le 6 janvier 1978, la loi n° 78-17, dite « Informatique et Libertés », est promulguée.

Cette loi crée la Commission nationale de l'informatique et des libertés (Cnil) et pose un certain nombre de principes toujours en vigueur, comme l'interdiction de prendre une décision administrative ou privée sur la seule base d'un profilage automatisé[12], le droit d'information et de rectification des personnes ou la spécificité de certaines données sensibles, celles qui évoquent les origines présumées des personnes, leurs opinions politiques, philosophiques, syndicales ou religieuses. La loi Informatique et Libertés crée surtout une distinction entre le secteur public et le secteur privé, en posant comme principe que les traitements opérés par les administrations et les délégataires de services publics doivent être autorisés par la loi ou par des textes réglementaires après un avis motivé de la Cnil, tandis que le secteur privé est soumis au régime de la déclaration.

Le principe de pertinence des données collectées au regard des finalités du traitement n'est pas inscrit d'emblée de façon explicite dans la loi mais le rapport « Tricot » le pose et la jurisprudence de la Cnil le prend en compte pour l'élaboration de ses avis, d'autant qu'en 1985, la convention 108 du Conseil de l'Europe entre en vigueur. Premier texte international consacré à la conciliation des traitements informatiques et des libertés individuelles, cette convention, qui porte sur la protection des personnes à l'égard du traitement automatisé des données à caractère personnel[13], demande dans son article 5 que les données individuelles soient traitées loyalement et licitement ; qu'elles soient enregistrées pour des finalités déterminées et légitimes et uniquement pour celles-ci ; qu'elles soient adéquates, pertinentes et non excessives par rapport aux finalités pour lesquelles elles sont enregistrées et qu'elles soient conservées sous une forme permettant l'identification des personnes concernées pendant une durée qui n'excède pas celle nécessaire aux finalités pour lesquelles elles sont enregistrées.

Ce cadre ferme, pour un temps, la porte aux rapprochements de données administratives à des fins de statistique publique ou de recherche en ne prévoyant pas explicitement des réutilisations possibles pour ces usages. Légalement, il faut alors les prévoir explicitement dans les actes règlementaires autorisant la création des fichiers visés, ce que le cloisonnement des administrations et leurs priorités propres ne facilitent guère. Le rapprochement de différentes sources sur la base du Nir est, de plus, juridiquement particulièrement contrôlé. Il implique un processus plus lourd à mettre en œuvre puisque le recours au RNIPP doit faire l'objet d'un décret en Conseil d'État après avis de la Cnil.

[10] Vidéo de l'INA en ligne sur le site internet de la Cnil le 14 octobre 2020.
[11] *Rapport de la commission informatique et libertés*, Documentation française, 1975.
[12] « *Aucune décision administrative ou privée impliquant une appréciation sur un comportement humain ne peut avoir pour seul fondement un traitement automatisé d'informations donnant une définition du profil ou de la personnalité de l'intéressé* » (article 2, 2ème alinéa).
[13] Convention signée le 18 janvier 1981 et ratifiée par la France le 24 mars 1983. Elle entre en vigueur en octobre 1985.

Encadré 2 • Extraits du rapport « Tricot » préparant la loi Informatique et Liberté (juin 1975)

« Au total, dans quelques années, l'informatisation devrait avoir franchi une nouvelle étape, caractérisée par une extension beaucoup plus grande et surtout par la constitution de vastes systèmes rassemblant au sujet de catégories humaines (élèves des lycées et étudiants, demandeurs d'emplois, travailleurs étrangers, jeunes enfants, assurés sociaux, malades, demandeurs de crédit, délinquants ou suspects, etc.) des données nombreuses, provenant d'origines diverses et destinées pour la plupart à plusieurs sortes d'utilisateurs. Il en ira de même pour les entreprises et sans doute pour les associations et autres groupements. Ces systèmes, pour partie du moins, ne resteront pas isolés les uns des autres et seront reliés selon leurs affinités et leurs complémentarités. […] »

« Le jour où, au sein de l'État, chaque fonctionnaire qui détient une parcelle de la puissance publique pourrait tout savoir de chaque homme, de chaque famille, de chaque entreprise, ne voit-on pas à quels risques l'administré serait exposé ? Chaque service a sa spécialité. Il a besoin pour exercer sa mission de certaines catégories d'informations qui correspondent aux critères que ce service est légalement autorisé à retenir pour adopter une attitude ou prendre une décision. Il ne faut pas paraître l'inciter à tenir compte d'autres facteurs qui, par hypothèse, ne peuvent pas légitimement intervenir en l'affaire. Ce serait préparer les abus et les détournements de pouvoir. Il nous paraît donc nécessaire de veiller à ce que les informations diffusées dans les services soient toutes celles, mais seulement celles, qui peuvent légitimement concourir à l'exercice de la mission propre à chacun de ces services. »

« Constatons enfin que, dans l'état actuel des choses, l'informatique est plus naturellement à la disposition des puissants qu'à celle des faibles. Elle est coûteuse, elle exige de grands moyens humains et matériels, elle a besoin pour s'alimenter de quantités plus ou moins grandes de données dont n'importe qui ne dispose pas. Ses utilisateurs privilégiés sont donc l'État, les grandes villes, les entreprises publiques, les banques et les assurances, les grandes entreprises privées, les groupements politiques et professionnels les plus puissants. »

1.2. Le retour des possibles

Si la loi de 1978 ne met pas fin à certaines opérations déjà entamées antérieurement par la statistique publique, qui se poursuivent parfois aux marges de la légalité[14], elle rend, de fait, difficile les nouveaux projets puisqu'ils impliquent de produire des textes législatifs ou règlementaires, donc une bonne coopération des administrations concernées, qui peuvent avoir d'autres priorités.

Juridiquement, l'usage de données administratives à des fins statistiques est cependant possible en y mettant les formes. Les années 1980 voient ainsi la mise en place de l'échantillon inter-régimes de retraites, pour évaluer le nombre de retraités et le montant de leurs retraites, après récupération des informations disponibles dans les différentes caisses de retraites, rapprochées sur la base du Nir, en utilisant une méthode en double aveugle et en se limitant à un échantillon de personnes[15]. Mais pour cela, il a fallu une loi, deux décrets et un arrêté pris avec l'avis de la Cnil[16]. En 2000, un échantillon inter-régimes des cotisants est également créé, là encore après le vote d'une loi[17]. En 2003, la constitution deux échantillons est même intégré dans la partie règlementaire du code de la sécurité sociale[18]. L'accès aux données administratives reste donc difficile.

[14] Voir Padieu R, « Exposé liminaire : Mobiliser les données existantes : enjeux et conditions » (pp. 9-16) et Riandey B., « La statistique 20 ans après la loi Informatique et libertés » (pp. 35-41), in *Dossier et recherche Ined*, n° 86, p. 9-34, Ined, 2000.
[15] Voir Faure J.-L. & Lacroix J, « Deux opérations pour mieux connaître la situation des retraités : un échantillon de retraités, des enquêtes sur les allocataires du Fonds national de solidarité », *Courrier des statistiques*, n°40, pp. 21-28, octobre 1986
[16] Loi n° 84-575 du 9 juillet 1984 portant diverses dispositions d'ordre social, décret n° 85-51 du 16 janvier 1985 autorisant l'utilisant du RNIPP pour la gestion des pensions de l'État, décret n° 85-420 du 3 avril 1985 autorisant l'utilisation du Répertoire national d'identification des personnes physiques par les organismes de Sécurité sociale.
[17] Article 27-II de la loi n° 2000-1257 du 23 décembre 2000 de financement de la sécurité sociale pour 2001.
[18] Décret n° 2003-686 du 22 juillet 2003 relatif à l'échantillon interrégimes de cotisants et à l'échantillon interrégimes de retraités et modifiant le code de la sécurité sociale (deuxième partie : Décrets en Conseil d'État).

Un premier assouplissement, fragile, a lieu en 1986 quand la loi du 7 juin 1951[19], qui encadre la statistique publique, est complétée d'un article 7bis autorisant la cession à l'Insee et aux services statistiques ministériels des données administratives détenues par les administrations ou par les établissements publics et personnes morales privées gérant un service public, cession possible à des fins exclusives de statistique publique[20]. Mais, de l'aveu même des rapporteurs du texte au parlement, l'objectif n'est pas tant de développer cette pratique que de protéger du risque pénal les statisticiens publics bénéficiant de transferts ponctuels de données administratives : « *Il ne s'agit nullement ici de créer un flux systématique de toutes les données détenues par les administrations, les établissements publics et les collectivités locales vers l'Insee, mais d'instituer un cadre légal pour des transmissions de données que nos statisticiens peuvent ponctuellement demander aux dépositaires en vue d'une opération statistique bien déterminée. Seront ainsi levées les incertitudes pesant sur la compatibilité entre ces transmissions et les diverses dispositions relative au secret professionnel que l'on trouve non seulement dans le code pénal mais aussi dans de nombreuses autres législations (code des procédures fiscales, code des douanes…)* »[21].

En réalité, la loi de 1986 ne règle pas vraiment la situation puisqu'elle ne modifie pas la loi Informatique et Libertés pour la rendre cohérente avec le nouvel article 7bis de la loi 1951, lequel ne concerne par ailleurs que le service statistique public, dont les contours ne sont pas encore précisément définis par la loi. L'usage des données administratives à des fins de recherche n'est pas encore prévu. Une première évolution effective dans ce sens a lieu en 1994 pour le seul champ des recherches dans le domaine de la santé. La loi du 1er juillet 1994[22] crée ainsi un chapitre spécifique dans la loi Informatique et Libertés, qui autorise l'utilisation à des fins de recherche dans ce domaine des données recueillies par ailleurs par les professionnels de santé. Pour cela, une procédure spécifique est instituée, qui prévoit un avis préalable d'un comité consultatif d'experts, appelé à évaluer la nécessité du recours aux données nominatives et la pertinence des données demandées par rapport à l'objectif de la recherche.

C'est finalement de l'Union européenne que vient la véritable ouverture avec la directive du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données[23]. Cette directive reprend pour l'essentiel des principes généraux déjà mis en œuvre en France par la Cnil mais elle va plus loin pour les chercheurs en autorisant les États à permettre des réutilisations à des fins historiques, statistiques ou scientifiques, même si la finalité initiale ne le prévoyait pas. Comme toute directive, elle n'est pas d'application immédiate et suppose une transposition dans le droit national, censée intervenir au plus tard en octobre 1998. La France le fait avec retard et en plusieurs temps. En 2000, la loi Informatique et Libertés est d'abord modifiée[24] pour autoriser la conservation de données nominatives au-delà de la durée nécessaire à la réalisation des finalités pour lesquelles elles ont été collectées ou traitées, afin de permettre des réutilisations « à des fins historiques, statistiques ou scientifiques », dans le cadre de la loi du 3 janvier 1979 sur les archives.

En 2004, la loi Informatique et Libertés est plus fortement modifiée pour transposer complètement la directive européenne de 1995[25]. À cette occasion, la possibilité ouverte en 1986 pour la statistique publique de traiter des données administratives à des fins statistiques dans le cadre de l'article 7bis de la loi de 1951 est inscrite dans la loi Informatique et Libertés. De façon plus générale, les traitements ultérieurs de données à des fins statistiques ou à des fins de recherche scientifique ou historique sont désormais considérés comme compatible avec les finalités initiales de la collecte des données. Une autorisation préalable de la Cnil reste cependant nécessaire pour l'interconnexion de fichiers aux

[19] Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.
[20] Article créé par la loi 86-1305 du 23 décembre 1986 portant modification de la loi du 7 juin 1951.
[21] Rapport n°30 sur le « Projet de loi portant modification de la loi n° 51-711 du 7 juin 1951 », fait au nom de la commission des lois constitutionnelles, de législation, du suffrage universel, du Règlement et d'administration générale, par le sénateur Jacques Thyraud, octobre 1986, p. 21.
[22] Loi n°94-548 du 1er juillet 1994 relative au traitement de données nominatives ayant pour fin la recherche dans le domaine de la santé et modifiant la loi no 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.
[23] Directive 95/46/CE du Parlement européen et du Conseil, du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données.
[24] Article 5 de la loi n° 2000-321 du 12 avril 2000 relative aux droits des citoyens dans leurs relations avec les administrations.
[25] Loi n° 2004-801 du 6 août 2004 relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel et modifiant la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Voir Isnard M., « Les apports récents de la loi », *Courrier des statistiques* n° 113-114, Insee, pp. 9-13, mars-juin 2005.

finalités initiales différentes ou appartenant à des organismes différents, de même que pour utiliser le Nir ou consulter le RNIPP.

L'assouplissement des contraintes juridiques s'accompagne d'une reprise des réflexions sur les appariements de données administratives au sein de la statistique publique. Ces réflexions sont aussi stimulées par les postures favorables à l'évaluation des politiques publiques, de plus en plus affichée comme une exigence. Avec elle, s'accroit l'intérêt des données de panels, que l'on pense plus faciles et moins couteux à mettre en œuvre quand ils s'appuient sur des données administratives, plutôt que sur des données d'enquêtes, chacune de ces deux approches ayant en réalité leurs avantages et leurs inconvénients. Dans son programme de moyen terme 1999-2003, le Conseil national de l'information statistique (Cnis), organisme sous tutelle de l'Insee, chargé de discuter des besoins sociaux en matière de statistiques publiques, demande ainsi que des panels soient mis en place pour faciliter l'évaluation des politiques publiques dans la sphère sociale, préoccupation qu'il renouvelle lors de la préparation du programme de moyen terme 2004-2008. En Juin 2004, Stefan Lollivier, directeur des statistiques démographiques et sociales de l'Insee, et Mylène Chaleix, présentent un rapport sur le sujet[26]. Même s'ils abordent aussi les panels par enquêtes, les auteurs suggèrent un recours accru aux données administratives, qui leur semble ouvrir des perspectives plus importantes, même si sa mise en œuvre implique des réflexions approfondies sur la manière de garantir les règles de confidentialité et de respect des personnes. Ils recommandent ainsi la mise en place d'un centre d'appariement sécurisé (« CASP ») et suggèrent d'élargir progressivement l'échantillon démographique permanent à d'autres sources administratives disponibles.

Pendant les années suivantes, les initiatives restent limitées en raison d'une frilosité relative de l'Insee qui anticipe des objections possibles de la Cnil. En 2007, la Dares ouvre une brèche en lançant un rapprochement d'extraits du fichier historique des demandeurs d'emploi de l'ANPE (devenu depuis Pöle emploi), avec les déclarations annuelles de données sociales, afin d'évaluer plus finement les effets de la mise en place du projet d'action personnalisé[27]. À vocation expérimentale, le rapprochement est ouvert à plusieurs équipes de chercheurs suite à un appel à projets. La Cnil valide ce dispositif d'observation ponctuel, construit uniquement à des fins d'études et de recherche, sans passer par une obligation légale[28].

La création du centre d'accès sécurisé distant (CASD), retenu comme « équipement d'excellence » dans le cadre du programme des « investissements d'avenir » dès 2010, permet de rendre plus effective l'assouplissement du cadre juridique en offrant aux détenteurs des données et aux chercheurs une solution technologique souple pour exploiter des données individuelles et opérer des rapprochements de fichiers, tout en garantissant la sécurité et la confidentialité de ces données. En autorisant le travail à distance, elle permet aussi aux chercheurs d'envisager de travailler sans convention de mise à disposition et sans travailler dans les locaux des services détenteurs des données.

L'élan donné à l'ouverture des données publiques par la loi du 7 octobre 2016 pour une République numérique et l'assouplissement des procédures juridiques permis par le nouveau règlement général sur la protection des données de 2016, complètent ce paysage devenu désormais favorable aux projets d'utilisations et de rapprochements de fichiers administratifs à des fins de recherche, tant pour la statistique publique *stricto sensu*[29] que pour les chercheurs.

[26] Lollivier S. & Chaleix M. *Rapport sur les outils de suivi des trajectoires des personnes en matière sociale et d'emploi*, Cnis, juin 2004.

[27] L'appariement des deux sources a été initié pour approfondir les travaux d'évaluation déjà réalisés par la Dares à partir du seul fichier historique des demandeurs d'emploi (Debauche E. & Jugnot S., « La mesure d'un effet global du projet d'action personnalisé », *Document d'études*, n°112, Dares, avril 2006), les DADS devant permettre de disposer d'une information plus précise que la seule sortie des listes des demandeurs d'emploi pour identifier les retours vers l'emploi salarié.

[28] Des résultats de ce rapprochement sont présentés dans Le Barbanchon T & Vicard A., « Trajectoire d'une cohorte de nouveaux inscrits à l'ANPE selon le FH-DADS », *Document d'études,* n°152, Dares, décembre 2009.

[29] Depuis la loi n° 2008-776 du 4 août 2008 de modernisation de l'économie, le service statistique publique est défini comme composé de l'Insee et des services statistiques ministériels.

1.3. Le cadre juridique en vigueur est plutôt favorable à la production de connaissance

1.3.1. Le règlement général sur la protection des données de l'Union européenne (RGPD)

Le règlement 2016/679 de l'Union européenne, dit règlement général sur la protection des données (RGPD), remplace la directive de 1995. Comme tous les règlements européens, le RGPD est directement applicable dans l'ordre juridique français. Pour plus de lisibilité, la loi Informatique et Libertés a été fortement réécrite par voie d'ordonnance en 2018 pour disposer d'une version conforme au RGPD. Cette nouvelle version de la loi Informatique et Libertés est entrée en application le 1er juin 2019.

Le RGPD a fait évoluer substantiellement l'ordre préexistant sur les procédures à respecter préalablement à tout traitement de données, en effectuant un transfert de responsabilité de la Cnil vers les organisations traitant des données individuelles[30]. Désormais, celles-ci doivent notamment mettre en place un délégué à la protection des données, chargé de s'assurer du respect du RGPD dans l'organisation. C'est avec ce délégué que les organisations doivent définir les règles de fonctionnement qu'elles doivent mettre en place pour respecter le RGPD et qu'elles doivent s'assurer que leurs traitements sont licites. Elles doivent aussi tenir un registre des traitements qu'elles opèrent et procéder à une analyse d'impact *« lorsque le traitement* est *susceptible d'engendrer un risque élevé pour les droits et libertés des personnes concernées* » (des critères et des exceptions sont prévus). Avec cette révolution, l'envoi d'une déclaration à la Cnil ne peut donc plus servir de paravent aux uns pour faire ou d'excuse aux autres, pour refuser de faire, alors que la Cnil ne disposait pas de moyens suffisants pour traiter sérieusement chacun des dossiers qu'elle recevait.

Le RGPD et la nouvelle loi Informatique et Libertés confortent les souplesses destinées à faciliter les traitements de données personnelles à des fins de recherche et de production statistique. En particulier :

- la limitation de la durée de conservation des données à la durée nécessaire aux finalités premières du traitement n'est pas opposable[31] ;
- des informations « sensibles » peuvent être collectées sous certaines conditions[32] ;
- des dérogations sont possibles aux obligations générales relatives aux droits des personnes sur le droit d'accès, le droit de rectification, le droit de limitation du traitement et le droit d'opposition[33], dans un cadre précisé par un décret en Conseil d'État[34] ;

Cependant, en contrepartie, pour bénéficier de ces dérogations, les traitements à des fins de recherche scientifique ou à des fins statistiques doivent respecter un certain nombre de garanties, dont « *la mise en place de mesures techniques et organisationnelles, en particulier pour assurer le respect du principe de minimisation des données ».*

[30] Plus exactement, les autorités ou organismes publics procédant à des traitements de données individuelles, ainsi que « les organismes ayant pour activité de base des opérations de traitement nécessitant le suivi régulier et systématique des personnes à grande échelle » et les « organismes ayant pour activité de base le traitement à grande échelle de données dites *sensibles* ou relatives aux condamnations pénales et infractions ».

[31] Article 5-1-e du RGPD (repris dans l'article 4-5° de la loi Informatique et Libertés) :« [Les données à caractère personnel doivent être] conservées sous une forme permettant l'identification des personnes concernées pendant une durée n'excédant pas celle nécessaire au regard des finalités pour lesquelles elles sont traitées ; les données à caractère personnel peuvent être conservées pour des durées plus longues dans la mesure où elles seront traitées exclusivement à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques conformément à l'article 89, paragraphe 1, pour autant que soient mises en œuvre les mesures techniques et organisationnelles appropriées requises par le présent règlement afin de garantir les droits et libertés de la personne concernée (limitation de la conservation) »

[32] Article 9-2-j du RGPD (repris dans l'article 6-II de la loi Informatique et libertés et l'article 44).

[33] Article 89-2 du RGPD (repris dans l'article 78 de la loi Informatique et libertés) : « Lorsque des données à caractère personnel sont traitées à des fins de recherche scientifique ou historique ou à des fins statistiques, le droit de l'Union ou le droit d'un État membre peut prévoir des dérogations aux droits visés aux articles 15 [droit d'accès], 16 [droit de rectification], 18 [droit de limitation du traitement] et 21 [droit d'opposition], sous réserve des conditions et des garanties visées au paragraphe 1 du présent article, dans la mesure où ces droits risqueraient de rendre impossible ou d'entraver sérieusement la réalisation des finalités spécifiques et où de telles dérogations sont nécessaires pour atteindre ces finalités ».

[34] Décret n° 2019-536 du 29 mai 2019 pris pour l'application de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (par son article 116).

Dans le cas du recours à des données administratives externes, ce principe de minimisation implique de limiter les données récupérées aux seules informations strictement nécessaires pour mener à bien l'opération statistique ou la recherche annoncée et de ne les conserver que le temps strictement nécessaire à la réalisation de cette opération. Le principe de minimisation oblige donc les statisticiens et les chercheurs à bien réfléchir d'emblée aux objectifs visés et aux moyens d'y parvenir de façon à calibrer correctement leurs demandes de données et les durées de conservation dès le départ, sous le regard des délégués à la protection des données impliquées. Dans l'esprit du RGPD, les chercheurs et les statisticiens ne doivent pas adopter une position maximaliste de confort en se gardant des marges de manœuvre importantes « au cas où ». Ils doivent réfléchir d'abord.

La « pseudonymisation » fait partie des mesures que les chercheurs et les statisticiens doivent mettre en œuvre à chaque fois que cela est possible dans le cadre du principe de minimisation[35]. Pour définir ce qu'est la pseudonymisation, le RGPD définit d'abord les « données à caractère personnel » comme des données permettant l'identification directe ou indirecte d'une personne physique, « notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale ».

Autrement dit, un fichier ne contenant ni l'identité des personnes, ni d'identifiant de gestion associé aux personnes peut rapidement constituer un fichier de données à caractère personnel par croisement d'informations précises sur, par exemple, le lieu de résidence, l'âge, le sexe et la nationalité, pour retenir des informations souvent présentes dans les données administratives ou les enquêtes sociales. Dans ce cas, la « pseudonymisation » consiste à conserver séparément les informations individuelles qui permettent directement ou facilement d'identifier les personnes ou une partie d'entre elles[36], d'une part ; des autres informations, qui ne peuvent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires externes, d'autre part. Des mesures techniques et organisationnelles, dont la gestion des droits d'accès, doivent alors garantir que ces deux ensembles de données ne sont pas facilement rapprochables, de sorte que les données « pseudonymisées » ne puissent être attribuées à une personne physique identifiée ou identifiable.

Pour mieux cadrer l'application du RGPD dans le cas des traitements de données à des fins de recherche scientifique hors santé, la Cnil a lancé une consultation auprès des chercheurs en juillet 2019. Clôturée en septembre 2019, cette consultation devait, selon les termes de la Cnil, « permettre une meilleure compréhension des traitements de données personnelles dans la recherche scientifique, clarifier le cadre juridique applicable et concevoir des fiches pratiques adaptées ». Au 31 mai 2021, la Cnil n'avait pas encore publié de documents, guides ou fiches pratiques sur ce sujet.

1.3.2. Des différences demeurent entre la statistique publique et les organismes de recherche pour l'utilisation du Nir

Les rapprochements de fichiers administratifs sont plus aisés à réaliser lorsque les fichiers à rapprocher utilisent un même numéro d'identification certifié. Les erreurs d'identification et le travail nécessaire pour préparer les fichiers en amont du rapprochement sont alors moindresque si l'on utilise tout ou partie des informations d'état civil (noms, prénoms, date et lieu de naissance), éventuellement complétées d'autres informations comme l'adresse. Dans ce cas, les erreurs de saisie, les informations parcellaires, les homonymies et les différences de format des variables entre les fichiers[37] sont sources de difficultés supplémentaires à traiter.

Dans la sphère sociale, de nombreux fichiers de gestion utilisent le numéro d'identification au répertoire des personnes physiques géré par l'Insee, plus communément appelé « numéro de sécurité sociale ». Le Nir est en particulier disponible dans les fichiers de gestion de Pöle emploi et dans la déclaration

[35] « *Ces mesures [de minimisation] peuvent comprendre la pseudonymisation, dans la mesure où ces finalités peuvent être atteintes de cette manière. Chaque fois que ces finalités peuvent être atteintes par un traitement ultérieur ne permettant pas ou plus l'identification des personnes concernées, il convient de procéder de cette manière* » - Article 89-1 du RGPD.

[36] Parmi les données concernées : les noms et prénoms, les coordonnées postales, téléphoniques et internet, mais aussi les variables de localisation détaillée à la commune ou infra communale (lieu de résidence, lieu de travail), l'identification du lieu de travail (Siren, Siret, raison sociale).

[37] Par exemple, utilisation ou non des majuscules et des caractères accentués, enregistrement du seul prénom d'usage ou de tous les prénoms, utilisation du code postal, du code du Insee ou du Libellé pour la commune de naissance, etc.

sociale nominative. Il est aussi techniquement possible de le récupérer auprès de l'Insee par la consultation du répertoire des personnes physique sur la base des informations d'état civil mais dans ce cas, l'existence d'informations parcellaires et d'erreurs de saisie dans les informations initiales peuvent aussi perturber la récupération du Nir. C'est notamment le cas pour les personnes nées à l'étranger. Malgré cette limite, le Nir reste une variable intéressante pour tout projet de rapprochement de fichiers quand elle est disponible ou récupérable.

Son utilisation reste toutefois très encadrée puisque, de façon générale, l'utilisation du Nir doit être autorisée par un décret en Conseil d'État pris après un avis motivé de la Cnil[38]. Sous certaines conditions, les traitements qui ont des finalités exclusives de recherche scientifique peuvent déroger à cette obligation, de même que les traitements qui relèvent de la statistique publique. Dans ce deuxième cas, le traitement ne doit pas porter sur des informations sensibles et doivent être mis en œuvre par le « service statistique public ». L'enquête Génération du Céreq, bien que labellisée « statistique publique », ne peut donc être couverte par cette deuxième exception puisque le Céreq n'est pas membre du service statistique public au sens de l'article 1-I de la loi n°51-711 du 7 juin 1951.

Ces dérogations ont été introduites dans la loi Informatique et Libertés par la loi du 7 octobre 2016 pour une république numérique[39], avant la transposition du RGPD. Elles concernent l'utilisation du Nir et la consultation du répertoire des personnes physiques. Pour profiter de ces dérogations, le Nir doit faire préalablement l'objet d'une opération cryptographique pour lui substituer un code non signifiant[40]. Sur ce point, le gouvernement a fait le choix de distinguer le service statistique public des organismes de recherche :

- pour les organismes de recherche, un cryptage spécifique doit être effectué pour chaque opération et le recours à des tiers de confiance est nécessaire, d'une part, pour opérer le cryptage ; d'autre part, pour effectuer l'interconnexion des fichiers sur la base du Nir crypté. Dit autrement, le responsable du traitement ne peut effectuer par lui-même, ni l'opération de cryptage, ni le rapprochement des fichiers et ce rapprochement ne peut être effectué non plus par l'opérateur de cryptage.
- les services statistiques publics peuvent, en revanche, utiliser une même clef de cryptage de sorte qu'un Nir donné fournisse le même « code statistique non signifiant » pour tous les fichiers. Il devra être modifié régulièrement, selon une fréquence définie par un décret en Conseil d'État qui l'a fixée à dix ans[41]. Une réflexion est aussi en cours au sein de l'Insee pour que ce code statistique non signifiant soit systématiquement introduit dans des fichiers jugés « pivots »[42], susceptibles d'être fréquemment mobilisés dans des appariements de fichiers. Sont notamment évoquées : les déclarations sociales nominatives, les enquêtes annuelles de recensement, les données fiscales sur les personnes et sur les logements. Il ne serait pas non plus illogique que les fichiers de Pöle emploi soient aussi considérés comme « pivot ».

Comme par un retournement de l'histoire, cette possibilité d'introduire un même identifiant dans tous les fichiers administratifs de données individuelles manipulés par le service statistique public réouvre les opportunités que l'Insee visait il y a cinquante ans avec son projet Safari. Alors que celui-ci fut à l'origine de polémiques qui portèrent la Cnil sur des fonds baptismaux marqués par une forte vigilance sur les activités de l'État, les opportunités les plus larges sont désormais réservées à des services constituant des administrations ministérielles, plutôt qu'aux organismes de recherche.

[38] Article 30 de la loi Informatique et Libertés issue de la transposition du RGPD.

[39] Cette innovation fait partie des points qui ont été les plus discutés par les internautes dans le cadre de dispositif participatif que le gouvernement avait mis en place pour recueillir l'avis du public sur les différentes mesures prévues par le projet de loi, voire de nouvelles propositions. Ce dispositif participatif permettait ainsi de voter pour chaque mesure envisagée et de les commenter. 706 votes ont été recueillis pour la mesure relative à l'assouplissement des règles encadrant l'usage du Nir à des fins de recherche et de statistiques publiques. Parmi ces votes, 410 « pour », 195 « contre » et 101 « mitigés ». Par comparaison, l'article 1er posant le principe d'open data des données publiques a recueilli le plus de votes, 2440, dont 2232 « pour », 49 « contre » et 159 « mitigés ».

[40] Cette opération de cryptographie est aussi dispensée de l'obligation d'un décret en Conseil d'État.

[41] La fréquence de dix ans a été fixée par l'article 31 du décret n° 2018-687 du 1er août 2018 pris pour l'application de la loi Informatique et Libertés modifiée par le RGPD. Cet article modifie le décret n° 2016-1930 du 28 décembre 2016 publié après le vote de la loi pour une République numérique, qui précise les modalités des opérations de cryptage du Nir, tant pour les traitements de la Statistique publique que pour les traitements à finalités de recherche. Ce texte est complété par l'arrêté du 28 septembre 2020 pris en application des articles 3 et 4 du décret n° 2016-1930.

[42] Voir par exemple l'intervention de C. Colin sur le code statistique non signifiant au bureau du Cnis du 9 décembre 2020.

En rendant possible dans la durée toute interconnexion entre fichiers disposant du code statistique non signifiant et en envisageant de l'introduire systématiquement dans des fichiers « pivots », les services ministériels concernés pourront aussi faire l'économie d'une réflexion *ex ante* sur les différents panels qu'il serait pertinent qu'ils constituent pour répondre aux problématiques récurrentes qu'ils ont à traiter. À rebours de l'esprit général du RGPD, la diffusion d'un Nir crypté partagé dans les fichiers administratifs permettra ainsi à la statistique publique de se concentrer sur les outils en reportant à plus tard la détermination des finalités statistiques envisagées. Elle permettra de rendre les choses possibles au moment où l'idée de faire viendra.

1.3.3. La loi Numérique encourage l'accès aux données administratives

Outre la simplification juridique de l'accès aux Nir pour les travaux de recherche et la statistique publique, la loi Numérique contribue aussi à rendre le contexte plus favorable à l'accès aux données administratives en posant dans son article 1er un principe d'open data des données publiques et la gratuité de cet accès entre administrations dès lors que l'accès est demandé dans le cadre de l'accomplissement des missions du demandeur[43].

L'administration détentrice des données peut alors, sans y être obligé, demander une saisine du comité du secret mis en place par la loi du 7 juin 1951, pour qu'il recommande, s'il le juge utile, le recours à une procédure d'accès sécurisé aux données, en tenant compte de la nature et de la finalité des travaux envisagés. Dans ses avis, ce comité tient compte des secrets protégés par la loi, notamment le secret statistique, la protection de la vie privée et la protection du secret des affaires[44]. Le passage au comité du secret est obligatoire pour accéder aux données individuelles des enquêtes labellisées « statistique publique » ainsi qu'aux autres données individuelles collectées par la statistique publique dans le cadre de ses missions. À l'initiative des administrations des impôts, il l'est aussi pour accéder aux données fiscales[45].

1.4. Eléments de réflexion pour le Céreq

Le cadre juridique actuel permet théoriquement au Céreq d'accéder aux fichiers administratifs nécessaires aux travaux de recherche relevant de son champ :

- D'abord, parce que la réalisation de ces traitements est « nécessaire à l'exécution d'une mission d'intérêt public ou relevant de l'exercice de l'autorité publique dont est investi le responsable du traitement » (article 6-e du RGPD et 5° de l'article 5 de la loi Informatique et Libertés). Pour l'enquête Génération, le fait que l'enquête soit une enquête de la statistique publique au sens de la loi n° 51-711 du 7 juin 1951 renforce ce point, comme sa mention dans le contrat d'objectifs et de performance signé par l'établissement avec ses ministères de tutelle.
- Ensuite, parce que le RGPD, et conséquemment la loi Informatique et Libertés, considèrent que le fait que l'exploitation à des fins de recherche scientifique ou de production statistique d'un fichier administratif n'a pas été prévue dans ses finalités initiales n'est pas opposable (article 5-1-b du RGPD[46] et article 4-2° de la loi Informatique et Libertés[47]).

---

[43] *« Sous réserve des articles L. 311-5 et L. 311-6 du code des relations entre le public et l'administration […], les administrations […] sont tenues de communiquer, dans le respect de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, les documents administratifs qu'elles détiennent aux autres administrations […] qui en font la demande pour l'accomplissement de leurs missions de service public. »*
« Les informations figurant dans des documents administratifs communiqués ou publiés peuvent être utilisées par toute administration […] qui le souhaite à des fins d'accomplissement de missions de service public autres que celle pour les besoins de laquelle les documents ont été produits ou reçus. »
« À compter du 1er janvier 2017, l'échange d'informations publiques entre les administrations de l'État, entre les administrations de l'État et ses établissements publics administratifs et entre les établissements publics précités, aux fins de l'exercice de leurs missions de service public, ne peut donner lieu au versement d'une redevance. »

[44] La procédure a été codifiée par le décret n° 2017-349 du 20 mars 2017 relatif à la procédure d'accès sécurisé aux bases de données publiques.

[45] III de l'article L135 D du Livre des procédures fiscales.

[46] « Le traitement ultérieur à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques n'est pas considéré, conformément à l'article 89, paragraphe 1, comme incompatible avec les finalités initiales (limitation des finalités) ».

[47] «Toutefois, un traitement ultérieur de données à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique, ou à des fins statistiques est considéré comme compatible avec les finalités initiales de la collecte des données, s'il

Cependant, les responsabilités données aux délégués à la protection des données et aux responsables des traitements peuvent voir naître une hétérogénéité des pratiques des organismes détenteurs des données primaires. Le champ des possibles ouvert aux chercheurs par la loi peut se retrouver restreint par les politiques de ces organismes, leur charge de travail et les perceptions subjectives de leur délégué à la protection des données et de leurs responsables des traitements sur les conditions pratiques de la mise en œuvre du RGPD, ces perceptions subjectives dépendant autant de leurs compétences que de leurs aversions au risque. L'effectivité du champ des possible peut aussi dépendre du point de vue du délégué à la protection des données et du responsables des traitements de l'organisation où les chercheurs exercent.

Enfin, le RGPD oblige à apporter des garanties organisationnelles et techniques en matière de gestion de la confidentialité pour assurer le respect du principe de minimisation. Il importe donc pour les organismes de recherche tel que le Céreq de progresser dans la sensibilisation et la formation des chargés d'études, en particulier pour l'organisation du stockage des données, la gestion des accès et des durées de vie des fichiers de données individuelles. La multiplication des formations et des séminaires sur ces sujets dans le monde de la recherche atteste de ce besoin.

est réalisé dans le respect des dispositions du règlement (UE) 2016/679 du 27 avril 2016 et de la présente loi, applicables à de tels traitements et s'il n'est pas utilisé pour prendre des décisions à l'égard des personnes concernées ; »

2. Le développement de systèmes d'information de plus en plus centralisés et articulés

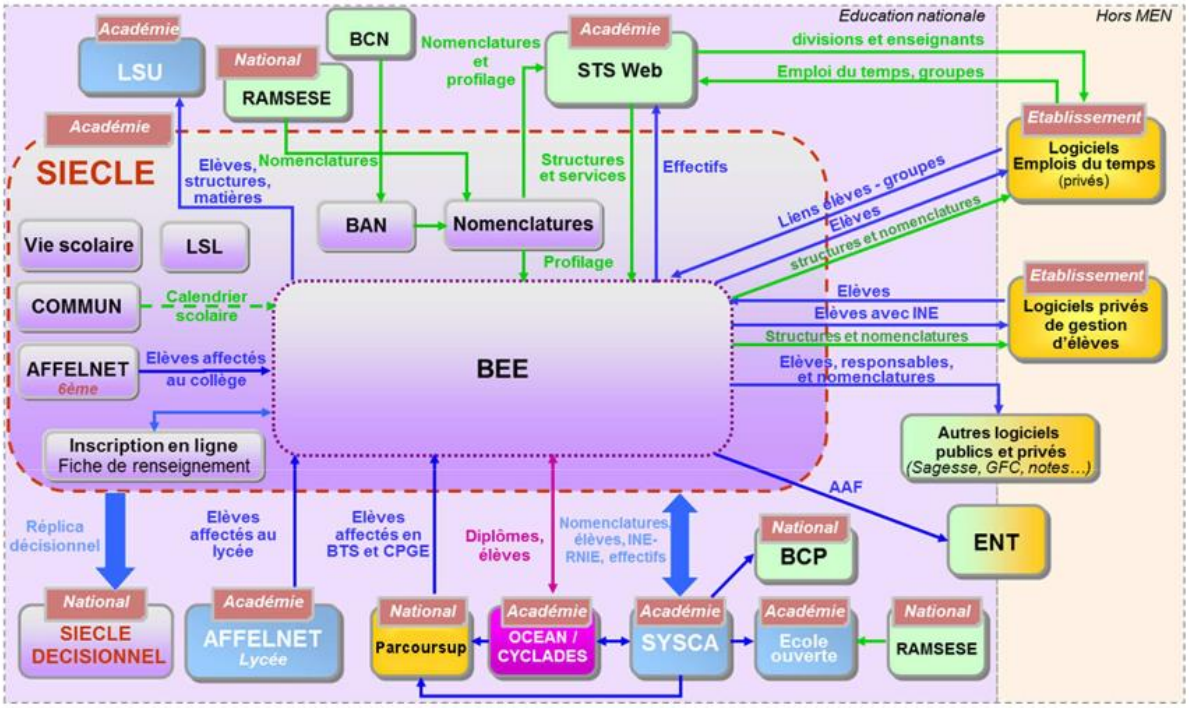
Dans les champs investigués par le Céreq, plusieurs systèmes d'information recueillant des informations de gestion tendent à se centraliser et à s'articuler davantage entre eux ces dernières années. C'est en particulier le cas pour les informations relatives aux parcours scolaires, avec la mise en place d'un identifiant unique certifié géré centralement : l'identifiant national des élèves et des étudiants (INE), identifiant mobilisé dans de nombreux outils de gestion des élèves, pour leur inscription, leurs choix d'orientation, leur réussite aux examens, etc. C'est aussi le cas dans le champ de la formation professionnelle continue, avec le déploiement du projet « Agora ».

Dans les deux cas, le champ des possibles théorique ne s'élargit cependant que de façon progressive en raison de contraintes techniques mais aussi du fait de la pluralité des acteurs impliqués. Par leur statut, ceux-ci sont plus ou moins directement liés par les choix techniques de l'État. Ils peuvent disposer de leurs propres outils et pratiques, ce qui ralentit les processus d'homogénéisation. Des débats éthiques, avec leurs implications politiques, peuvent aussi freiner leur déploiement.

2.1. Dans le champ scolaire, l'identifiant national des élèves et étudiants certifié (INE) devrait permettre de suivre les parcours scolaires dans tous ses aspects

Depuis plusieurs décennies, le suivi des parcours scolaires des élèves du primaire à l'enseignement supérieur est une priorité des services statistiques ministériels de l'Éducation et de l'Enseignement supérieur – l'idée germait déjà dès le projet Safari. La mise en place de l'identifiant national des élèves et étudiants (INE), appuyée sur un répertoire géré par la Depp, doit répondre à cet objectif mais cet identifiant est aussi et d'abord un identifiant largement utilisé dans les applications de gestion du système scolaire, où il tient le même rôle que le Nir dans la sphère sociale.

Schéma 1 • Articulation des systèmes d'information du second degré



Source : ministère de l'Éducation nationale, guide utilisateur de la BEE, 2018.

En particulier, dans les établissements relevant de l'Éducation nationale, l'INE est nécessaire aux chefs d'établissements pour inscrire les élèves, au premier degré et au second degré. Il est ensuite demandé lors des inscriptions dans l'enseignement supérieur. Il a aussi diffusé dans de nombreuses applications de gestion du système scolaire, comme Affelnet (pour les orientations en sixième et en troisième), ABP puis Parcoursup (pour l'orientation post-bac), Ocean et Cyclades (pour la gestion des inscriptions et des résultats aux examens nationaux et aux concours, dont le brevet et le baccalauréat) ou le livret scolaire numérique.

2.1.1. En 1995, l'INE est créé pour gérer les élèves du secondaire des établissements relevant du ministère de l'Éducation nationale

Depuis 1995, les élèves entrant dans le secondaire public se voient attribuer un identifiant national unique. Cet identifiant, appelé alors « numéro de matricule national », fait partie des données mentionnées dans l'arrêté du 22 septembre 1995 portant création du traitement « Scolarité ».

Ce traitement est d'abord destiné à la gestion administrative, pédagogique et financière des élèves au niveau de l'établissement, ainsi qu'au pilotage et à la gestion au niveau des rectorats et des inspections d'académie. Il doit enfin permettre de faciliter le pilotage au niveau de l'administration centrale en permettant à la Depp d'accéder à des extraits à des fins statistiques, notamment pour alimenter des panels d'élèves sur des échantillons réduits[48]. Destiné aux établissements publics, le système d'information Scolarité est également accessible aux établissements privés qui le souhaitent.

Ce système d'information et ses applications associées articulent des bases qui restent au niveau de l'établissement (les bases Elèves Etablissements, BEE), des bases situées au niveau académique et une base centrale de pilotage pour l'administration centrale. Seules certaines informations peuvent circuler entre les différents niveaux[49]. Le numéro de matricule national est accessible à tous les niveaux. Il permet d'articuler les informations relatives à un même élève. Ce traitement a fait l'objet d'un avis favorable de la Cnil, sans réserve notable (délibération 93-074 du 7 septembre 1993).

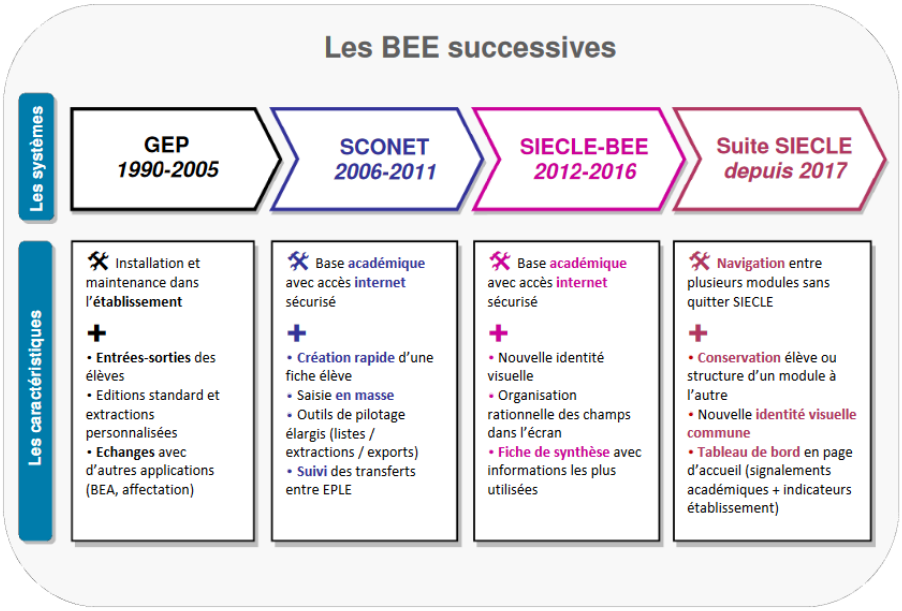
L'arrêté du 10 décembre 2002 change la dénomination du numéro matricule national pour l'appeler « identifiant national élève (INE) ». Cet « INE BEA » est généré automatiquement au niveau de l'Académie à partir de l'état civil complet des élèves lors de leur entrée dans le secondaire[50]. Les éventuelles anomalies ou doublons apparents sont traités manuellement par des gestionnaires au niveau de l'académie. Jusqu'à la fin du lycée, l'INE reste un identifiant de gestion interne que les familles n'ont pas besoin de connaître. Pour les élèves provenant d'un établissement public de leur académie, l'INE est récupéré par les collèges à l'entrée en sixième avec les dossiers élèves transmis informatiquement en aval de l'application d'affectation Affelnet déployée au sein de chaque académie. Pour leurs élèves provenant d'une autre académie, les établissements doivent normalement récupérer l'INE attribué dans l'académie de départ, mais l'existence réitérée de cette consigne dans les manuels d'utilisation de l'application témoigne qu'en pratique, ils peuvent aussi faire générer un nouvel INE pour ne pas bloquer la suite du processus, notamment la constitution de leurs classes. Pour les élèves venant d'établissements n'attribuant pas d'INE, notamment pour les élèves venus de l'étranger, la génération d'un INE doit être aussi demandé avant de pouvoir finaliser les inscriptions.

[48] L'échantillon doit concerner au plus 5 % des élèves du champ étudié. Pour ces panels, la Depp peut notamment récupérer dans Scolarité : l'état civil et la nationalité de l'élève, ses liens de parenté avec ses responsables légaux, les catégories socioprofessionnelles du père et/ou de la mère, la situation scolaire de l'année en cours (formation, options), la situation scolaire relative à l'année précédente, les vœux d'affectation.

[49] Le système SCONET (Scolarité sur le Net) a pris la relève au milieu des années 2000 (expérimentation à partir de 2004, généralisation en janvier 2006). À partir de 2012, il est remplacé par le système d'information pour les élèves de collèges, de lycées et pour les établissements (SIECLE). SIECLE est interfacé à diverses applications nationales (base centrale des nomenclatures, APB puis Parcoursup) et des applications gérées au niveau académique (BEA, Ocean, Affelnet). Il peut aussi s'interfacer à des logiciels privés, utilisés par exemple pour gérer les notes ou les absences des élèves.

[50] Code sur 11 positions : code de l'académie (2 digits), année de codification de l'élève (2 digits), numéro d'ordre (6 digits) et lettre clé de contrôle (1 digit). Il semblerait qu'auparavant, l'Éducation nationale utilisait le Nir dans ses fichiers de gestion (voir l'intervention d'Alain Goy de la Depp à la réunion de la formation Education du Cnis du 20 avril 2005, en réponse à Michel Théry).

Schéma 2 • Les bases Elèves – Etablissements successives dans le second degré



Source : ministère de l'Éducation nationale, guide utilisateur de la BEE, 2018.

2.1.2. Au début des années 2000, la volonté d'étendre l'INE au premier degré se heurte à des contestations

Au début des années 2000, le ministère de l'Éducation prépare puis déploie un système d'information analogue à celui du secondaire pour les établissements du premier degré, maternelles et écoles primaires. Comme Scolarité dans le secondaire, cette Base Elèves du 1er degré (BE1D) est d'abord un outil d'aide à la gestion administrative des élèves au niveau de l'établissement, qui permet aux directeurs d'école de préparer la rentrée scolaire et de réaliser différents actes de gestion au cours de l'année (saisie des admissions et des radiations, constitution des classes, édition des certificats de scolarité, enregistrement des passages au niveau supérieur, etc.), sans couvrir pour autant l'ensemble des sujets[51]. Il permet aussi des remontées d'informations au niveau académique à des fins de gestion et de pilotage et des remontées d'informations individuelles à la Depp.

Mise en place de façon expérimentale dans certains établissements à partir de 2004, la BE1D commence à être déployée en 2007. 40 % des élèves du champ sont couverts à la rentrée 2007, 80 % à la rentrée 2008 et la totalité à la rentrée 2009.

La base nationale des identifiants élèves (BNIE) est créée en même temps que la BE1D et en articulation avec celle-ci, pour attribuer un INE aux élèves dès leur entrée en primaire, quel que soit l'établissement, public ou privé. Normalement, les élèves scolarisés à domicile sont également immatriculés. Cette base nationale est gérée par la Depp. La BE1D est donc interfacée avec la BNIE de façon à attribuer à tout nouvel entrant un INE unique. Là encore, les traitements sont automatiques, avec des gestionnaires dans les académies pour traiter des anomalies[52]. À l'occasion de la déclaration de la BE1D fin 2004, il est aussi acté avec la Cnil que le l'INE doit être différent du Nir[53] et que c'est un INE crypté qui doit être utilisé pour rapprocher plusieurs fichiers successifs afin d'étudier des parcours scolaires.

[51] Par exemple, comme Scolarité, il n'intègre pas de fonctionnalité sur la notation des élèves.
[52] Lettre du directeur des affaires juridiques du ministère de l'Éducation nationale à la Cnil du 15 février 2006 (DAJ-A3/04-452). Réponse de la Cnil demandant des précisions techniques le 12 juin 2006 (AT/SV/SN/LBA/D1062227). Réponse du ministère détaillant le fonctionnement pratique le 8 février 2007. Récépissé de la Cnil du 27 février 20O7.
[53] Lettre du ministère de l'Éducation nationale à la Cnil du 24 décembre 2004 pour déclarer le traitement BE1D. L'arrêté créant le traitement n'est signé que quatre ans plus tard, après le déploiement du système d'information (arrêté du 20 octobre 2008 portant création d'un traitement automatisé de données à caractère personnel relatif au pilotage et à la gestion des élèves de l'enseignement du premier degré).

Même si la création de la BNIE et l'INE est explicitement inscrite dans la perspective de suivre l'élève pendant tout son parcours scolaire, de la maternelle à l'enseignement supérieur[54], ces ambitions initiales sont remises en cause par les polémiques publiques qui accompagnent leur déploiement. Des syndicats professionnels, les associations de parents d'élèves et la Ligue de droits de l'Homme dénoncent de conserve le fichage des élèves et les utilisations détournées possibles de certaines informations individuelles que l'Éducation nationale souhaite enregistrer dans ses bases, notamment sur les origines des élèves.

Ces polémiques se développent surtout au cours des années 2007-2008[55]. Des procédures judiciaires sont même engagées pour contester la mise en place de ces outils informatiques[56]. Pour y mettre fin, le ministre de l'Éducation nationale, Xavier Darcos, annonce le retrait de l'enregistrement de certaines informations initialement prévues et validées par la Cnil, notamment les origines sociales des parents, la nationalité, l'année d'arrivée en France et la langue parlée à la maison. Il annonce aussi que la durée maximale de conservation des données n'excédera pas le terme de l'année civile au cours de laquelle l'élève n'est plus scolarisé dans le premier degré[57] alors que le projet initial prévoyait une durée de conservation totale de 35 ans[58]. Ces décisions ministérielles sont actées dans l'arrêté du 20 octobre 2008 créant le traitement Base Elèves du premier degré. Elles conduisent à retarder l'extension de l'utilisation de la BNIE au secondaire et au supérieur sans que l'Administration n'abandonne ses objectifs. Ce n'est que partie remise.

2.1.3. En 2012, la création du répertoire national des identifiants élèves et étudiants relance l'objectif de suivre les élèves tout au long de leur parcours scolaire

En 2010, la Depp annonce au Cnis la transformation future de la BNIE en RNIE, « répertoire national des identifiants élèves et étudiants ». Ce changement de nom doit faciliter son acceptation en faisant oublier la mobilisation sociale hostile à la BNIE et en mettant en avant la notion de répertoire afin de « tradui[re] la volonté d'affirmer le statut du fichier, qui est avant tout un répertoire et non une base de données informative. L'objectif est en effet de pouvoir attribuer un numéro (INE=Identifiant national élève-étudiant) unique à tout élève scolarisé et de retrouver ce numéro tout au long de sa scolarité. Ce numéro doit servir à améliorer la gestion du système éducatif (élimination des doubles inscriptions…) et sera inclus, sous une forme cryptée, dans les fichiers à finalité statistique pour permettre l'étude de trajectoires d'élèves »[59].

Le nouveau répertoire est déclaré à la Cnil le 6 octobre 2011 puis créé officiellement par l'arrêté du 16 février 2012[60]. Il reste sous la responsabilité de la Depp et couvre l'ensemble de la scolarité, de la maternelle au supérieur, dans la finalité de « faciliter la gestion du système éducatif et [de] permettre le suivi statistique des élèves, des étudiants et des apprentis ». Son champ couvre les élèves et étudiants

[54] Voir par exemple, l'avant-projet de programme statistique 2006 présenté lors de la réunion du 20 avril 2005 de la formation Education du Cnis et le compte-rendu de la réunion. La mise en œuvre de la BNIE y est annoncée pour la rentrée 2005/2006 pour le 1er degré, sans visibilité précise pour les autres niveaux.

[55] À titre d'exemple, un communiqué du collectif Appel du collectif « STOP FICHAGE 13 » du 7 octobre 2008 est présenté en annexe 1. Le site internet de la ligue des droits de l'Homme de Toulon propose aussi un dossier sur le sujet, toujours actualisé (voir https://section-ldh-toulon.net/-le-ministere-de-l-EN-et-les-.html).

[56] Dans sa décision du 19 juillet 2010 (n°317182 & 323441), le Conseil d'État donne partiellement raison à la ligue des droits de l'Homme et au syndicat national unitaire des instituteurs, professeurs des écoles et PEGC de l'Isère sur la légalité de l'arrêté du 20 octobre 2008 créant la base Elèves du premier degré. Il reproche notamment à l'État l'impossibilité pour les parents de refuser l'enregistrement de données personnelles, même pour des motifs légitimes, et le déploiement trop précoce de la base, plusieurs mois avant la réception du récépissé de la Cnil. Dans une autre décision, relative à la BNIE (n° 334014), le Conseil d'État censure la durée de conservation des données alors fixée à 35 ans, durée qu'il juge excessive au regard des finalités déclarées du traitement. Cependant, le Conseil d'État valide les deux systèmes d'information qu'il considère important pour le bon fonctionnement du service public.

[57] Dépêche AFP du 12 juin 2008 s'appuyant sur une lettre adressée à la PEEP par le ministre, reprise sur le site internet de la ligue des droits de l'Homme de Toulon. Voir aussi rapport IGAENR n° 2015-054, *Adaptation des systèmes d'information à la gouvernance du premier degré et au pilotage des écoles*.

[58] « Cette durée, qui correspond à la somme de la durée de conservation des données du Système d'information du 1er degré (15 ans), du second degré (10 ans) et de l'enseignement supérieur (10 ans), est prévue pour permettre le suivi de l'ensemble des formations reçues par les élèves jusqu'à la sortie du système éducatif ainsi que les formations reçues dans renseignement supérieur, en intégrant d'éventuelles interruptions suivies d'une reprise d'études dans renseignement supérieur par exemple » (lettre du ministère de l'Éducation nationale à la Cnil du 8 février 2007).

[59] Avant-projet du programme statistique 2011 présenté à la commission Services publics du Cnis le 26 mai 2010.

[60] *Journal officiel* du 23 mars 2012.

inscrits dans les établissements d'enseignement scolaire, d'enseignement supérieur ou dans les centres de formation d'apprentis, qui relèvent des ministères en charge de l'Éducation nationale, de l'Enseignement supérieur, de l'Agriculture, de l'Apprentissage, de la Défense et de la Mer[61].

Le principe général de fonctionnement ne change pas : l'attribution d'un identifiant à un nouvel entrant ou la recherche de celui d'un élève ou étudiant déjà inscrit, s'appuie sur le nom de famille, le nom d'usage, les prénoms, le sexe, la date de naissance, le lieu de naissance[62], ainsi que le numéro d'identification du dernier établissement fréquenté, la date d'admission et la date de radiation de l'élève ou de l'étudiant dans le dernier établissement fréquenté. En cas de litiges, le traitement reste effectué au niveau académique.

Toute sortie du système éducatif couvert par le RNIE pendant plus de cinq ans entraine la radiation du répertoire. Un retour postérieur entraine donc la création d'un nouvel INE.

Pour les établissements relevant du ministère chargé de l'Éducation nationale, les élèves déjà immatriculés dans les bases académiques ont été intégrés au nouveau répertoire par la Depp à la rentrée 2015, avec un traitement des litiges par les services statistiques académiques. Peu de doublons ont été constatés lors de cette phase. À la rentrée 2016, le travail d'immatriculation a été mené en doublon, à la fois dans l'ancien système et dans le nouveau système. Depuis la rentrée scolaire 2017, le nouvel INE, adossé au RNIE, est pleinement en œuvre.

Ce déploiement s'est fait parallèlement au déploiement du système d'information statistique consolidé académique (SYSCA), qui remplace le système Scolarité et les BEA. Concomitamment, dans le premier degré, l'Outil numérique pour la direction d'école (ONDE[63]) a remplacé la BE1D. Désormais, l'INE attribué aux élèves dans le premier degré suivra les élèves dans le second degré puis dans le supérieur.

Pour l'apprentissage, l'utilisation du nouvel INE est en place depuis la rentrée 2018, après mise en place d'un portail web d'échange entre les CFA et la Depp. Pour les établissements relevant du ministère de l'Agriculture, l'utilisation du nouvel INE est effectif depuis la rentrée 2019[64]. Toutes les formations ne sont pas encore couvertes. L'utilisation du RNIE implique en effet une articulation des bases de gestion des établissements avec le répertoire, pour récupérer l'INE des élèves ou étudiants inscrits, ou demander l'immatriculation des nouveaux entrants qui n'en ont pas. Cette exigence complique plus ou moins le déploiement du nouvel INE. Pour les établissements relevant du ministère de l'Agriculture, il s'est agi d'ajuster leur système d'information commun. Pour l'apprentissage et l'enseignement supérieur, l'exercice est plus difficile du fait de la multiplicité des logiciels de gestion, donc des modèles de bases de données, utilisées.

2.1.4. L'articulation entre le second degré et l'enseignement supérieur devrait s'améliorer

Jusqu'à la mise en place du RNIE, l'étudiant devait transmettre lui-même le numéro d'identification qui lui avait été attribué dans l'enseignement secondaire à son premier établissement d'inscription dans l'enseignement supérieur. En fin de lycée, les futurs étudiants ont à connaître pour la première fois de ce numéro parce qu'ils en ont besoin pour accéder à la plateforme d'affectation post-baccalauréat, APB puis Parcoursup. À cet effet, l'INE figure sur le relevé de note des épreuves anticipées du baccalauréat et, souvent, sur les bulletins scolaires. À défaut, il peut être demandé à son établissement scolaire. Pour les étudiants qui ne disposent pas d'identifiant, par exemple parce qu'ils ont passé leur bac à l'étranger ou parce qu'ils ont connu une longue interruption de scolarité, c'est l'établissement de première

[61] Situation au 8 octobre 2020. Un arrêté modificatif du 1er septembre 2016 a ajouté les ministères chargés de la Défense et de la Mer dans la liste des tutelles couvertes.
[62] Mention du code commune pour les personnes nées en France ou indication du pays de naissance pour les personnes nées à l'étranger.
[63] La mise en place de l'outil numérique pour la direction d'école (ONDE) a conduit à modifier l'arrêté du 20 octobre 2008 créant la BE1D : changement du nom du système d'information du premier degré, réintroduction de le la profession des parents parmi les variables collectées, possibilité d'élargir le champ au Centre national d'enseignement à distance (CNED) et aux établissements français à l'étranger relevant de l'Agence pour l'enseignement français à l'étranger (arrêté du 13 janvier 2017).
[64] Cette mise en œuvre est concomitante au déploiement d'un nouveau système d'information de gestion des élèves pour les établissements d'enseignement agricole public et privé, Fregata, interfacé avec SYSCA pour la gestion de l'INE. Comme pour les établissements du ministère de l'Éducation nationale, l'ensemble des élèves déjà inscrits et disposant d'un identifiant spécifique à leur sphère (l'INA) ont été immatriculé au RNIE par la Depp préalablement à la bascule.

inscription dans le supérieur qui attribuait un INE en respectant certaines règles, conformément à une norme dite « base 36 ». L'INE de l'étudiant, qui doit notamment figurer sur ses certificats de scolarité, était ensuite supposé suivre l'étudiant pendant tout son parcours dans l'enseignement supérieur sur le territoire national.

À l'occasion de la mise en place du RNIE par la Depp, le Sies a développé le webservice INES[65] pour permettre aux services gestionnaires des établissements de l'enseignement supérieur d'échanger des informations relatives à leurs nouveaux étudiants inscrits avec le RNIE, afin de certifier l'INE des étudiants qui en disposent et immatriculer ceux qui n'en disposent pas. Ces nouveaux protocoles permettent d'interrompre le système « base 36 » d'attribution décentralisée d'un INE pour les étudiants n'en disposant pas. L'accès à INES concerne les établissements qui relèvent du champ SISE[66]. Il est aussi ouvert aux autres établissements de l'enseignement supérieur qui le souhaitent. Le dispositif a été mis en service fin 2020.

À la différence des champs couverts par la Depp, pour lesquelles tous les élèves ont basculé dans le nouveau répertoire, il n'y a pas de bascule globale dans le nouveau système pour les étudiants de l'enseignement supérieur. Cette bascule s'effectue au fil de l'eau. Les étudiants déjà inscrits conservent leur INE antérieur, issu des BEA ou de la norme dite « base 36 ». En revanche, à partir de la rentrée 2018, les nouveaux bacheliers qui s'inscrivent pour la première fois dans le supérieur disposent du nouvel INE unique certifié, puisqu'ils ont basculé dans le nouveau système quand ils étaient dans le secondaire. Depuis la mise en place du webservice INES, les nouveaux étudiants, qui, du fait de leur parcours antérieur, ne disposent pas déjà d'un INE, basculent aussi dans le nouveau système.

2.1.5. Perspectives pour le Céreq

Les possibilités techniques théoriques ouvertes par le déploiement de l'INE se heurtent jusqu'à présent à deux principales difficultés pour reconstituer les parcours scolaires. La première difficulté tient au champ couvert par l'INE. La seconde tient à son unicité.

Concernant le champ, une partie des établissements scolaires et d'enseignement supérieur ne sont pas couverts. L'obligation de l'usage de l'INE a d'abord été limité aux établissements relevant du ministère de l'Éducation nationale, avant de s'étendre au-delà. Son extension s'est notamment heurtée à des difficultés techniques pour les établissements libres de choisir leurs systèmes d'information car l'adossement au répertoire national d'immatriculation des élèves et étudiants suppose que les applications de gestion administrative soient en mesure d'échanger avec ce répertoire. Dans ces conditions, les parcours qui peuvent être reconstitués à partir des données administratives disponibles à la Depp et au Sies sur la base de l'INE, comportent des trous d'observation dont on ne peut dire s'ils sont liés à des arrêts de scolarité, des scolarités à l'étranger ou des passages dans des établissements localisés en France mais hors champ. Ces trous tendent cependant à se réduire avec le temps avec l'extensions du champ couvert par l'INE et des remontées de données individuelles aux services statistiques ministériels concernés.

L'utilisation d'un même numéro d'immatriculation pour un élève donné tout au cours de son parcours scolaire est un autre impératif pour pouvoir reconstituer l'intégralité des parcours scolaires. Cette unicité de l'INE n'est pas assurée avant la mise en place de l'INE certifié par l'adossement des systèmes de gestion au répertoire national des identifiants élèves, étudiants et apprentis. Une première difficulté tient au fait que l'identifiant attribué dans l'enseignement du premier degré ne suivait pas l'élève lors de son entrée dans le secondaire. Cette discontinuité est sensée disparaître avec la mise en place du RNIE. Le caractère trop décentralisé de la gestion de l'INE accentuait la difficulté puisque les élèves changeant d'académie dans le secondaire pouvaient se voir attribuer un nouvel INE dans leur nouvelle académie. Cela pouvait aussi arriver pour des étudiants en multi-inscription ou changeant d'établissements dans le supérieur. Aucun chiffrage n'est cependant disponible sur l'ampleur réel de ces immatriculations multiples.

[65] Voir présentation à la réunion du 4 octobre 2016 de la commission Services publics du Cnis.
[66] Le système d'information sur le suivi de l'étudiant (SISE) est présenté partie 4.2.

À l'avenir, un meilleur adossement des systèmes de gestion au RNIE doit garantir l'attribution d'un INE certifié unique et sa diffusion progressive à l'ensemble du champ éducatif. Cette évolution, combinée à l'extension du champ des établissements couverts et l'utilisation de l'INE dans de nombreux outils de gestion, permet d'envisager de rendre effective à moyen terme la possibilité de mieux suivre les parcours à partir d'appariements de données administratives individuelles, rapprochées à partir de l'INE.

Pour le Céreq, cette évolution peut être une opportunité importante d'amélioration de son dispositif d'enquêtes « Génération », collectées régulièrement dans le cadre de la statistique publique, pour observer le déroulement d'insertion sur le marché du travail des jeunes en fonction de leur parcours de formation initiale. La mobilisation des données administratives sur les parcours scolaires peut permettre en effet d'envisager trois axes de progrès :

- l'allègement de la charge d'enquête pesant sur les enquêtés par une réduction de la taille du questionnaire, certains sujets abordés étant couverts par les sources administratives ;
- l'enrichissement des informations disponibles pour les répondants, par l'apport d'informations plus précises sur les parcours scolaires, que le questionnaire ne pouvait aborder en détail pour des raisons de durée de questionnement ;
- une amélioration du traitement de la non-réponse par la disponibilité d'informations administratives sur les parcours des personnes échantillonnées mais non répondantes (en préalable, il faudrait examiner les éventuelles différences de profil entre les répondants et les non répondants).

Ces potentialités ne seront pleinement mobilisables que dans plusieurs années, après le plein déploiement de l'INE unique certifié, quand la mobilisation des informations administratives sur les parcours scolaires sera possible pour l'ensemble des jeunes enquêtés dans le dispositif Génération, qu'ils sortent du secondaire sans diplôme, d'un master ou d'un doctorat. Cependant, il peut être utile d'entamer dès aujourd'hui des expérimentations ; d'abord, pour s'approprier les bases de données administratives et leurs subtilités liées aux règles de gestion ; ensuite, pour commencer à examiner les apports possibles de ces sources, en se limitant pour commencer aux niveaux de sortie et aux catégories d'établissements les mieux couverts. Les deux expérimentations initiées par le Céreq en 2021, d'un rapprochement de l'enquête Génération 2017 avec, d'un côté les fichiers de la Depp présentés dans la partie 4.1, et de l'autre, ceux de l'enseignement supérieur présentées dans la partie 4.2, vont dans ce sens.

Ces réflexions devront tenir compte de la mise en œuvre à venir de l'équipement d'excellence « Innovation, Données et Expérimentation en Education » (IDEE), retenu dans le cadre du troisième programme[67] des « investissements d'avenir » (voir encadré 3). L'axe 1 de cet équipement, porté par la Depp, vise en effet la mise en place d'un centre sécurisé permettant d'accéder aux données administratives relevant de l'Éducation nationale, à des fins de recherche. Même si les thématiques d'étude du Céreq ne constituent pas forcément le cœur du projet IDEE, très tourné vers les sciences de l'éducation, la mise en place de cet outil pourrait influencer l'organisation des rapprochements utilisant des données administratives scolaires, voire permettre de disposer de nouvelles données, en fonction de ses contours qui restent à définir. Par exemple, les données versées à des fins de recherche seront-elles exhaustives ou sur des échantillons, sélectionnés par exemple sur leur date de naissance comme dans plusieurs panels de la statistique publique ? Est-ce que les données de l'enseignement supérieur auront vocation à intégrer le dispositif ? Est-ce que des informations sur les débuts de trajectoire professionnelles seront proposées, à partir de données administratives telles que la déclaration sociale nominative ou les fichiers de Pöle emploi[68] ?

[67] Ouvert en janvier 2020, l'appel à projet porté par l'Agence nationale de la recherche a été clos le 19 juin 2020 et ses lauréats désignés le 18 décembre 2020. Selon les termes de cet appel à projet, « EquipeEx+ » vise à soutenir des équipements structurants pour la recherche d'envergure nationale. Ils doivent être proposés par des établissements d'enseignement supérieur ou de recherche avec une priorité pour les équipements fortement mutualisés. Ces équipements doivent être ouverts à toutes les communautés scientifiques concernées ainsi qu'aux entreprises, sur la base d'une tarification permettant d'assurer le fonctionnement, la mise à jour et, en partie, le renouvellement de ces équipements. Les financements des investissements d'avenir doivent permettre l'acquisition initiale des équipements et, le cas échéant, une contribution à leur maintenance. Sauf exception, ils ne doivent pas couvrir les frais de fonctionnement.

[68] Cette perspective pourrait être une extension logique du déploiement du dispositif Inserjeunes, présenté dans la partie 4.3, qui vise à produire des indicateurs d'insertion à six, douze, dix-huit et vingt-quatre mois sur les sortants des formations professionnelles par l'apprentissage ou par la voie scolaire du niveau CAP au BTS, en s'appuyant pour l'instant sur la seule déclaration sociale nominative.

Encadré 3 • L'equipex « Innovation, Données et Expérimentation en Education » (IDEE)

Le projet IDEE résulte d'une initiative conjointe de l'école normale supérieure de Paris et du laboratoire J-PAL de l'école d'économie de Paris. Il reprend un projet plus ancien porté par Stanislas Deheane et Esther Duflo, dans le cadre du conseil scientifique de l'Éducation nationale. Il vise à faciliter la recherche expérimentale en éducation, notamment dans le cadre méthodologique porté par le J-PAL. Il comprend trois axes.

- La mise en place d'un système permettant d'accéder de manière sécurisée et à distance aux données administratives de la Depp, notamment pour pouvoir apparier des bases de données individuelles. Les principaux progiciels statistiques seraient disponibles sur ce serveur. Des « identificateurs uniques individuels » seraient créés pour faciliter les appariements de données. Un effort serait mené sur la documentation de ces données.
- La construction, la validation et le partage d'un ensemble de protocoles de mesure facilement réplicables et la mise à disposition, en location, d'équipements d'observations et de mesures, comme par exemple, un jeu de tablettes pour les enquêtes sur gros échantillons, des casques d'électroencéphalogramme, des dispositifs de suivi des yeux, etc.
- L'organisation et l'animation de rencontres entre les porteurs de projets, les chercheurs et les équipes pédagogiques.

Six institutions portent formellement le projet : le département d'études cognitive de l'école normale supérieure de Paris, le J-PAL, Neurospin (grande infrastructure de recherche spécialisée dans l'imagerie cérébrale), le Laboratoire de Recherche sur les Apprentissages en Contexte (LaRAC) de l'Université Grenoble-Alpes, le Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP/OSC) de Science Po Paris et le CNRS.

Cette dernière perspective n'est pas la plus évidente faute d'identifiant commun entre les données individuelles administratives scolaires, qui mobilisent l'INE, et les données individuelles administratives relatives au marché du travail, qui utilisent le NIR. Le rapprochement des données de deux sphères implique donc de repartir des informations d'état civil pour procéder à la construction d'une table de passage entre les données des deux sphères, en respectant une organisation et des conditions de sécurité strictes. Cette complexité pourrait cependant être un argument pour justifier que ce travail soit effectué de façon mutualisée par la statistique publique, pour que les producteurs publics de statistiques et les chercheurs puissent ensuite en profiter dans le cadre d'un accès sécurisé aux données.

2.2. Dans le champ de la formation continue, le projet « Agora » doit permettre d'articuler les informations individuelles issues des différentes parties prenantes[69]

La formation professionnelle fait intervenir de nombreux acteurs : des prescripteurs, des financeurs et les organismes de formation. Chacun dispose de ses systèmes d'information, avec des interconnexions limitées et des saisies souvent multiples. Cette variété ne facilite pas non plus le suivi national des politiques publiques de formation. L'ambition affichée du projet « AGORA » est de remédier à cela en permettant l'échange des informations entre les différents acteurs, d'abord à des fins de gestion, ensuite pour le suivi, le pilotage et l'évaluation des politiques publiques de formation professionnelle. Il doit permettre, par exemple, de répondre aux questions suivantes : quelles actions de formation ont été réalisées au bénéfice de quels publics ? Quels ont été les impacts de ces formations sur l'accès à l'emploi et les parcours professionnels ?

[69] Voir notamment :
- https://www.moncompteformation.gouv.fr/espace-public/sites/default/files/2020-05/AGORA%20-%20Cadre%20de%20r%C3%A9f%C3%A9rence%20-%2026%20mai%202020.pdf (cadre de référence fixé par la DGEFP en mai 2020)
- https://www.ccomptes.fr/sites/default/files/2018-07/20180704-formation-demandeurs-d-emploi.pdf (partie « Systèmes d'information » du rapport de la Cour des comptes de mai 2018 sur la formation professionnelle)
- https://www.cpformation.com/agora/ (article « Tous sur le projet Agora », 27 mars 2018).

Ce projet avance lentement, notamment en raison des évolutions du contexte institutionnel de la formation professionnelle, qui a fortement changé ces dernières années, mais aussi en raison du nombre de systèmes d'information à mettre en relation. Au sein du ministère chargé de l'Emploi, la Délégation générale à l'emploi et à la formation professionnelle (DGEFP) assure la maîtrise d'ouvrage du projet, dont la maîtrise d'œuvre a été confiée à la Caisse des dépôts, par extension de ses responsabilités initiales sur le compte personnel de formation (CPF). La Dares, service statistique du ministère du Travail, fait partie des destinataires des données à des fins statistiques. Elles doivent notamment lui permettre de remplacer sa base régionalisée des stagiaires de la formation professionnelle (« BREST »)[70]. Les textes prévoient par ailleurs que les chercheurs peuvent accéder aux données en conventionnant avec la Dares.

2.2.1. Un peu d'historique

La réforme de la formation professionnelle 2014, entrée en vigueur le 1er janvier 2015, met en place plusieurs nouveaux dispositifs : l'entretien professionnel, le conseil en évolution professionnel et le compte personnel de formation (CPF), qui se substitue au droit individuel à la formation (DIF). La gestion du CPF est confiée à la Caisse des dépôts et consignation. Pour cela, celle-ci met en place un système d'information dédié avec des financements de la DGEFP. Opérationnel depuis le 1er janvier 2015 et destiné à la fois aux titulaires des comptes et aux financeurs, il prend la forme d'un portail donnant accès à plusieurs fonctionnalités : un site d'information, la gestion des listes de certifications éligibles au CPF, la gestion des opérations de mobilisation du compte en débit et en crédit, ainsi que l'accès au dossier des formations suivies dans le cadre du CPF.

En juillet 2015, la commission Systèmes d'Informations du Conseil national pour l'emploi, la formation et l'orientation professionnelles (Cnefop)[71] demande la création d'un système d'échanges d'informations, qu'il intitule « Agora ». L'idée est d'obliger les organismes de formation à l'utiliser pour informer les financeurs des entrées et des sorties de formation et pour permettre aux différents acteurs de la formation professionnelle de partager ces données entre eux. Dans la foulée, la DGEFP lance une étude d'opportunité pour une plateforme sur les « entrées et sorties de la formation professionnelle ».

Cette volonté est inscrite dans le code du travail par la loi El Khomri du 8 août 2016. Celle-ci crée l'article L. 6353-10 qui donne un cadre juridique à la création de cette plateforme, en prenant appui sur son utilité pour la mise en œuvre du CPF. Cet article dispose ainsi que :

> *« Les organismes de formation informent les organismes qui financent la formation, dans des conditions définies par décret, du début, des interruptions et de l'achèvement de la formation, pour chacun de leurs stagiaires, et leur communiquent les données relatives à l'emploi et au parcours de formation professionnelle dont ils disposent sur ces stagiaires. »*
>
> *« Les organismes financeurs, l'organisme gestionnaire du système d'information du compte personnel de formation mentionné au III de l'article L. 6323-8 et les institutions et organismes chargés du conseil en évolution professionnelle mentionnés à l'article L. 6111-6 partagent les données mentionnées au premier alinéa du présent article, ainsi que celles relatives aux coûts des actions de formation, sous forme dématérialisée et dans des conditions définies par décret en Conseil d'État pris après avis de la Commission nationale de l'informatique et des libertés. »*

La loi El Khomri met aussi en place le compte personnel d'activité (CPA), qui doit permettre aux actifs de suivre leur compte personnel de formation, leur compte de prévention de la pénibilité et leur compte d'engagement citoyen. Rattaché directement au salarié quel que soit son statut, le CPA assure que ses droits le suivent tout au long de son parcours professionnel. C'est encore la Caisse des dépôts et consignations qui est désigné comme opérateur pour en assurer la gestion, *via* un portail et une plate-forme numérique développés avec des financements de la DGEFP et du programme des « investissements d'avenir ».

[70] BREST est une base de données sur la formation professionnelle des personnes en recherche d'emploi produite par la Dares depuis 2003, à partir des fichiers de gestion de rémunération ou de protection sociale des stagiaires. Chaque observation de la base correspond à un stagiaire rémunéré, qu'il le soit par l'État, par un conseil régional ou par Pôle emploi (pour son propre compte, celui de l'Unédic ou celui de l'État). Si le stagiaire n'a pas droit à une rémunération, il figure néanmoins en tant que bénéficiaire d'une protection sociale. Pour en savoir plus, voir par exemple, l'encadré présentant la source dans Noémie Cavan (2015), « La formation professionnelle des personnes en recherche d'emploi en 2013 », *Dares Analyse*, n°30, avril.

[71] Le Cnefop a été supprimé et ses missions absorbées par France compétences à sa création, le 1er juillet 2019.

En mars 2017, un rapport de l'Inspection générale des affaires sociales[72] recommande d'élargir les ambitions du projet AGORA pour qu'il devienne un « hub » d'échange d'informations « qui simplifierait les démarches administratives d'entrées et sorties en formation ou de suivi des stagiaires », et qui « éclairerait, aussi, significativement, la connaissance des trajectoires de formation ». La même année, l'État demande à la Caisse des dépôts et consignations de faire converger les portails du CPA et du CPF en un portail unique à l'horizon du premier trimestre 2018 pour aboutir à la plateforme d'échanges demandée par la loi El Khomri. Le projet « Agora » se met ainsi progressivement en place, juridiquement en tant que système d'information du CPF, le « SI CPF ».

La loi du 5 septembre 2018 pour la liberté de choisir son avenir professionnel transforme en profondeur le CPF, notamment en passant de droits en heures à des droits en euros. Elle conforte la Caisse des dépôts comme gestionnaire du dispositif qui en assure, seule, les différentes composantes, de l'inscription des titulaires des droits à formation au paiement des organismes de formation**.** Elle oblige à une évolution du « SI CPF ». Les bénéficiaires du CPF doivent pouvoir disposer d'un accès direct à l'offre de formation et gérer en totale autonomie leur dossier de prise en charge. Cette nouvelle version du service dématérialisé doit ainsi permettre aux titulaires de comptes de s'inscrire directement aux formations qu'ils auront choisies et de payer les organismes de formation avec les droits dont ils disposent, sans mobiliser d'intermédiaire.

Encadré 4 • Quelques précisions sur le cadre juridique d'Agora (système d'information du CPF)

Les objectifs et les ayants droit de la plateforme sont définis dans la partie règlementaire du code du travail. Ils sont précisés par l'arrêté du 11 octobre 2019 relatif à la mise en œuvre du traitement automatisé de données à caractère personnel dénommé « Système d'information du compte personnel de formation » :

- son annexe 1 liste les données à caractère personnel pouvant être enregistrées ;
- son annexe 2 liste les organismes « contributeurs », dont les agents sont habilités à accéder aux données à caractère personnel ;
- son annexe 3 liste les organismes « consommateurs », dont les agents sont habilités à être destinataires des données à caractère personnel ;
- son annexe 4 liste les traitements automatisés pouvant alimenter le SI-CPF et ceux pouvant être mis en relation avec lui.

Ce cadre juridique prévoit notamment que l'identification des personnes bénéficiaires des formations s'appuie sur leur nom, prénom et date de naissance mais aussi, *in fine,* sur leur Nir.

Outre les informations relatives aux prescriptions de formation, à leur réalisation, à leur financement et à leurs bénéficiaires, ces textes permettent l'intégration d'informations sur « la situation d'emploi » à 6, 12 et 18 mois après la formation. À ce stade, cependant, ils ne prévoient pas que des données administratives puissent être utilisées pour caractériser cette situation d'emploi, que ce soit le fichier historique de Pôle emploi ou la déclaration sociale nominative.

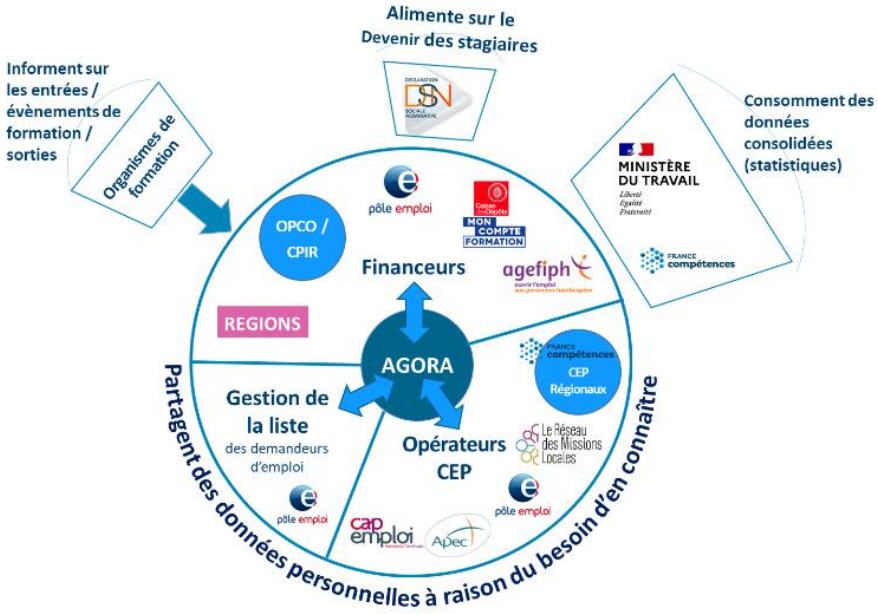
Parmi les destinataires des données, la Dares et les « organismes qu'elle mandate au moyen de conventions de recherche » peuvent « exploiter les données à des fins statistiques destinées à la recherche ou à l'évaluation du SI-CPF » et les utiliser pour « réaliser des enquêtes et des études statistiques sur l'emploi, le travail et la formation professionnelle en France afin d'éclairer la conception et la mise en œuvre des politiques publiques dans ces domaines, notamment par le suivi et l'évaluation des résultats des politiques menées ». France compétences bénéficie également d'un accès assez large aux données dans le cadre de ses missions. La DGEFP, en revanche, ne peut accéder qu'à des « indicateurs » de gestion et de pilotage.

[72] Nicolas Amar et Anne Burstin (2017), *La transformation digitale de la formation professionnelle continue*, rapport de l'IGAS n° 2016-055R.

2.2.2. Une plateforme articulant les systèmes d'information des parties prenantes

La plateforme Agora ne se substitue pas aux systèmes d'information des différents acteurs, mais s'interface avec eux pour leur permettre de mettre en œuvre leurs obligations d'informations. Elle leur permet aussi de recueillir les informations disponibles déjà déposées par d'autres acteurs en s'inscrivant dans la logique du « dites-le nous une fois », qui facilite les échanges d'information au motif d'éviter les saisies multiples. C'est donc une sorte de « hub » de données où les organismes contributeurs intègrent des informations sur la plateforme au fil de leurs actions (voir schéma 3). Par ailleurs, des « consommateurs », comme les financeurs, la Dares ou France compétences accèdent à certaines informations à des fins de suivi, de pilotage ou d'évaluation.

Schéma 3 • La structuration d'Agora



Source : ministère du Travail, cadre de référence d'Agora, version de mai 2020.

Concrètement**,** l'action de formation constitue l'observation de base. Cette action de formation est caractérisée par son prescripteur, son bénéficiaire, sa nature, son coût, sa temporalité et le devenir du stagiaire à son issue (voir encadré 5). Toutes les actions de formation professionnelle prises en charge par les organismes financeurs de la formation professionnelle, qu'elles soient financées en totalité ou seulement partiellement, doivent normalement entrer dans son champ, quel que soit le statut de la personne, qu'elle soit en activité ou à la recherche d'un emploi. L'intégration des différentes parties prenantes est cependant progressive puisqu'elle implique que les systèmes de gestion des différents acteurs soient en mesure de dialoguer avec Agora. Pour le moment, le système couvre ainsi plutôt bien le CPF et les formations des personnes en recherche d'emploi. Les remontées des dossiers gérés par les régions sont encore mal couvertes[73] même si des efforts sont en cours dans le cadre du plan d'investissement dans les compétences pour finaliser cette extension de champ. Les échéances semblent moins certaines pour l'alternance, la formation des salariés financés par les OPCO, l'accès à la certification et les conseils en évolution professionnelle.

[73] Voir l'encadré sur les « données mobilisées » in Balmat C. & Corazza E. (2020), « Le compte personnel de formation en 2018 », *Dares Résultats*, n°9, février : « Les dossiers gérés par les conseils régionaux sont plus difficiles à suivre. Les systèmes d'information des régions diffèrent beaucoup entre régions et au sein d'une même région, entre types de formation et, parfois, entre les anciennes régions qui la composent. Certaines régions ne collectent pas systématiquement les informations nécessaires à la complétude du système d'information du compte personnel de formation (SI-CPF), qui nécessite notamment une traçabilité nominative des bénéficiaires ».

Pour alimenter le « hub », un tronc commun de variables a été défini, avec les référentiels à utiliser pour les renseigner. Des « règles d'alimentation » ont également été fixées pour déterminer quel type de partenaires peut intervenir à chaque évènement du cycle de vie d'un dossier et quelles informations chacun doit, ou peut selon les cas, apporter, modifier ou supprimer.

Si le financeur de l'action de formation est unique, celui-ci est l'unique contributeur à Agora qu'il alimente au fur et à mesure du cycle de vie du dossier. En pratique, le financeur peut déléguer tout ou partie de la gestion d'un dossier. Le système d'information distingue donc le déléguant du délégataire, qui « apporte » l'information. En cas de cofinancements, l'un des financeurs est désigné fournisseur « pivot ». Dans ce cas, ses données priment sur celles fournies par les autres, à l'exception des données de coûts, qui restent à la seule main de chacun des financeurs concernés. En pratique, le système n'empêche pas les contributeurs non « pivot » d'apporter leurs données, mais, elles ne seront retenues par Agora.

Encadré 5 • Les informations prévues dans Agora

Selon les spécifications disponibles au troisième trimestre 2020, le tronc commun d'Agora prévoit de collecter les informations suivantes :

- ***L'identité du déclarant et celle du déléguant*** (conseil régional, OPCA, opérateur de rémunération, acteur du CEP, Pôle emploi, Agefiph) : l'information est obligatoire, par la mention du SIRET (non vérifié) et d'un libellé sur listes.
- ***L'identification du dossier de formation*** : l'identifiant unique généré par la Caisse des dépôts lors de l'initialisation du dossier (« numéro AGORA ») est obligatoire dans tous les flux, afin de pouvoir relier entre elles toutes les informations d'un même dossier. En complément, une autre variable est prévue de façon facultative pour permettre au déclarant de mettre son propre identifiant de gestion s'il le souhaite.
- ***Le statut du dossier de formation*** : obligatoire, cette information indique le statut du dossier (en cours d'élaboration, validé, rejeté, entrée en formation, sortie en formation, clos après réalisation partielle, après réalisation totale ou non-réalisation). Le statut détermine le caractère obligatoire ou non de certaines informations.
- L'i***dentité du bénéficiaire :*** obligatoire, elle est renseignée par les noms, prénoms, sexe et date de naissance du bénéficiaire, ainsi que son Nir (ou numéro d'identification d'attente pour les étrangers sans Nir dans l'attente de sa certification). L'articulation avec le SNGI permet de contrôler l'existence de la personne dans le répertoire, qui alimente ensuite le hub avec ses informations d'état civil certifiées.
- ***Le bénéfice d'une obligation d'emploi travailleur handicapé (BOETH)*** : cette information est obligatoire mais, pour le moment, elle n'est que déclarative et sans contrôle externe.
- ***Les moyens de contacter le bénéficiaire*** (adresse, courriel, téléphone). Pour ces informations, c'est la dernière information fournie qui est retenue. Les adresses doivent respecter les formats de la base d'adresses nationale (BAN) pour faciliter les suivis effectués par zonage, dont les quartiers prioritaires de la politique de la ville. Un référentiel permet d'attribuer le Code Insee de la commune de domiciliation sur la base du code postal et du libellé de commune.
- ***La situation de la personne à l'entrée en formation*** : obligatoire, cette information prend deux modalités : « en recherche d'emploi » ou « en activité ». D'autres variables sont renseignées ensuite selon la modalité prise. Pour les personnes « en recherche d'emploi », une variable complémentaire obligatoire permet de préciser leur situation d'inscription à Pöle emploi (inscrite, non inscrite, inconnue). Pour les personnes « en activité », une variable complémentaire obligatoire permet de préciser le statut du bénéficiaire (salarié, fonctionnaire, contractuel de droit public, travailleur en ESAT, intermittent du spectacle, non salarié, étudiant, bénévole, non connu). D'autres variables complémentaires sont obligatoires pour les personnes déclarées « salariées » : la catégorie socioprofessionnelle, le type de contrat (CDI, CDD, CSP, contrat d'apprentissage, contrat de professionnalisation, intérimaire), le fait que l'employeur est un particulier ou non et, s'il n'est pas un particulier, son SIRET, sa raison sociale, sa commune d'implantation (l'adresse complète est facultative).

- ***L'identification de la formation*** : l'intitulé « littéraire » est obligatoire. Le recours à un « code Offre info » (par les financeurs) ou « EdOF » (par les offreurs de formations) sont facultatifs. Le domaine de la formation doit être renseigné, soit selon la nomenclature des spécialités de formation (NSF) en 100 sous-domaines, soit selon Formacode (thesaurus de l'offre de formation du Centre info) en 63 champs. Les objectifs de la formation doivent aussi être précisées selon des modalités prédéfinies. L'articulation aux référentiels d'Offre Info et EdOF doit permettre de préremplir certains champs une fois la formation identifiée.
- ***Si la formation a une dimension certifiante***, le code Certif info de la certification visée doit être mentionné, ce qui permet de récupérer son intitulé et son niveau grâce au référentiel Certif info. Si le bénéficiaire obtient la certification, cela doit être mentionné.
- ***Les modalités de formation*** : les modalités pédagogiques de la formation doivent être précisées (à distance, en présentiel, mixte // collective ou individuelle), de même que sa durée (nombre d'heures total affiché par l'offreur de formation, nombre d'heures prévisionnel de la formation pris en charge, nombre d'heures effectivement pris en charge).
- ***La temporalité de la formation*** : les dates d'entrée et de sortie, prévisionnelles et effectives, sont obligatoires, de même que le motif de sortie selon une liste préétablie (fin prévue de l'action, retour à l'emploi avant la fin prévue, entrée dans une autre formation, abandon, etc.).
- ***L'offreur de formation*** : il est obligatoire de renseigner le fait que la formation est réalisée en interne ou non et, si elle est réalisée en externe, le SIRET de l'organisme de formation.
- ***Les coûts de la formation*** : de nombreuses variables sur les coûts et les financements sont obligatoires. Elles permettent d'identifier le dispositif principal de formation assurant son financement, l'ensemble des financeurs et les montants pris en charge.
- ***Le devenir du stagiaire*** : il est possible d'indiquer la situation d'activité du stagiaire (« salarié », « en emploi non salarié », « en recherche d'emploi ») à 3 mois, 6 mois et 12 mois. Très logiquement, une piste de réflexion porte sur l'utilisation de la déclaration sociale nominative et du fichier de Pöle emploi pour fournir une partie de ces informations.

La plateforme intègre des contrôles qualité sur le flux d'information arrivante et des contrôles de cohérence intertemporelle. Certains contrôles entraînent un rejet du flux, qui doit donc être corrigé pour être pris en compte. C'est notamment le cas quand les anomalies portent sur la structuration des données transmises, le respect des formats attendus et certaines incohérences internes (date d'entrée avant la date de naissance, par exemple). D'autres contrôles ne génèrent qu'une alerte du fournisseur qui peut ne pas en tenir compte (par exemple lorsque les ordres de grandeurs de coûts ou de durées sont jugés anormaux par rapport à des seuils prédéfinis) : comme pour tout système d'information centralisé, des arbitrages doivent être faits entre la fiabilité et la comparabilité des données, d'une part, et les opérations de gestion qu'il ne faut pas trop complexifier pour les services gestionnaires, d'autre part.

Des données externes peuvent aussi être mobilisées pour renseigner certaines variables, comme :
- le référentiel SIRENE, pour fiabiliser les informations sur les entreprises formatrices et les organismes de formation ;
- la base Offre Info des Carif-Oref relative à l'offre de formation professionnelle, notamment celles financées par les Régions, l'État, Pôle emploi, les Opca, Opacif ou Agefiph ;
- la base Certif Info relative aux diplômes et certifications reconnues.

Agora s'appuie aussi sur le système national de gestion des identifiants géré par la Caisse nationale d'assurance vieillesse pour s'assurer de la bonne identification de la personne bénéficiaire par la récupération de son numéro d'immatriculation au répertoire des personnes physique, le Nir.

2.2.3. Perspectives pour le Céreq

Compte tenu de ses missions, le Céreq devrait être conduit à s'appuyer régulièrement sur les données administratives contenues dans Agora, en articulation avec les travaux propres de la Dares et de France compétences, pour des exploitations statistiques diverses à des fins d'études et d'évaluations nationales, sectorielles ou territoriales. Dans le cadre strict des textes en vigueur, de tels travaux

nécessitent cependant de conventionner avec la Dares ou s'effectuer en sous-traitance pour France compétences. Des sous-traitances devraient être aussi possibles avec les différents autres acteurs, dans le cadre de leurs droits d'accès, qui sont moins larges. La possibilité d'accéder aux données dans le cadre de l'exception de recherche et de production statistique prévu au RGPD pourrait également être examinée, mais *in fine*, l'accès effectif suppose l'accord *de facto* des ayants droit prévus par les textes.

La formulation retenue pour définir les finalités de l'accès de la Dares dans l'arrêté du 11 octobre 2019 suggère que les travaux menés sous conventionnement avec elle peuvent concerner des appariements avec des sources externes de données d'entreprises, par exemple pour récupérer des informations financières, économiques ou productives sur les entreprises des salariés formés, ou pour relier les informations enregistrées dans Agora à des enquêtes auprès des entreprises, interrogées par exemple sur leur politique et leurs pratiques de formation. Dans ce même cadre, des extraits d'Agora devraient pouvoir servir de base de sondage, après des entreprises ou des bénéficiaires à des fins d'enquête. Il devrait également être possible de rapprocher des données individuelles extraites d'Agora avec d'autres sources individuelles externes. Dans la mesure où un conventionnement avec la Dares est indispensable pour mener ce type de travaux et parce que ses moyens sont contraints, il est plus réaliste de penser que les opportunités d'articuler des extraits d'Agora à d'autres sources ou de les utiliser comme base de sondage devraient être surtout examinées dans le cadre d'appels à projets de la Dares ou pour la production d'enquêtes relevant de la statistique publique. Aussi, l'accent devrait être mis prioritairement, dans un premier temps au moins, sur des projets d'études exploitant directement et uniquement des extraits d'Agora.

Encadré 6 • L'enquête de la Dares auprès des stagiaires de la formation professionnelle

La Dares a mis en place en 2019 une enquête « Post-Formation » pour recueillir l'avis des stagiaires de la formation professionnelle sur le déroulement de leur formation[74]. Cette enquête doit permettre de produire des indicateurs qu'Agora devrait pouvoir fournir à terme, comme le taux d'accès effectif à la certification, le taux d'abandon, etc. Elle doit aussi permettre d'aborder des points de vue subjectifs absents des données administratives, car, au-delà de la question de l'insertion professionnelle, il s'agit aussi de mieux comprendre les motivations et les attentes des personnes formées.

Cette enquête est trimestrielle (1ère vague auprès des sortants de formation au quatrième trimestre de 2018). Le questionnaire de 6 pages est adressé par courrier avec une enveloppe pré-affranchie mais l'enquêté peut aussi faire le choix de répondre par internet. Toutes les personnes sorties de formation professionnelle ayant bénéficié de cette formation au titre de leur recherche d'emploi sont dans son champ, que la formation ait été à son terme ou pas. L'enquête a lieu six à neuf mois après la sortie. L'échantillon est tiré aléatoirement dans les fichiers administratifs des stagiaires de formation professionnelle, avec l'objectif d'atteindre 120 000 répondants par an, soit 30 000 par trimestre, avec un taux de réponse de l'ordre de 35 %. Cette enquête n'est pas labellisée « statistique publique ».

[74] Voir https://dares.travail-emploi.gouv.fr/dares-etudes-et-statistiques/enquetes/article/enquete-post-formation

3. Les premiers panels sociodémographiques alimentés par des données administratives ou pseudo-administratives

Dès la fin des années 1960, l'Insee a mis en place deux premiers grands panels dans la sphère des statistiques démographiques et sociales : l'échantillon démographique permanent et un panel de salariés, dont le champ s'est progressivement élargi jusqu'à devenir aujourd'hui un panel « Tous salariés ». Ces deux panels sont alimentés en tout ou partie par des sources administratives ou pseudo administratives[75]. Outre leur ancienneté, ces panels ont joué un rôle pivot, qu'ils conservent, dans la mise en place d'autres panels thématiques alimentés par les sources administratives. En effet, en articulant leur échantillonnage sur ceux des deux premiers panels, les nouveaux dispositifs peuvent profiter des informations que ces deux premiers panels collectent, en particulier le niveau de formation et la structure familiale, pour l'échantillon démographique permanent, et une partie des trajectoires professionnelles pour le panel « Tous salariés ». Il est donc intéressant de s'arrêter un instant sur ces deux panels. Parce qu'il a aussi un rôle pivot pour le suivi de certains aspects des trajectoires professionnelles, nous présenterons aussi le fichier historique des demandeurs d'emploi mis en place une trentaine d'années plus tard par l'agence nationale pour l'emploi (ANPE), en partenariat avec la Dares, à partir des données de gestion de l'agence et de l'Unédic[76].

La présentation de ces trois panels sera aussi l'occasion d'illustrer au fil de l'eau certaines difficultés que peut présenter l'usage des données administratives et des panels qu'elles alimentent, en particulier les points suivants :

- la définition d'une situation ou d'une réalité administrative peut différer du concept statistique que l'on cherche à approcher[77] ;
- le champ couvert par la source administratives ne couvre pas forcément l'ensemble du champ souhaité par le statisticien ;
- certaines informations collectées administrativement sont plus fiables que d'autres, parce qu'elles sont plus centrales au processus de gestion, donc plus contrôlées ;
- certaines informations pertinentes pour les analyses, voire nécessaire, ne sont pas disponibles dans les sources administratives mobilisées[78] ;
- l'évolution des procédures administratives, mais aussi de la charge de travail et des priorités du producteur du panel peut induire des ruptures dans les suivis de trajectoires.

[75] Les « bulletins d'état civil », qui alimentent l'échantillon démographique permanent, sont remplis dans le cadre de démarches administratives qui permettent notamment d'assurer la gestion du répertoire national des personnes physiques géré par l'Insee. Ils ont aussi un usage statistique et sont traités en pratique comme des questionnaires d'enquêtes de la statistique publiques, dont l'opportunité est discutée au Cnis et la conformité validée par le comité du Label.

[76] Le fichier historique est désormais produit par Pôle emploi, créé en 2008.

[77] Par exemple, la résidence fiscale n'est pas le lieu de résidence au sens du recensement. Il en est de même pour la composition des familles. L'inscription à Pöle emploi, y compris dans la seule catégorie A, ne signifie pas que le demandeur d'emploi est chômeur BIT. L'enregistrement du versement d'une rémunération une date donnée ne signifie pas que le bénéficiaire est en emploi salarié à la même date.

[78] Par exemple, l'information sur le diplôme, qui joue un rôle structurant sur le marché du travail français, n'est pas disponible dans la déclaration sociale nominative. Elle est présente dans le fichier de Pôle emploi, mais la qualité de cette information, enregistrée sur une base déclarative, n'est pas assurée, notamment en cas de diplômes multiples. Les aspects relatifs aux opinions, aux ressentis, aux motivations des choix, sont aussi absents.

Ces trois panels permettent aussi d'illustrer certains des arbitrages que les services statistiques doivent effectuer lorsqu'ils construisent des bases d'étude mobilisant des données administratives et quand ils articulent plusieurs sources entre elles (certains de ces arbitrages sont aussi nécessaires pour les données d'enquête), notamment :

- Faut-il proposer les données brutes, telles qu'elles sont dans la source administrative d'origine ou les retraiter, voire les synthétiser, pour en faciliter l'utilisation par des non-spécialistes de cette source ?
- Faut-il procéder à des contrôles de cohérence et des corrections des informations disponibles dans la source primaire ? Dans ce cas, faut-il proposer aux utilisateurs les seules informations redressées ou également les informations brutes, ou, au moins, des indicateurs de redressement ?
- En cas de sources multiples, faut-il traiter les éventuelles incohérences observées entre les différentes sources pour faciliter le travail des utilisateurs et assurer une homogénéité des traitements ou, au contraire, laisser le plus d'informations possibles aux utilisateurs ?

3.1. Le Panel « Tous salariés »

Producteur	Insee
Sources mobilisées	Déclarations annuelles de données sociales / déclarations sociales nominatives
Champs	- Salariés du secteur privé et semi-public non agricole et hors particulier employeurs depuis 1967 + salariés de la fonction publique hospitalière depuis 1984 + salariés des deux autres fonctions publiques depuis 1988 + salariés agricoles depuis 2002 + salariés des particuliers employeurs depuis 2009 - nés en octobre les années paires de 1967 à 2001 ; tous les nés en octobre de 2002 à 2011 ; nés en octobre les années paires ou nés un jour « EDP » depuis 2012 (2 au 5 janvier, 1er au 4 avril, 1er au 4 juillet, 1er au 4 octobre). - de métropole depuis 1967 + des départements d'outre-mer depuis 2002
Période couverte	Depuis 1967 (pour les salariés de métropole du secteur privé et semi public non agricole, hors particulier employeur)
Variables d'identification	Nir

Produit par l'Insee, le noyau historique du panel « tous salariés » a été mis en place au milieu des années 1960 sur le champ des salariés du secteur privé et semi-public non agricole en exploitant une déclaration administrative obligatoire : la déclaration de salaire 24.60, effectuée par les employeurs à destination de l'administration fiscale (cette déclaration était utilisée pour déterminer le montant d'une taxe assise sur les salaires due par les employeurs et pour contrôler l'exactitude des déclarations individuelles de revenus)[79].

Profitant de l'introduction du numéro d'identification individuel dans cette déclaration fiscale à partir de 1964, l'Insee a d'abord compilé les déclarations liées à une même personne sur une année donnée dans des fichiers « salariés ». Il s'agissait alors surtout de mieux mesurer les revenus salariaux d'un travailleur dans une entreprise donnée, l'entreprise pouvant déclarer ces revenus de façon éclatée, dans des déclarations distinctes, par exemple en isolant les primes. Pour des raisons de capacité de traitement, cette exploitation s'est initialement limitée à un échantillon de personnes, sélectionnées sur un critère simple, le mois de naissance : les personnes nées en octobre d'une année paire, soit environ un vingt-quatrième de la population. Le panel est né dans la foulée par la juxtaposition de ces fichiers annuels « salariés » pour permettre les études longitudinales et étudier ainsi les effets de l'ancienneté, les promotions individuelles, les possibles déclassements en cas de changement d'employeur…

D'abord restreint aux personnes nées en octobre des années paires, le panel double sa dimension à partir de 2002 en s'élargissant aux personnes nées en octobre des années impaires. Cette extension d'échantillon n'est cependant pas d'un grand intérêt pour les travaux longitudinaux puisque l'Insee modifie à nouveau son plan de sondage dès les données 2012 : les personnes nées en octobre des

[79] Perrot Marguerite (1969), « Où retrouver les statistiques de salaires ? », *Economie et Statistique*, n°5, Insee, pp. 61-71. Voir aussi Blanchemanche Madeleine (1968), « Les salaires dans l'industrie, le commerce et les services en 1966 (Exploitation statistique des déclarations « modèle 2.460 ») », *Etudes et conjoncture*, n°23-7, Insee, pp. 3-45.

années paires restent dans le panel mais, à la place de celles nées les années impaires, l'Insee intègre désormais celles nées l'un des jours de l'échantillon démographique permanent (voir partie suivante). Ce choix permet de poursuivre les utilisations longitudinales sur longue période sur le cœur historique du panel, constitué des personnes nées en octobre des années paires, tout en permettant de déployer progressivement un nouveau panel pouvant s'articuler avec l'échantillon démographique permanent et ainsi, s'enrichir d'informations sociodémographiques complémentaires.

Outre les évolutions d'échantillonnage, le panel a également connu une évolution de son champ au cours du temps. D'abord limité aux salariés du secteur privé et du secteur semi-public non agricole, il intègre les salariés de la fonction publique hospitalière en 1984, puis les agents des deux autres fonctions publiques[80] à partir de 1988, les salariés agricoles à partir de 2002, les salariés des particuliers employeurs à partir de 2009. On parle alors de panel « Grand format » ou « Tous salariés ».

Avec le temps, les obligations déclaratives ont aussi évolué et des fusions ont été opérées dans le cadre d'une démarche récurrente de simplification administrative[81]. Sans entrer dans les détails, il faut néanmoins citer la fusion en 1970 de la déclaration fiscale avec l'obligation qu'avaient les employeurs de déclarer aussi aux organismes de sécurité sociale les rémunérations versées[82]. Cette fusion donne naissance à la déclaration annuelle de salaire (DAS). En 1984, la fusion avec l'attestation d'activité salariée aboutit à la déclaration annuelle de données sociales (DADS). Depuis 2017, la déclaration sociale nominative (DSN) remplace la DADS, en intégrant encore une fois diverses obligations déclaratives supplémentaires. Cette dernière évolution modifie le rythme des déclarations des entreprises, puisque la DSN est mensuelle alors que la DADS était annuelle.

Ces évolutions se sont accompagnées de nombreuses modifications dans l'organisation des circuits de traitements de l'information et dans les informations disponibles, sans en remettre en cause le noyau principal constitué par le montant des rémunérations, les caractéristiques de l'employeurs (secteur d'activité, taille de l'entreprise, localisation) et les principales caractéristiques du salarié (sexe, âge, localisation, PCS, conditions d'emploi).

Le panel n'intègre pas toutes les informations disponibles dans les sources initiales. Celles-ci sont sélectionnées et synthétisées par l'Insee pour en faciliter l'usage. En particulier, l'observation de base du panel de l'Insee est un salarié donné, dans une entreprise donnée, au cours d'une année civile donnée. Il peut arriver que le salarié ait travaillé au cours de l'année dans plusieurs établissements[83] d'une même entreprise[84]. Dans ce cas, les informations qualitatives sur l'emploi occupé (secteur d'activité, taille de l'employeur, lieu de travail, PCS) sont celles relatives à l'établissement pour lequel la durée de paie est la plus longue[85]. En revanche, les informations relatives aux rémunérations, aux heures travaillées et à la durée de paie associées aux différents établissements concernés sont cumulées. Si le salarié a travaillé dans deux entreprises différentes au cours d'une même année civile, il est présent dans le fichier avec deux observations.

Le passage des déclarations administratives à des fichiers « statistiques » s'accompagnent aussi de contrôles et de redressements qu'il est important que l'utilisateur connaisse. À titre d'exemples[86] :

- des retards de déclaration peuvent générer des trous de collecte. Des contrôles sont donc effectués en comparant les déclarations reçues d'une année sur l'autre, sachant que la disparition d'un déclarant peut aussi résulter d'une fermeture ou d'une fusion avec un autre déclarant ;
- des primes peuvent faire l'objet de déclaration distinctes des salaires. Dans une approche de repérage des emplois, il importe alors d'identifier les rémunérations qui correspondent à un emploi et celles qui correspondent à des rémunérations « annexes ». Des traitements

[80] Fonction publique d'État et fonction publique territoriale.
[81] Voir tableau page 28 in Jugnot Stéphane (2014), « La constitution de l'échantillon démographique permanent de 1968 à 2012 », *Document de travail*, n°F1406, Insee.
[82] Pour le contrôle des cotisations sociales employeurs et salariés dues.
[83] Identifiés par le numéro Siret.
[84] Identifiée par le numéro Siren.
[85] Si deux établissements sont ex aequo, c'est l'établissement qui a fourni la rémunération la plus forte qui est retenu.
[86] Ces exemples s'appuient principalement sur le fonctionnement en vigueur avant l'intégration de la déclaration sociale nominative.

spécifiques sont effectués pour traiter statistiquement ce sujet, sur la base de critères et de seuils ;

- les déclarations de salaires sont accompagnées de déclarations de périodes d'emploi, donc d'une durée d'emploi et, depuis 1995, d'un nombre d'heures. Ces informations ne sont pas toujours bien renseignées dès lors qu'elles ne sont pas centrales pour les usages administratifs. Les logiciels de gestion de paie peuvent parfois préremplir par défaut les déclarations. Le déploiement des 35 heures a aussi pu perturber certaines pratiques déclaratives en raison de la possibilité d'avoir des régimes d'horaires hebdomadaires habituels différents, combinés à un nombre variable de jours de réduction du temps de travail. Des contrôles sont donc effectués sur le salaire mensualisé et le salaire horaire pour repérer des anomalies dans le niveau de rémunération et/ou les durées de travail déclarées. Ces contrôles, comme les redressements qui en dérivent, sont effectués sur la base de critères et de seuils qui sont, notamment, définis en s'appuyant sur les pratiques observées comme statistiquement habituelles dans la profession concernée. Les règles de contrôle et de redressement évoluent dans le temps ;
- des entreprises peuvent déclarer de façon groupée les salariés de différents établissements, induisant une mauvaise localisation du lieu de travail, ce qui est susceptible de perturber les analyses territorialisées.

Du fait de sa dimension longitudinale, l'utilisation du panel sur longue période peut être aussi compliquée par l'évolution des nomenclatures, notamment celles relatives à l'activité économique[87] et à la catégorie socioprofessionnelle[88]. Les contraintes et les arbitrages du service statistique producteur pèsent aussi. L'évolution de l'échantillonnage a déjà été évoquée. Elle déforme la représentation de la structure par âge des salariés suivis. L'échantillonnage retenu peut aussi perturber des analyses de politiques publiques où des effets de seuil liés à l'âge interviennent. Parfois, les arbitrages sur la charge de travail perturbent aussi la construction du panel. Ainsi, pour des raisons de priorités et en raison de la charge de travail induite par le traitement des DADS, l'Insee a fait le choix de ne pas les traiter certaines années (1981, 1983 et 1990). Ces années sont donc absentes du panel.

Un panel sert d'abord à suivre des trajectoires individuelles mais il peut aussi avoir pour fonction complémentaire de proposer des successions de photographies représentatives, permettant ainsi aux utilisateurs de faire l'économie d'un accès aux fichiers exhaustifs successifs ayant servis à constituer le panel. Cette seconde fonction, qui n'est pas proposée par tous les panels issus de sources administratives, l'est dans le panel « Tous salariés ». Chaque année, lors de l'intégration des informations d'une année d'observations supplémentaires, des individus sont ainsi intégrés dans le panel uniquement pour lui permettre de donner une image représentative en photographie annuelle. En effet, certaines déclarations ne comportent pas le Nir. Un tirage au sort est donc effectué parmi ces déclarations. Ces personnes complémentaires se voient attribuer un « Nir fictif » dans le panel, qui ne peut être suivi dans le temps

3.2 L'échantillon démographique permanent[89]

Producteur	Insee
Sources mobilisées	- Depuis l'origine : bulletins d'état civil (naissance de la personne et de ses enfants, mariage, décès), bulletins du recensement de la personne. - Ajouts depuis la refonte des années 2010 : fichier électoral (avec récupération de l'historique) + DADS/DSN via le panel « tous salariés » (depuis 2012, avec récupération de l'historique depuis 1967 pour les nés en octobre des années paires) + sources fiscales et sociales via Filosofi (depuis 2011)

[87] NAP73, puis NAF à partir de 1993. La NAF est révisée en 2003, puis en 2008.
[88] CSP de 1954, puis PCS de 1982, révisée en 2003, puis la PCS-ESE à partir de 2009.
[89] Voir notamment Jugnot Stéphane (2014), « La constitution de l'échantillon démographique permanent de 1968 à 2012 », *Document de travail*, n° F1406, Insee, et Durier Sébastien (2018), « l'échantillon démographique permanent à 50 ans : retours sur un dispositif statistique original », *actes des journées de méthodologie statistique 2018,* Insee. Des informations, dont une présentation détaillée des fichiers, sont également disponibles sur le site internet https://utiledp.site.ined.fr (site dédié aux utilisateurs de l'EDP, créé dans le cadre du projet « BigStat » financé par l'agence nationale de la recherche).

Champs	- Personnes nées du 1er au 4 octobre depuis 1968 + personnes nées du 2 au 5 janvier, du 1 au 4 avril, du 1 au 4 juillet ou du 1 au 4 octobre (depuis 2004 pour l'état civil et 2008 pour le recensement) - En métropole depuis 1968 + DOM
Période couverte	Depuis 1968, pour les sources fondatrices (état civil et recensement)
Variables d'identification	Nir, directement ou après identification à la BRPP sur la base de l'état civil complet[90] (nom, prénoms, date et lieu de naissance) avec un traitement manuel partiels des rejets d'identification

Également produit par l'Insee, l'échantillon démographique permanent a été créé à la fin des années 1960 pour faciliter les analyses démographiques longitudinales, par exemple sur la descendance finale des hommes et des femmes ou l'espérance de vie par catégories socioprofessionnelles[91]. Pendant longtemps, il s'est limité aux personnes nées du 1er au 4 octobre. À partir de 1968, tous leurs bulletins d'état civil (naissance, mariage, décès et naissance de leurs enfants) et tous leurs bulletins de recensement étaient collectés et rassemblées dans un fichier unique. Comme pour le panel DADS, le choix d'un critère simple sur la date de naissance facilitait la production du panel et permettait l'entrée permanente de nouveaux individus dans la base.

Les bulletins d'état civil et ceux du recensement ne comportent pas directement le Nir, mais disposent de l'état civil complet de la personne (nom, prénom, date et lieu de naissance). Une recherche au répertoire des personnes physiques permet donc de le récupérer. Le Nir est ensuite utilisé pour assurer le suivi longitudinal. Réattribuer un Nir n'est cependant pas toujours facile du fait d'informations incomplètes, erronées ou difficilement lisibles sur les bulletins. Pour gérer ces difficultés, des traitements automatiques et manuels sont effectués afin d'essayer de retrouver la bonne personne. Pour les personnes nées à l'étranger, ces recherches complémentaires ont longtemps été plus systématiques et moins conclusives parce que les agents de l'Insee n'accédaient pas au fichier de la sécurité sociale, chargée de leur attribuer leur Nir par délégation. De ce fait, le suivi longitudinal est plus compliqué pour les personnes nées à l'étranger.

Dans les années 2000, la mise en œuvre d'une nouvelle méthode de « recensement » a obligé l'Insee à refondre profondément l'échantillon démographique permanent. Désormais, le recensement, jusqu'alors exhaustif mais ponctuel et espacé dans le temps, est remplacé par une enquête annuelle, collectée sur un échantillon d'adresses. Plus exactement, chaque année, un cinquième des communes de moins de 10 000 habitants est recensé exhaustivement, de sorte que toutes les petites communes sont couvertes sur cinq ans. Dans les grandes villes, seule 8 % environ de la population est recensée chaque année, de sorte qu'environ 40 % des logements sont enquêtés sur cinq ans. Chaque année, les informations collectées aux cours des cinq dernières enquêtes annuelles sont combinées pour produire les résultats du « recensement » de l'année[92]. Pour l'échantillon démographique permanent, ce changement de méthode a des conséquences majeures. Il implique qu'il n'est plus possible d'actualiser « exhaustivement », à intervalle plus ou moins réguliers, les informations socio-démographiques des personnes qui y figurent. Une personne non mobile résidant dans une petite commune peut être recensée tous les cinq ans alors qu'une personne habitant une grande ville peut ne jamais l'être.

Deux évolutions majeures ont été mises en œuvre pour répondre à ce défi. La première, qui ne résout pas vraiment la difficulté, a consisté à quadrupler la taille de l'échantillon, avec l'ajout des personnes

[90] Les bulletins d'état civil et ceux du recensement ne comportent pas le Nir et nécessitent donc une recherche du Nir sur la base de l'état civil complet.

[91] La Cnil est saisie en 1980 de l'existence de l'échantillon démographique permanent. La procédure aboutit en 1984, avec la publication du décret n° 84-393 du 23 mai 1984 et de l'arrêté du 23 mai 1984. Le décret autorise l'utilisation du RNIPP, tandis que l'arrêté « définit » de façon particulièrement vague l'échantillon démographique permanent. Son article 1er donne comme objectif unique « l'élaboration de statistiques démographiques et sociales » et l'article 2 précise que « les informations enregistrées concernent les personnes nées du 1er au 4 octobre de chaque année. Ces informations sont issues des recensements de la population successifs, des bulletins statistiques de l'état civil et du fichier électoral ». Voir Jugnot S., « La constitution de l'échantillon démographique permanent de 1968 à 2012 », *Document de Travail*, n° F1406, Insee, 2014.

[92] Le « recensement » est daté de l'année médiane des cinq enquêtes annuelles utilisées pour le construire. Par exemple, le recensement 2018 est produit avec les enquêtes annuelles collectées de 2016 à 2020. Des ajustements seront opérés pour les recensements suivants en raison de l'épidémie du Sras-Cov2 survenue à partir de mars 2020, qui a conduit à annuler l'enquête annuelle de recensement 2021.

nées du 2 au 5 janvier, du 1er au 4 avril et du 1er au 4 juillet[93]. La seconde, plus cruciale, a consisté à faire le choix d'intégrer des nouvelles sources administratives susceptibles de fournir des informations sociodémographiques plus ou moins proches des thématiques couvertes par les bulletins du recensement. L'Insee a ainsi ajouté des informations :

- sur l'inscription électorale, à partir du fichier national des électeurs[94] (récupération de l'historique depuis 1967) ;
- sur l'exercice d'un emploi salarié, à partir du panel « Tous salariés » déjà évoqué (informations disponibles à partir de l'année 2012, avec récupération de l'historique depuis 1967 pour les personnes nées les quatre premiers jours d'octobre une année paire et depuis 2002 pour les personnes nées les même quatre jours une année impaire) ;
- sur la composition du ménage déclaré au fisc, les sources de revenus, leurs montants et les allocations sociales perçues, à partir des données socio-fiscales exhaustives agrégées par l'Insee dans son dispositif Filosofi[95] (informations disponibles à partir des revenus 2010).

Les informations issues de ces sources administratives ne sont pas utilisées pour « compléter » les données manquantes induites par la mise en place d'un « recensement » sur échantillon. Une telle option aurait été complexe à mettre en œuvre. Il aurait fallu arbitrer parfois entre plusieurs sources administratives, par exemple pour le lieu de résidence, entre le lieu d'inscription électorale, l'adresse déclarée aux services fiscaux et celle mentionnée sur les feuilles de paie. Il aurait aussi fallu tenir compte des écarts possibles entre les concepts statistiques, les déclarations spontanées des personnes au recensement et les informations administratives – des écarts entre les sources déclaratives, dont le recensement, et les sources administratives sont par exemple, régulièrement documentés pour les informations relatives à l'emploi[96]. De plus, si l'Insee avait effectué ces arbitrages et ces traitements, ils n'auraient pas forcément été ceux qu'un utilisateur aurait retenus pour ses travaux spécifiques.

Au contraire, chaque nouvelle source fait l'objet de tables spécifiques complémentaires, qui s'articulent aux autres tables relatives au noyau historique de l'échantillon démographique permanent, laissant l'utilisateur libre de choisir les sources d'information qu'il souhaite privilégier. Sur le plan technique, cette structuration facilite aussi l'élargissement considérable des informations proposées dans l'échantillon démographique permanent (voir encadré 7).

Au fur et à mesure que le recul disponible pour ces nouvelles sources administratives s'allongera, il est probable que les informations à trou du recensement seront de plus en plus laissées de côté par les utilisateurs pour leurs analyses de trajectoires, au profit des seules données administratives, exhaustives sur leur champ. Les données du recensement garderont cependant un intérêt pour des travaux plus méthodologiques de comparaison avec les sources administratives, pour étudier le taux de couverture de ces sources ou leurs biais d'observation possibles.

Les informations issues des bulletins du recensement gardent aussi un intérêt pour les champs d'investigation du Céreq puisque c'est la seule source de l'échantillon démographique permanent à proposer une information sur le niveau de diplôme. Sur ce point, il est probable que, d'ici quelques années et avec l'accord de la Cnil, le service statistique public décide d'ajouter dans l'EDP des informations sur les parcours scolaires à partir des fichiers administratifs récupérés par les services statistiques des ministères de l'Éducation nationale et de l'Enseignement supérieur[97]. Ce schéma ne semble pas envisagé pour le moment parce que la Cnil a demandé que l'identifiant de gestion utilisé dans la sphère de l'Éducation soit clairement distinct du Nir et parce que les controverses publiques sur le déploiement de l'INE sont encore trop récentes. Cependant, comme la première partie le rappelle, il existe un mouvement tendanciel de libéralisation du champ des possibles légaux, notamment pour les

[93] Ce choix revient à retenir les personnes nées les 4 premiers jours de chaque trimestre, sans retenir le 1er janvier.
[94] Ce fichier, géré centralement par l'Insee, permet d'éviter qu'une même personne s'inscrive dans deux communes en même temps. Jusqu'à récemment, ses informations pouvaient différer des listes électorales car celles-ci restaient gérées et produites au niveau local, avec des échanges d'informations avec l'Insee. La loi n°2016-1048 du 1er août 2016 a modifié cette organisation en transformant le fichier de l'Insee en « répertoire électoral unique » qui sert désormais aussi à produire les listes électorales.
[95] Filosofi est produit par l'Insee pour lui permettre de proposer des indicateurs localisés à des niveaux infra-départementaux sur le revenu disponible des ménages (niveau moyen, distribution, indicateurs de pauvreté, etc.). Pour une présentation détaillée, voir « Fichier Localisé Social et Fiscal », *Sources et méthodes*, Insee, janvier 2020.
[96] Voir par exemple, Massif Jean-Benoit (2016) « De quelle mesure de l'emploi le Recensement de la population est-il le nom ? La place du Recensement de la population dans le système de suivi de l'emploi », *Economie et* statistique, n° 483-484-485, Insee.
[97] Cette remarque vaut aussi pour le panel « Tous salariés ».

traitements à finalité de recherche ou de production statistique. Ce mouvement est renforcé par l'assouplissement des sensibilités éthiques et politiques à la question des rapprochements de fichiers. Depuis 2016, l'Insee peut à nouveau envisager le développement d'appariements de sources administratives à grande échelle tel qu'il l'envisageait dans son projet Safari sur la base de l'identifiant statistique non signifiant distinct du Nir. Or, il est techniquement possible d'associer ce même identifiant statistique à l'INE *via* les informations d'état civil de l'élève ou de l'étudiant disponibles dans le répertoire national de l'INE que la Depp gère.

Comme le panel « Tous salariés », l'échantillon démographique présente des limites liées aux sources utilisées et à sa dimension longitudinale sur longue période. Par exemple, le critère d'appartenance à l'échantillon n'est pas homogène selon les sources[98]. L'absence d'une personne dans l'échantillon peut ainsi résulter d'un trou de collecte ou traduire un départ hors de France, un changement de situation administrative[99], un décès non repéré. Des trous de collecte peuvent aussi résulter d'une priorisation négative liée à la charge de travail de l'Insee au détriment des traitements d'identification des personnes. Par ailleurs, pour une source donnée, la richesse des informations proposées peut évoluer fortement dans le temps (c'est notamment le cas pour les informations issues des recensements) et selon les sources, les informations proposées sont des informations déclarées ou retraitées. Parfois, les deux informations, brutes et retraitées, sont disponibles. Parfois, seul un indicateur de retraitement l'est. Parfois, il n'est pas possible de savoir si l'information proposée est retraitée ou brute. Dans l'idéal, les utilisateurs conséquents auraient évidemment besoin de savoir *a minima* si l'information proposée est retraitée ou brute, ainsi que la nature des contrôles et retraitements effectués.

Enfin, il faut noter que l'échantillon démographique permanent a acquis un rôle pivot dans la statistique publique. Plusieurs panels alimentés par des sources administratives exhaustives se calent ainsi sur sa méthode d'échantillonnage en ciblant les personnes nées un « jour EDP » afin d'y puiser ensuite des informations complémentaires aux sources administratives qu'ils mobilisent. C'est en partie le cas du panel « Tous salariés » de l'Insee. C'est aussi le cas de deux panels de la DREES, l'échantillon inter-régimes des cotisants et l'échantillon inter régimes des retraités, ou du panel « Trajam » de la Dares présenté *infra*.

[98] Selon la source : personnes nées un « jour EDP » recensées en France, ou ayant leur résidence fiscale en France, ou salariées en France, ou inscrites sur les listes électorales…

[99] Par exemple, en cas de sortie des tables issues du panel « Tout salariés » : passage à l'emploi non salarié, à l'inactivité ou au chômage.

Encadré 7 • Informations issues des sources intégrées à l'échantillon démographique permanent

Historiquement, l'échantillon démographique permanent collectait les informations issues des bulletins d'état civil et des bulletins de recensement concernant les personnes nées un « jour EDP ». D'autres sources ont été mobilisées à l'issue d'une restructuration importante de l'échantillon au début des années 2000 : le fichier électoral, le panel « Tous salariés » et Filosofi. Actuellement, les informations suivantes sont intégrées :

- ***Pour l'état civil*** : l'acte de naissance de l'individu EDP, ses actes de mariage, les actes de naissance de ses enfants et son acte de décès.

- ***Pour les recensements*** : des informations collectées à l'occasion des recensements généraux de population (1968, 1975, 1982, 1990 et 1999) puis des enquêtes annuelles de recensement depuis 2004. Le contenu précis des informations retenues dans le recensement s'est fortement élargi avec le temps. Initialement centré sur les informations du bulletin individuel de l'individu EDP, il propose désormais aussi des informations issues de sa feuille de logement et des bulletins individuels de toutes les personnes recensées dans le logement.

- ***Pour le fichier électoral*** : les séquences d'inscription disponibles, avec la date d'inscription, le lieu d'inscription, le type d'inscription (qui permet de distinguer notamment les inscriptions d'office des inscriptions volontaires) et, s'il y a lieu, la date de radiation. Le type de liste est également mentionné (liste principale, liste consulaire, liste complémentaire pour les élections européennes, liste complémentaire pour les élections municipales) – cette information peut permettre de repérer une petite partie des départs de Français vers l'étranger.

- ***Pour le panel « Tous salariés »*** : l'échantillon démographique permanent ne reprend qu'une partie des informations disponibles dans le panel « Tous salariés » pour ne proposer qu'une seule observation d'informations synthétiques pour chaque croisement « personne née un jour EDP » / « année civile ». Sont notamment disponibles : le nombre d'entreprises différentes dans lesquelles la personne a travaillé durant l'année, le nombre total d'heures travaillées dans l'ensemble de ces entreprises, le nombre total de jours payés et le salaire total perçu. D'autres informations ne sont disponibles que pour l'emploi relatif à l'entreprise dans laquelle le salarié a travaillé le plus longtemps : lieu de résidence de la personne, lieu de travail, type de contrat de travail, catégorie socioprofessionnelle, secteur d'activité, domaine d'emploi (fonction publique d'État, territoriale, hospitalière, etc.), taille de l'entreprise et de l'établissement en nombre de salariés. Pour les travaux de comparaisons avec le recensement, il faut noter que la date de référence de la situation d'emploi, du lieu de résidence ou du lieu de travail n'est pas forcément celle du recensement.

- ***Pour Filosofi*** : les informations disponibles concernent les ménages fiscaux ordinaires. Elles sont issues des déclarations de revenus, de la taxe d'habitation. Elles portent aussi sur les prestations sociales versées par la caisse nationale des allocations familiales, la caisse nationale de l'assurance vieillesse et la mutualité sociale agricole. Une partie des revenus du ménage est imputée par l'Insee en s'appuyant sur une modélisation nationale construite à partir de l'enquête Patrimoine (c'est le cas des revenus financiers de certains produits non soumis à déclaration de revenu, comme le livret A, le livret d'épargne populaire, l'assurance-vie ou les PEA). Les informations proposées sont réparties entre une table « individu », une table « logement », une table « revenu de l'individu » et une table « revenu du ménage ». C'est au niveau du ménage que les informations sur les revenus, les allocations et les minimas sociaux perçus sont le plus détaillées. Des indicateurs de positionnement du ménage par rapport à l'ensemble de la population sont aussi proposés (pauvreté du ménage, niveau de vie, centile de revenu…). Concernant le logement, des informations sur son confort sont disponibles mais elles proviennent du fichier des propriétés bâties et de la taxe d'habitation, dont l'actualisation fait l'objet de débats publics récurrents depuis plusieurs décennies.

3.3 Les fichiers historiques de Pôle emploi[100]

Producteur	Pôle emploi
Sources mobilisées	Fichiers de gestion de l'Unedic et de l'ANPE, puis de Pöle emploi
Champs	Exhaustif
Période couverte	Dix années glissantes
Variables d'identification	Nir, Identifiant Pôle emploi.
Remarques	

Le fichier historique statistique des demandeurs d'emploi (FHS) a été développé à partir de la fin des années 1990 par l'Agence nationale pour l'emploi, devenu depuis Pôle emploi. Ce développement s'est fait en partenariat avec la Dares et l'Insee. Destiné à des travaux d'analyses statistiques, il est issu d'un retraitement du fichier historique administratif des demandeurs d'emploi de Pôle emploi (FHA), qui répond, lui, à des besoins opérationnels.

Le fichier historique permet de suivre les trajectoires des demandeurs d'emploi dans les fichiers de gestion de Pôle emploi au cours des dix dernières années : les périodes d'inscription, avec le motif d'entrée et le motif de sortie, l'exercice d'une activité réduite ou non chaque mois pendant ces périodes d'inscription, le bénéfice ou non d'un droit à une allocation chômage, les entretiens réalisés dans le cadre du projet d'action personnalisé (PAP) puis du projet personnalisé d'accès à l'emploi (PPAE), etc. Il comporte aussi des informations sociodémographiques renseignées lors de l'inscription du demandeur d'emploi ou actualisées ultérieurement, comme le sexe, l'âge, le lieu de résidence, la situation familiale, ainsi que la nature de l'emploi et le type de contrat recherché. Le diplôme déclaré à l'inscription fait partie des informations enregistrées. Comme pour l'échantillon démographique permanent de l'Insee, ces informations sont réparties entre plusieurs tables thématiques : caractéristique du demandeur d'emploi, caractéristiques des demandes d'emploi successives, historique des mois d'activité réduite…

La capacité à suivre les trajectoires d'inscription sur dix ans constitue l'apport principal de cet outil. Il permet ainsi de repérer, à des fins de recherche ou de production d'indicateurs, les sorties durables des listes de Pöle emploi ou, au contraire, les allers-retours fréquents[101]. Les personnes concernées peuvent être caractérisées, tant par différents attributs sociodémographiques que par des informations sur le niveau de service proposé par Pôle emploi. Avec l'accord de Pôle emploi, des enrichissements ponctuels sont aussi possibles avec le fichier historique activité pour disposer d'informations plus particulières sur l'accompagnement des demandeurs d'emploi : les entretiens réalisés, les actions conseillées, réalisées ou non, les mises en relation avec les entreprises, les formations des demandeurs d'emploi indemnisés.

Un bémol doit cependant être apporté dans la capacité à suivre les trajectoires du fait de l'utilisation de l'utilisation de l'identifiant « Assedic » dans la production régulière du FHS. De ce fait, dès qu'un individu change de zone Assedic, sa demande d'emploi est interrompue dans le territoire de départ pour recommencer dans le territoire d'arrivée, avec un identifiant différent. Pôle emploi dispose cependant aussi d'un identifiant national des demandeurs d'emploi, donc d'un moyen pour surmonter ponctuellement cette difficulté.

Le FHS est exhaustif sur son champ. Il permet donc d'envisager des analyses territoriales ciblées. Pour les projets de recherche qui le justifient et avec l'accord de Pôle emploi, il est aussi possible d'envisager des appariements avec des données externes, comme la statistique publique l'a déjà fait avec les

[100] Cette partie s'appuie notamment sur la version 1.4 de février 2011 de la « Documentation générale » du fichier historique des demandeurs d'emploi, produite par la direction des études, des statistiques et des prévisions de Pôle emploi, ainsi que sur la version 3.0 d'aout 2011 du « Dictionnaire des données ».

[101] Voir par exemple, Debauche Etienne et Jugnot Stéphane (2006), « La mesure d'un effet global du projet d'action personnalisé », *Document d'études*, n°112, Dares.

déclarations annuelles de données sociales[102] ou l'enquête Emploi[103]. Pour les DADS, l'appariement s'est appuyé sur le Nir. Pour l'enquête Emploi, qui ne dispose pas du nom de famille, un algorithme a été utilisé pour identifier les rapprochements potentiels sur la base du prénom, du sexe, de la date de naissance et de la commune de résidence.

Le fichier historique statistique est actualisé chaque trimestre. Adossé à un fichier de gestion, les informations les plus récentes peuvent évoluer d'une version à l'autre en fonction des délais de mise à jour et de retards ponctuels possibles. Les motifs de sortie sont généralement mal connus. Il ne permet donc pas de repérer directement les sorties directes vers l'emploi mais seulement les sorties durables des listes, ce qui constitue déjà en soi une information intéressante compte tenu de l'importance de Pôle emploi dans le service public de l'Emploi.

Il convient aussi de rappeler que les statistiques de Pôle emploi rendent compte des bénéficiaires de ce service public, selon les règles de gestion qui lui sont propres et qui peuvent évoluer dans le temps. Elles ne rendent pas compte du chômage au sens du bureau international du travail ou même du chômage déclaratif tel qu'il peut être identifié dans les enquêtes auprès des personnes. La notion de « demandeurs d'emploi sans activité réduite » est elle-même trompeuse[104]. Elle ne peut être considérée comme une approximation du chômage au sens du bureau international du travail. Des divergences importantes existent entre ces différentes notions, à plusieurs reprises documentées[105]. En particulier, ces écarts ne sont pas répartis de façon homogène sur toute la population, en particulier en fonction de l'âge.

[102] Le Barbanchon Thomas et Vicard Augustin (2009), « Trajectoire d'une cohorte de nouveaux inscrits à l'ANPE selon le FH-DADS », *Document d'étude*, n°152, Dares. Pour un exemple d'utilisation, voir par exemple, Fontaine Maëlle et Rochut Julie (2014), « L'activité réduite : quel impact sur le retour à l'emploi et sa qualité ? Une étude à partir de l'appariement FH-DADS », *Document d'études*, n°183, Dares.

[103] Hameau Alexis *et ali* (2019), « Appariement entre l'enquête Emploi et le fichier historique de Pôle emploi sur la période 2012-2017 », *Document d'études*, n°233, Dares.

[104] Jugnot Stéphane (2015), « Améliorer la publication mensuelle des statistiques du « chômage » pour faciliter le débat public. Quelques propositions », *Document de travail*, n°03-2015, IRES.

[105] Voir, par exemple, Marchand Olivier (1991), « Statistiques du chômage : les écarts se creusent depuis cinq ans », *Economie et Statistique*, n°249, Insee, pp. 7-14 // Coder Yohan et *ali* (2019), « Les chômeurs au sens du BIT et les demandeurs d'emploi inscrits à Pôle emploi : une divergence de mesure du chômage aux causes multiples », *Insee Références Emploi, chômage, revenu du travail*, Insee, pp.71-85.

4. Une multiplication des projets d'appariement

Depuis quelques années, les projets d'appariements de données administratives dans les champs d'études du Céreq se multiplient sous l'effet croisé de plusieurs facteurs. L'assouplissement du cadre juridique n'en est qu'une condition nécessaire, voire une conséquence. La pression de la contrainte budgétaire joue aussi, le recours à des données administratives préexistantes étant perçu comme moins couteux que la mise en œuvre de grosses enquêtes par échantillon, sans que des véritables analyses comparées des coûts et des avantages des deux approches n'aient été réellement conduites[106].

Les opportunités techniques et la commande politique d'indicateurs « de performance » à un échelon fin peuvent aussi expliquer l'appétence pour les données administratives. Du côté des opportunités techniques, la capacité à traiter facilement des gros volumes de données ainsi que la centralisation de systèmes d'information auparavant éclatés sont des facteurs facilitateurs. Du côté de la commande politique, certains projets récents, comme nous le verrons, visent à répondre à des commandes politiques d'indicateurs dit « de performance », dont on ne sait s'il s'agit en réalité d'indicateurs d'évaluation, d'aide au pilotage ou de communication.

Nous ne présenterons ici que des projets récents mis en œuvre par la statistique publique qui peuvent concerner les champs de la formation initiale, de la formation continue et des parcours professionnels.

4.1. Les fichiers anonymisés des élèves pour le suivi des parcours scolaires dans le champ de la Depp

Le projet de fichiers anonymisés d'élèves pour la recherche et les études (Faere) a été initié par la Depp au début des années 2000 pour faciliter les études statistiques, notamment celles relatives au suivi du parcours des élèves quel que soit leur cursus. Pour cela, la Depp centralise des fichiers de données individuelles anonymisées provenant de différentes sources. Un fichier est produit pour chaque année scolaire.

Le fichier relatif à l'année scolaires *n/n+1* donne des informations sur tous les élèves inscrits dans les établissements couverts par ses deux sources principales :

- Le système d'information « scolarité » (remplacées depuis 2018 « Sysca Sco »), qui couvre l'ensemble des élèves inscrits un moment au cours de l'année scolaire considérée dans les établissements du second degré, publics ou privés sous contrat, du ministère chargé de l'Éducation nationale. Pour certaines académies, l'information est aussi disponible pour les établissements privés hors contrat.
- Le système d'Information sur la formation des apprentis, intégré depuis 2008, qui couvre les jeunes inscrits au 31 décembre en centres de formation d'apprentis (CFA) et des sections d'apprentissage (SA).

Au total, 3 % des élèves du second degré et environ 15 % des élèves du post-bac seraient exclus des fichiers des apprenants. En particulier, les élèves scolarisés dans les établissements relevant du ministère de l'Agriculture, dans les lycées militaires et dans la majorité des établissements privés hors contrat ne sont pas couverts.

[106] Le coût d'une bonne appropriation des données, qui implique de connaître les processus de gestion à l'origine des fichiers administratifs, la liste des informations faisant l'objet de contrôles, bloquants ou non, voire de corrections, et leurs évolutions, est souvent négligé. Le coût du passage de données administratives brutes à des données à usages statistiques, qui implique des contrôles et des redressements, est aussi sous-estimé. Il faudrait aussi tenir compte du coût nécessaire à la tenue d'une documentation complète et à jour de ces données, pour que les utilisateurs puissent les exploiter à bon escient.
Pour une analyse complète des coûts, les plus radicaux pourraient aussi chercher à tenir compte des effets des décisions publiques qui pourraient être prises sur la base d'analyses et d'évaluations biaisées du fait du recours à des données administratives qui ne permettent pas d'observer pleinement les réalités que l'on souhaite mesurer, notamment en matière d'insertion professionnelle et de retour à l'emploi. L'évaluer serait cependant illusoire.

Pour les élèves couverts, diverses informations sont disponibles :

- des caractéristiques de l'élève (sexe, date de naissance, département de naissance, nationalité agrégée, commune de résidence, PCS des adultes « responsables » et lien entre l'élève et ces adultes) ;
- des caractéristiques de l'établissement de scolarisation l'année *n/n+1* (numéro d'établissement, département et académie, type d'établissement, caractère public ou non) ;
- quelques caractéristiques de la scolarité suivi l'année *n/n+1* (identifiant de la classe, boursier ou non, formation suivie, régime scolaire, matières aux choix, date et motif de sortie) ;
- les résultats aux examens à la session *n+1* pour le diplôme national du brevet, le baccalauréat, le CAP et BEP (inscription, présence, résultats) ;
- les caractéristiques de l'établissement de scolarisation et de la scolarité l'année antérieure (*n-1/n*) *;*
- un identifiant non signifiant, obtenu par cryptage de l'INE, qui permet de relier les données relatives à un même élève d'une année sur l'autre pour suivre son parcours.

Pour être plus précis, les informations disponibles ne sont pas strictement les mêmes selon le système d'information d'origine, donc selon la nature de l'établissement fréquenté, difficulté inévitable lorsque l'on agrège des données de sources administratives diverses. En particulier **:**

- le département de naissance est absent des fichiers SIFA jusqu'en 2016 inclus ;
- la nationalité peut être détaillée pour les élèves des fichiers Scolarité mais pas pour ceux des fichiers SIFA (pour eux, distinction « Union européenne » / » Autre ») ;
- la PCS est collectée pour deux personnes « responsables » dans les fichiers Scolarité et une seule dans les fichiers SIFA, selon une nomenclature « Éducation nationale », intermédiaire entre les nomenclatures Insee de 24 et 42 postes. Dans SIFA, la non-réponse est proche de 50 % et dépasse 60 % pour les apprentis du supérieur ;
- dans les fichiers SIFA, l'établissement de scolarisation des apprentis correspond à l'unité administrative d'immatriculée (UAI) où la formation est dispensée et peut donc correspondre à un établissement au sens juridique ou à l'une de ses parties (antenne de CFA, section d'apprentissage…). Pour les élèves des établissements du champ « Scolarité », y compris parfois des apprentis, il n'y a pas de distinction entre les différentes sections de l'établissement ;
- l'identifiant de la classe, le fait d'être boursier et le régime scolaire (situation par rapport à l'hébergement ou la demi-pension) ne sont disponibles que pour les fichiers Scolarité ;
- pour les élèves quittant leur établissement en cours d'année scolaire, la date de sortie et le motif ne sont disponibles que dans les fichiers Scolarité.

Les informations relatives aux examens sont récupérées dans les systèmes d'information dédiés. Depuis la session 2017, le rapprochement avec le fichier CYCLADES permet ainsi de récupérer des informations concernant le diplôme national du brevet pour les élèves concernés. Les fichiers OCEAN[107] fournissent ces informations pour le baccalauréat, le certificat d'aptitude professionnelle, le brevet d'études professionnelles, le brevet professionnel, les mentions complémentaires de niveau V et IV ainsi que certains diplômes supérieurs (BTS et diplômes comptables). Les résultats aux examens des spécialités agricoles ne sont pas couverts.

Pour des raisons informatiques, liées aux volumes des données mobilisées, la recherche d'informations sur un examen donné n'est faite que pour les élèves relevant du « vivier » théorique de cet examen[108]. Une fois ce vivier défini, la recherche de la présence d'informations pour les élèves concernés dans les bases de données sur les examens se fait d'abord sur la base de l'INE. Comme des INE fictifs étaient présents dans les bases, un rapprochement était ensuite réalisé sur la base d'une variable de profil compilant le sexe, la date et département de naissance, la formation suivie et l'établissement fréquenté (le département de naissance, absent des bases SIFA, n'est évidemment pas utilisé pour les apprentis).

Ce dernier point rappelle que la capacité à reconstituer les parcours dépend de la qualité du renseignement de l'INE. Selon la documentation des fichiers anonymisés des apprenants, environ 15 %

[107] À terme, le système d'information CYCLADES est aussi appelé à remplacer OCEAN.

[108] Les viviers sont détaillés dans la documentation du fichier. Par exemple, le vivier du baccalauréat comprend les inscrits dans des classes de terminales (générales, technologiques ou professionnelles), en 1ère année de BTS ou DTS, en classes de mise à niveau BTS, UPI-ULIOS, MODAL Niveau IV et FCII niveau IV.

des apprentis n'avaient pas d'INE renseigné avant que SIFA ne soit déployé dans le cadre de la mise en œuvre de l'INE unique certifié. Dans ce cas, un INE fictif était généré pour conserver l'enregistrement individuel dans la base annuelle des apprenants. Ces observations ne peuvent pas être reliées à celles des années scolaires antérieures ou suivantes, ce qui perturbe donc le suivi des parcours. Le volume d'INE fictifs variait fortement d'une académie à l'autre. De 2013 à 2016, six académies affichaient des taux d'INE fictifs supérieur à 20 %, dont Paris et Versailles. La proportion d'INE renseignés est aussi très variable dans les bases de données sur les examens selon le type d'examen. Par exemple, en 2009, la proportion d'identifiants renseignés serait de près de 100 % pour les candidats au bac mais n'atteignait que 17 % pour les inscrits au CAP. À partir de l'année 2017, il n'y a plus d'INE fictifs du fait du déploiement de l'INE certifié par le RNIE[109].

4.2. Le système d'information sur le suivi de l'étudiant du Sies

Le système d'information sur le suivi de l'étudiant (SISE) a été mis en place au milieu des années 1990 par le Sies, service statistique du ministère chargé de l'Enseignement supérieur, pour recueillir des données individuelles sur les étudiants à des fins d'études et de production statistique. Il comprend deux composantes, l'une sur les inscriptions (« SISE-Inscriptions »), l'autre sur la réussite aux examens (« SISE-Résultats »).

À la différence de la Depp pour l'essentiel de l'enseignement primaire et secondaire, le Sies ne peut pas s'appuyer sur des systèmes d'information mutualisés et articulés pour faire remonter les informations administratives des établissements de l'enseignement supérieur. L'élaboration des fichiers SISE-Inscription et SISE-Résultats s'appuie donc sur plusieurs dispositifs de remontées d'information, chacun alimenté, à la base, par les établissements à partir d'extractions de leurs logiciels de gestion. Juridiquement, certains de ces dispositifs constituent des remontées administratives, pour les établissements sous tutelle de l'Éducation nationale ou de l'Enseignement supérieur. D'autres sont collectées dans le cadre d'enquêtes statistiques relevant de la loi du 7 juin 1951 (voir encadré 8). À ce titre, dans leur cas, les données individuelles collectées sont couvertes par le secret statistique et leur accès devrait impliquer un passage au comité du secret.

Le recueil des informations s'effectue chaque année en deux temps, d'abord sur les inscrits en janvier (collectées entre octobre et janvier de l'année considérée), puis sur les résultats (collectées en mai de l'année suivante). Ces informations individuelles doivent remonter avec l'INE[110] afin de permettre de relier les fichiers annuels successifs pour suivre les parcours des étudiants dans l'enseignement supérieur.

[109] Une table de passage a été créée pour permettre de relier les observations désormais identifiées avec l'INE certifié avec celles enregistrées dans les fichiers annuels antérieurs, qui ne disposait pas encore de cet INE.
[110] L'état civil complet des étudiants ne figure pas dans ces remontées. Seul l'INE sert d'identifiant.

Encadré 8 • Des enquêtes de la statistiques publique alimentent SISE

L'élaboration des fichiers SISE s'appuie sur des remontées d'informations issues des logiciels de gestion des établissements. Juridiquement, certains dispositifs constituent des remontées administratives. D'autres prennent la forme d'enquêtes présentées au Cnis :

- *l'enquête sur les effectifs étudiants et leur diplomation des écoles d'ingénieurs privées*. Elle couvre 90 écoles d'ingénieurs privées et 63 000 étudiants (incluses dans la ligne « SISE-ingénieur » du tableau infra). Elle est collectée d'octobre n-1 à janvier n pour les inscriptions de l'année n-1/n et en mai-juin n pour les résultats de la même année scolaire ;

- *l'enquête sur les effectifs d'étudiants et leur diplomation auprès des écoles de commerce et de gestion et autres établissements de l'enseignement supérieur.* Elle couvre 370 écoles de commerce et de gestion, instituts catholiques, écoles vétérinaires, écoles de journalisme, écoles administratives et juridiques qui ne relèvent pas des ministères chargés de l'Éducation, de l'Enseignement supérieur ou de la Culture, soit 230 000 étudiants (pour l'essentiel, répartis entre les lignes « SISE-management », « SISE-univprivées » et « SISE-26bis » du tableau infra). Elle est collectée selon le même calendrier que l'enquête « ingénieur privé » ;

- *l'enquête sur les effectifs d'étudiants et leur diplomation auprès des établissements d'enseignement supérieur artistiques et culturels.* Elle couvre 300 établissements sous tutelle du ministère chargé de la Culture, soit 78 000 étudiants (ligne « SISE-Culture » du tableau infra). Elle est collectée uniquement d'octobre N-1 à janvier N pour les inscriptions de l'année scolaire courante et les résultats de l'année scolaire précédente.

Le taux de réponse des établissements à ces enquêtes serait de l'ordre de 95 % à 100 %.

D'abord centré sur les universités publiques, le champ couvert par SISE a été progressivement élargi :
- les IUFM et instituts catholiques en 1999,
- les écoles d'ingénieurs en 2001,
- les écoles normales supérieures et grands établissements (comme l'Inalco) en 2004,
- les écoles de commerce, de gestion et de « management » en 2006,
- les écoles nationales vétérinaires et l'école nationale supérieure du paysage de Versailles en 2009,
- une partie des écoles d'enseignement supérieur issue de l'enquête dite « 26 » en 2016.

Certains champs ne sont pas encore couverts par SISE. Les établissements concernés transmettent des données au Sies pour lui permettre de produire des statistiques sur les inscriptions et sur les diplômés, mais ne transmettent pas de données individuelles identifiées par l'INE. C'est le cas des formations supérieures des lycées agricoles, qui transmettent des données individuelles sans identifiant. C'est aussi le cas des formations supérieures du secteur paramédical et social et certains établissements de l'enquête 26[111], qui ne transmettent que des données agrégées. Au total, plus d'un millier d'établissements, rassemblant plus de 200 000 étudiants, ne sont pas couverts par SISE, soit 25 % des établissements et 7 % des étudiants (voir tableau ci-dessous[112]).

[111] Cette enquête est désormais aussi une enquête de la statistique publique, qui collecte des informations agrégées auprès des établissements privés qui n'ont pas basculé dans l'un des dispositifs SISE. La collecte des données agrégées s'effectue d'octobre *n*-1 à janvier *n* par un questionnaire en ligne *via* une application internet (fiche d'opportunité présentée à la commission Services publics du Cnis le 12 mars 2020).
[112] Source : Sies - Diaporama présenté à la commission Services publics du Cnis le 12 mars 2020.

Tableau 1 • Champ couvert par SISE-Inscriptions

Synthèse des remontées d'inscriptions de l'enseignement supérieur

Année scolaire 2018-2019

Enquêtes	Nombre d'établissements		Effectifs 2018-2019 (hors DI)		Type de données
Lycées (CPGE-STS)	2 431	56,9%	346 000	12,9%	Données individuelles
SISE-Université	87	2,0%	1 661 000	62,0%	Données individuelles
SISE-ENS	15	0,4%	15 000	0,6%	Données individuelles
SISE-Ingenieur	121	2,8%	142 000	5,3%	Données individuelles
SISE-Culture	300	7,0%	80 000	3,0%	Données individuelles
SISE-Management	142	3,3%	181 000	6,8%	Données individuelles
SISE-univprivées	5	0,1%	32 000	1,2%	Données individuelles
SISE- 26Bis	88	2,1%	25 000	0,9%	Données individuelles
Lycées agricoles	275	6,4%	20 000	0,7%	Données individuelles mais sans INE
Paramédical	386	9,0%	103 000	3,8%	Données agrégées (N-1)
Social	128	3,0%	32 000	1,2%	Données agrégées (N-1)
Enquête 26	293	6,9%	41 000	1,5%	Données agrégées
Total	**4 271**	**100,0%**	**2 679 000**	**100,0%**	

Source : MESRI-SIES

Chaque année le Sies s'efforce de faire basculer des établissements supplémentaires de l'enquête 26 vers SISE. Cependant, un reliquat devrait persister parce que des petits établissements n'ont pas les moyens suffisants pour fournir facilement des données individuelles dans le format demandé et parce que, régulièrement, des nouveaux établissements sont créés. Pour ces derniers, il importe d'abord de disposer de données même agrégées, avant d'envisager de récupérer des données individuelles conformes aux normes SISE[113].

Les informations demandées pour SISE-Inscriptions, toutes issues des logiciels de gestion, concernent :

- l'identification de l'étudiant (INE) et de son établissement ;
- des caractéristiques sociodémographiques de l'étudiant (sexe, nationalité, date et lieu de naissance, PCS des parents, lieu de résidence de l'étudiant et des parents) ;
- des informations sur le passé scolaire de l'étudiant (bac, sa série, son année, son académie, dernier diplôme obtenu antérieurement dans l'enseignement supérieur, 1ère inscription dans l'enseignement supérieur, 1ère inscription dans l'enseignement supérieur français, 1ère inscription dans l'établissement) ;
- des informations sur l'inscription actuelle (diplôme préparé, cursus, niveau dans le diplôme, existence d'un cursus aménagé, conventionnement, régime d'inscription…)

Les informations demandées pour SISE-Résultats, également issues des logiciels de gestion**,** se limitent à l'identification de l'étudiant (INE), de son établissement, du diplôme obtenu et sur, certains champ, l'existence d'une mobilité à l'étranger. Le Sies complète ensuite ces données en utilisant des informations recueillies pour SISE-Inscription.

Des extraits des fichiers annuels SISE-Inscriptions et SISE-Résultats sont accessibles aux chercheurs par la plateforme Quételet-Progedo Diffusion, et plus particulièrement *via* l'Adisp, sa composante chargée de la diffusion des données de la statistique publique[114]. L'ensemble des données individuelles sont accessibles aux chercheurs sous convention avec le Sies, dans des conditions d'accès techniques strictes.

[113] Voir intervention du Sies en commission Services publics du Cnis, en mars 2020.
[114] Le site de l'Adisp met à disposition des guides détaillés sur SISE-Inscriptions et SISE-Résultats pour les universités. Ces guides présentent le dispositif et les variables disponibles dans les fichiers proposés. http://www.progedo-adisp.fr/

Sans entrer dans une présentation détaillée du contenu des fichiers et de leur production, plusieurs points méritent attention avant d'envisager les opportunités offertes par ces fichiers pour le suivi des trajectoires des étudiants, outre les limites déjà évoquées liées au champ couvert (voir *supra*) et à la qualité de l'INE (voir partie 2.1) :

- Les noms et prénoms de l'étudiant, son adresse précise (au-delà de la commune de résidence) ou ses coordonnées téléphoniques ou mail ne sont pas collectées dans les enquêtes SISE.

- Certaines informations sont données par l'étudiant lors de son inscription. Elles ne sont pas toujours bien renseignées (par exemple, la mention attribuée au bac est rarement présente), ni forcément actualisées chaque année pour un élève déjà inscrit dans l'établissement (PCS des parents par exemple).

- SISE couvre tous les inscrits en formation initiale ou en formation continue diplômante d'au moins un an, y compris les apprentis et les formations par alternance. Sont aussi compris les étudiants français et étrangers inscrits dans le cadre de conventions et d'échanges internationaux, dès lors qu'ils ont payé des droits d'inscription pour une formation d'un an, même s'ils sont amenés à suivre une formation à l'étranger durant l'année universitaire. Les auditeurs libres et les stagiaires français ou étrangers venant suivre des cycles de formation d'une durée inférieure à l'année universitaire doivent être exclues.

- Les étudiants inscrits dans le cadre d'une convention entre un institut catholique et une université publique doivent être recensés dans les deux types d'établissement.

- Un étudiant ayant plusieurs inscriptions est présent avec plusieurs observations dans le fichier SISE-Inscriptions et, s'il y a lieu, dans le fichier SISE-Résultats.

- Le Sies choisit une inscription principale pour les étudiants ayant plusieurs inscriptions dans un même établissement. Les autres inscriptions sont dites « secondes ». Il donne priorité aux diplômes nationaux par rapport aux diplômes d'université, puis au diplôme de niveau terminal le plus élevé (avec des exceptions précisées dans le guide utilisateur[115]). L'inscription principale permet de produire des statistiques sur les personnes sans double compte[116].

- Pour un diplôme donné, l'étudiant ne peut être inscrit que dans un établissement. En cas de co-habilitation, l'étudiant ne doit être recensé que dans l'établissement où il a payé ses frais d'inscription. Il en est de même quand l'établissement délègue à un autre une partie de la formation à un diplôme donné.

- Certains redressements sont effectués mais il n'est pas possible d'identifier les informations corrigées dans les fichiers mis à disposition[117]. Par exemple, l'écart entre l'année de naissance et l'année du baccalauréat ne peut être inférieur à 14 ans. En cas de non-réponse[118] sur l'année de première inscription à l'université, une imputation est réalisée (c'est l'année du baccalauréat qui est retenue pour les bacheliers). L'année de naissance peut aussi être redressée en fonction de l'année du baccalauréat et de sa série, etc.

- Pour les élèves bacheliers, l'INE attribué lors de sa première inscription dans le supérieur est celui qui, issu de la BEA, a été utilisé pour l'inscription au bac. Pour les élèves non bacheliers, ceux ayant eu une longue interruption depuis le bac ou ayant passé le bac dans un lycée français à l'étranger, l'INE est attribué par l'université de première inscription**.** La mise en oeuvre du RNIE entrainera une diffusion progressive de l'INE certifié, au fur et à mesure des nouvelles premières inscriptions.

[115] En particulier, les diplômes d'ingénieur et les préparations au concours d'enseignement priment sur les autres diplômes nationaux. À l'opposé, les habilitations à diriger des recherches (HDR) ont un niveau de priorité moindre que les autres diplômes nationaux. Les diplômes d'établissement identifiés spécifiquement (tels que le diplôme de l'IEP de Paris) sont prioritaires par rapport aux diplômes d'université génériques. Les diplômes sélectifs, comme le DUT, aussi.
[116] De ce fait, par exemple, un étudiant peut avoir plusieurs « niveaux d'études » puisque cette variable est calculée en nombre d'années du niveau atteint pour l'inscription considérée par rapport au baccalauréat.
[117] Le Sies dispose des variables brutes.
[118] « Quelques non-réponses » selon le guide utilisateur.

- SISE-Résultats inclut les personnes ayant obtenu leur diplôme même si elles n'entrent pas dans le champ SISE-Inscriptions, par exemple à l'issue d'une formation continue de moins d'un an.

4.3. Le dispositif Inserjeunes sur la performance des formations professionnelles par apprentissage ou par voie scolaire (Depp, Dares)[119]

Le dispositif d'observation « Inserjeunes », produit par la Depp en partenariat avec la Dares, a été développé pour répondre à une obligation légale visant la publication systématique de taux d'insertion des jeunes passées par des formations professionnelles, avec un degré de détail fin allant jusqu'au niveau des établissements.

L'article L. 6111-8 du code du travail, créé en 2016, prévoie ainsi la publication annuelle des « résultats d'une enquête nationale qualitative et quantitative relative au taux d'insertion professionnelle à la suite des formations dispensées dans les centres de formation d'apprentis, dans les sections d'apprentissage et dans les lycées professionnels ». En 2018, l'article 24 de la loi pour la liberté de choisir son avenir professionnel a précisé la commande en demandant une publication de données au niveau de l'établissement de formation[120].

Le dispositif Inserjeunes doit permettre de répondre à cette commande en permettant de calculer et diffuser des taux d'insertion à six, douze, dix-huit et vingt-quatre mois, ainsi que des taux de poursuite d'études, des taux d'interruption en cours de formation et une « valeur ajoutée » par établissement. Ces indicateurs, qui prétendent rendre compte de la perfomance des formations et des établissements, doivent aider les élèves et leur famille dans leurs choix d'orientation. Les premiers résultats, encore partiels, ont été diffusés sur un site internet dédié[121] au début de l'année 2021, avec des taux d'insertion limités à six mois. Le recul est appelé à s'élargir avec le temps.

Les objectifs du législateur rappellent ceux qui ont conduit à la mise en place d'un dispositif d'observation harmonisé sur l'insertion professionnelle des étudiants diplômés de l'enseignement supérieur à partir de 2010, à la suite de la loi du 10 aout 2007 relative aux libertés et responsabilités des universités[122]. La façon d'y répondre est cependant complètement différente, ce qui ne permettra pas de comparer les indicateurs des deux champs. Pour le champ de l'enseignement supérieur, le Sies utilise ainsi des données d'enquêtes, collectées par les établissements d'enseignement supérieur auprès de leurs anciens élèves, sur la base d'un questionnaire et d'un protocole harmonisés[123]. Dans le champ des centres d'apprentissage et des lycées professionnels, la Depp et la Dares ont fait, au contraire, le choix d'un dispositif qui mobilise des sources administratives en appariant les données scolaires sur les inscriptions et sur la réussite aux examens avec les données de la déclaration sociale nominative[124].

Le dispositif Inserjeunes se substitue donc aux enquêtes sur l'insertion professionnelle dans la vie active des lycéens (IVA) et celles sur l'insertion des apprentis (IPA), qui ne permettaient pas de répondre aux nouvelles obligations légales du fait d'un taux de réponse trop faible et d'un manque de recul (le taux

[119] Voir notamment la présentation faite par la Depp à la commission Services publics du Cnis du 12 mars 2020 et le site internet présentant les résultats : https://www.inserjeunes.education.gouv.fr/diffusion/accueil.

[120] « Chaque année, pour chaque centre de formation d'apprentis et pour chaque lycée professionnel, sont rendus publics quand les effectifs concernés sont suffisants : 1° Le taux d'obtention des diplômes ou titres professionnels ; 2° Le taux de poursuite d'études ; 3° Le taux d'interruption en cours de formation ; 4° Le taux d'insertion professionnelle des sortants de l'établissement concerné, à la suite des formations dispensées ; 5° La valeur ajoutée de l'établissement. Pour chaque centre de formation d'apprentis, est également rendu public chaque année le taux de rupture des contrats d'apprentissage conclus ».

[121] https://www.inserjeunes.education.gouv.fr/diffusion/accueil

[122] L'article 20 de la loi du 10 aout 2017 a complété l'article L. 612-1 du code de l'éducation par l'alinéa suivant : « Les établissements dispensant des formations sanctionnées par un diplôme d'études supérieures rendent publiques des statistiques comportant des indicateurs de réussite aux examens et aux diplômes, de poursuite d'études et d'insertion professionnelle des étudiants ».

[123] Depuis 2011, l'enquête auprès des masters bénéficie d'un avis d'opportunité favorable du Cnis. Depuis 2020, elle bénéficie aussi d'un avis de conformité délivré par le comité du Label (réunion du 30 septembre 2020), qui lui permet de disposer du label d'intérêt général et de qualité statistique, ainsi que du caractère obligatoire.

[124] Pour Inserjeunes, ce sont les données issues de la déclaration sociale nominative détenues par la Dares qui sont mobilisés.

d'insertion n'est observé qu'à sept mois). Cette substitution de données administratives à des données d'enquête conduit cependant à des pertes d'information, notamment sur la situation des jeunes qui ne sont pas retrouvés dans les déclarations sociales nominatives et sur les démarches entreprises par les jeunes sans emploi.

Le rapprochement entre les données scolaires et la déclaration sociale nominative ne peut pas s'appuyer sur un identifiant commun, les premières utilisant l'INE et la seconde le Nir. Dans un premier temps, l'INE sert à construire la base des élèves potentiellement sortants en permettant le rapprochement des fichiers des élèves inscrits en voie professionnelle scolaire ou en apprentissage une année donnée avec les fichiers des inscrits de l'année suivante dans les mêmes formations pour repérer les poursuites d'études. Des fichiers d'inscrits dans d'autres types de formation qui utilisent aussi l'INE sont aussi mobilisés pour réperer d'autres cas de poursuites d'études, par exemple dans l'enseignement supérieur avec les fichier SISE. L'INE permet enfin de repérer, pour les années terminales, les élèves inscrits aux examens et leur réussite. Le rapprochement avec la déclaration sociale nominative s'effectue ensuite *via* à un algorithme sur la base du nom, du prénom, de la date et du lieu de naissance. L'algorithme calcule un indicateur de proximité entre les identités des individus présents dans les fichiers scolaires et les identités des individus présents dans la déclaration sociale nominative pour tenir compte d'éventuelles erreurs ou d'incomplétudes lors de la saisie de ces informations dans les systèmes d'information de gestion. Un seuil permet ensuite d'accepter ou non le rapprochement entre un individu des fichiers scolaires donné et un individu de la déclaration sociale nominative. Ces traitements ont été construits pour que les rapprochements soient réalisés automatiquement, sans traitement manuel d'éventuels cas litigieux[125]. Un traitement manuel sur échantillon permet cependant de s'assurer de la qualité du rapprochement automatique.

La base de données individuelle constituée pour produire les indicateurs d'Inserjeunes pourra être accessible sous certaines conditions aux chercheurs. Ces bases comportent des informations sur l'année de sortie des élèves (formation suivie, réussite aux examen) mais pas l'ensemble des informations disponibles sur leur parcours scolaire antérieur. De même, une partie seulement des informations disponibles dans la déclaration sociale nominative est récupérée. Seules des photographies sur la situation à six, douze, dix-huit et vingt-quatre mois sont prévues et non l'ensemble du parcours professionnel salarié sur cette fenêtre temporelle. Pour ces dates, les informations récupérées de la déclaration portent essentiellement sur la profession occupée (PCS), le niveau de rémunération, des caractéristiques du contrat de travail (nature, convention collective applicable) et des caractéristiques de l'employeur (secteur d'activité, taille, commune d'implantation). Techniquement, pour des travaux de recherche, il est possible de compléter les données d'inserjeunes, soit par des informations de parcours scolaires détenues par la Depp, soit par des informations de parcours professionnels détenues par la Dares. Cependant, les exigences de la Cnil de bien séparer la sphère d'usage de l'INE et celle du Nir, garanties par une gestion des identifiants stricts, interdisent un enrichissement simultanée dans les deux sphères.

Deux autres limites sont à souligner :

- Seuls les sortants des formations professionnelles des établissements relevant de l'Éducation nationale et de l'Agriculture sont couverts (niveau V, IV et BTS pour le niveau III). À la différence des enquêtes IVA-IPA, l'apprentissage dans l'enseignement supérieur (niveau II et niveau I et une partie du niveau III) n'est pas couvert. Les sortants des établissements relevant d'autres ministères, notamment la Santé, les Affaires sociales, la Jeunesse et les Sports ne sont pas non plus couverts.
- Aucune information n'est disponible sur la situation des jeunes considérés comme sortants et absents de la déclaration sociale nominative. Or cette double absence peut correspondre à une grande variété de cas : une poursuite d'études non reperée, une absence d'identification par l'algorithme de rapprochement des sources, un emploi hors champ de la

[125] L'algorithme permet de calculer une mesure de proximité entre deux identités. Pour un traitement exclusivement automatique, un seuil unique est déterminé à regard d'experts. D'un côté de ce seuil, les identités sont considérées comme identiques. De l'autres, elles ne le sont pas. Quand des moyens existent pour traiter manuellement les cas litigieux, deux seuils sont fixés, l'un en deçà duquel l'égalité des identités est automatiquement rejetée ; l'autre au-delà duquel l'égalité des identités est automatiquement acceptée. Si la mesure de proximité se situe entre les deux seuils, un traitement manuel est effectué pour trancher. Le choix du seuil ou des seuils dépend en partie des objectifs du traitement. Un rapprochement réalisé à des fins de contrôle, par exemple, n'a pas les mêmes exigences et les mêmes critères qu'un rapprochement réalisé à des fins statistiques.

déclaration sociale nominative (emploi non salarié, emploi à l'étranger), une période de chômage au sens du BIT ou un passage à l'inactivité.

4.4. Le dispositif Force sur l'efficacité de formations données aux demandeurs d'emploi (Dares)[126]

Le dispositif Force (« formation, chômage et emploi ») est un projet de rapprochement de grande ampleur de données administratives, effectué dans le cadre de l'évaluation du Plan d'investissement dans les compétences (PIC). Il vise à reconstituer les trajectoires de formation et d'emploi des personnes en recherche d'emploi, par appariement de plusieurs sources administratives :

- un extrait du fichier historique des demandeurs d'emploi, complété d'extraits du FHA, pour suivre le parcours des demandeurs d'emploi auprès du service public de Pöle emploi ;

- des informations de la base régionalisée des stagiaires de la formation professionnelle (BREST), qui rassemble les caractéristiques des formations suivies par les personnes en recherche d'emploi, qu'elles soient inscrites ou non à Pöle emploi ;

- des informations de la base I-MILO, consacrée aux jeunes en contact avec une mission locale, notamment ceux concernés par un programme d'accompagnement renforcé (Garantie jeunes, PACEA, etc.).

- des informations sur les contrats de travail issues de la base mouvements de main-d'œuvre, elle-même alimentée par la déclaration sociale nominative (données utilisées aussi pour le dispositif Inserjeunes déjà évoqué).

Le champ du dispositif est exhaustif : toutes les personnes présentes dans le fichier historique des demandeurs d'emploi, BREST et la base I-Milo sont concernées. Les informations de la déclaration sociale nominative ne sont récupérées que pour ces personnes.

[126] Dispositif évoqué succinctement dans le programme de travail 2020 des producteurs de la statistique publique présenté à la commission Emploi, qualification et revenus du travail le 23 avril 2020. Voir aussi avec le CASD, où les données sont mises à disposition pour les chercheurs habilités.

Encadré 9 • Les informations contenues dans le dispositif FORCE

La base de données Force comprend tous les demandeurs d'emploi présents dans le fichier historique statistique de Pöle emploi, soit tous ceux qui ont été inscrit à Pôle emploi au moins une fois au cours des dix dernières années. Pour ces demandeurs d'emploi, les informations disponibles concernent les caractéristiques des demandeurs d'emploi (sexe, date de naissance, qualification/diplôme, lieu de résidence...) ; les caractéristiques des demandes d'emploi (dates et motif d'inscription, catégorie d'inscription, type d'emploi recherché, date et motif de sortie des listes) ; l'existence de droits à une indemnisation chômage ; des informations sur les contacts et entretiens du demandeur d'emploi avec Pöle emploi ; des informations sur les prestations et services dont le demandeur d'emploi a bénéficié, notamment les formations suivies par le demandeurs d'emploi indemnisés.

Sont aussi intégrés tous les stagiaires de la formation professionnelle figurant dans le fichier BREST depuis janvier 2017. Ce fichier est construit à partir des fichiers de gestion de rémunération ou de protection sociale des personnes en recherche d'emploi qui bénéficie à ce titre d'une formation professionnelle, qu'elles soient inscrites ou non à Pöle emploi. Chaque nouvelle entrée en formation fait l'objet d'une nouvelle observation, avec des informations sur le bénéficiaire (sexe, âge, lieu de résidence, plus haut diplôme obtenu) et sur la formation suivie (dates, durée, domaine, objectifs).

Enfin, sont également intégrés tous les jeunes de 18 à 25 ans ayant été inscrits au moins une fois auprès d'une mission locale depuis le mois de janvier 2016, via la base I-MILO. Les informations disponibles portent sur les caractéristiques du bénéficiaire du suivi, dont le diplôme le plus élevé à l'entrée de chacun des dispositifs dont il a bénéficié ; ces dispositifs dont il a bénéficié (Civis, garantie jeune, parcours contractualisé d'accompagnement vers l'emploi et l'autonomie, programme local…), avec les dates d'entrée et de fin ; et les contacts qu'a eu le bénéficiaire avec la mission locale.

Les informations de la déclaration sociale nominative disponibles dans le champ de la base des mouvements de main-d'œuvre sont disponibles à partir de 2017 pour toutes les personnes figurant dans l'une ou l'autre des trois bases précédentes. Ici, le contrat de travail est l'observation de base. Contrairement à d'autres bases utilisant la déclaration sociale nominative, il n'y a pas de sélection de certaines périodes d'emploi privilégiées. Pour chaque contrat, quelques caractéristiques du salarié et du contrat sont disponibles, notamment la nature du contrat, la date de début et de fin, la PCS, le salaire, le SIRET de l'établissement employeur, le SIRET de l'établissement utilisateur pour les contrats d'intérimaires.

Comme le Nir n'est pas toujours disponible dans les fichiers mobilisés, le centre d'accès sécurisé aux données (CASD), qui intervient comme tiers de confiance, procède à leur rapprochement de façon différente selon les cas :

- pour associer les données de Pôle emploi et celles de la déclaration sociale nominative, le CASD utilise le Nir haché ;

- pour rapprocher les données sur les stagiaires de la formation professionnelle et celles relatives aux jeunes en contact avec les missions locales avec les données de Pôle emploi, le CASD utilise un algorithme mobilisant les noms, prénoms et date de naissance ;

- le même algorithme est utilisé pour rapprocher les données de la déclaration sociale nominative avec celles des stagiaires de la formation professionnelle et celles relatives aux jeunes en contact avec les missions locales qui n'ont pas été identifiés dans les données de Pôle emploi lors de la phase précédente. Pour les autres, le rapprochement avec la déclaration sociale nominative a été fait lors de la première phase.

Les noms, prénoms et date de naissance des personnes ne sont pas conservés dans les fichiers Force. Dans ces bases, les individus ne peuvent être repérés que par les identifiants non signifiants propres à chacune des sources utilisées et par l'identifiant non signifiant unique créé par le CASD pour pouvoir articuler les extraits de ces différentes sources.

Difficile à évaluer sans information disponible sur le sujet, la qualité de l'articulation des extraits issus des différentes sources est nécessairement variable selon les sujets traités car elle dépend nécessairement de la nature des informations identifiantes mobilisées pour rapprocher les données (le Nir ou une partie de l'état civil). Elle dépend aussi de la qualité de la saisie de ces informations identifiantes dans les systèmes de gestion.

L'accès aux fichiers Force est possible pour les chercheurs sélectionnés par le comité scientifique du Plan d'investissement dans les compétences. Il s'effectue par l'intermédiaire du CASD après passage au du comité du secret.

4.5. Le dispositif Trajam sur les trajectoires des jeunes bénéficiaires des politiques d'emploi (Dares)

Développé par la Dares, le panel Trajam[127] doit permettre de suivre et d'évaluer les trajectoires professionnelles des jeunes bénéficiaires des politiques d'emploi en faveur de l'insertion, des emplois aidés et des mesures d'accompagnement : passage par l'Epide, les écoles de la deuxième chance, les missions locales, Pôle emploi.

Pour les démarches que la Dares a réalisées auprès de la Cnil, le projet Trajam est inscrit dans le cadre de la mise en œuvre du plan de la Garantie européenne pour la jeunesse (GEJ). Adopté en 2013, ce plan incite les États à mettre en place un système d'accompagnement des jeunes qui ne sont ni en emploi, ni inscrits dans le système éducatif ou en formation (« NEET »). La Commission européenne attend aussi de chaque État qu'il rende compte du devenir des jeunes après l'entrée dans une solution d'emploi ou de formation.

Le panel Trajam est construit par rapprochement des fichiers administratifs des dispositifs pris en compte, du fichier historique des demandeurs d'emploi et du panel DADS « Tous salariés » de l'Insee.

Seule une sélection restreinte d'informations est retenue dans chacun des fichiers sources, parce que l'objectif premier est d'utiliser les informations relatives aux différents dispositifs de façon conjointe ou comparative. Il ne s'agit pas d'une base de données destinée à des travaux d'évaluation approfondis segmentés qui cibleraient uniquement tel ou tel dispositif. D'ailleurs, le panel se limite aux bénéficiaires des dispositifs âgés de 16 à 25 ans, dont la trajectoire professionnelle doit être ensuite suivie jusqu'à l'âge de 35 ans. Il ne couvre donc pas l'ensemble des bénéficiaires de certains dispositifs ciblés, comme le contrat unique d'insertion, le contrat de professionnalisation, l'insertion par l'activité économique ou accompagnement par Pôle emploi.

Le dispositif Trajam doit ainsi, par exemple, permettre d'estimer le nombre de jeunes aidés sans double compte, le nombre de jeunes bénéficiant de plusieurs dispositifs ou l'identification d'éventuels parcours d'accompagnement type.
L'appariement au panel « Tous salariés » conduit à restreindre le champ aux jeunes nés en octobre des années paires ou un jour « EDP », soit *grosso modo* 1/12^{e} de la population.

Les fichiers sont rapprochés sur la base du Nir, disponible directement ou récupéré indirectement. Pour les fichiers qui ne disposent pas du Nir, une recherche préalable est effectuée au répertoire national d'identification des personnes physiques sur la base de l'état civil complet[128], après un travail de mise au propre et de normalisation de ces informations. Naturellement, le Nir et les données directement identifiantes ne sont pas maintenues dans les bases une fois que le panel est construit. Un Nir « haché »

[127] TRAjectoires des Jeunes Appariées aux Mesures actives du marché du travail.
[128] Noms, prénoms, date de naissance, sexe, commune de naissance ou pays de naissance pour les étrangers.

leur est substitué. Le code statistique non signifiant prévu par l'article 34 de la loi pour une République numérique est appelé à remplacer ce Nir haché.

Les individus du panel sont affectés d'un poids puisqu'il s'agit d'un échantillon. La pondération permet aussi de redresser les données afin de tenir compte de la plus ou moins grande facilité à retrouver le Nir pour certaines catégories de jeunes[129] et pour assurer un calage sur les effectifs disponibles dans les bases exhaustives.

Les trajectoires d'emploi sont observées, ici comme dans d'autres dispositifs, sous le prisme des données administratives mobilisées. Les emplois hors champ du panel « Tous salariés » ne sont donc pas repérés, notamment les emplois non salariés ou à l'étranger. Les périodes de chômage sans inscription à Pôle emploi, plus fréquentes chez les jeunes qu'à d'autres tranches d'âge, ne le sont pas non plus, pas plus que les périodes d'inactivité, certains retours en formation initiale ou en formation continue. Corrélativement, les périodes de « NEET » au sens d'Eurostat ne sont donc pas directement identifiables. Encore une fois, cette limite n'est pas rédhibitoire compte tenu de l'importance de l'emploi salarié dans l'emploi total et de la place de Pôle emploi.

L'accès aux données sera possible aux chercheurs par l'intermédiaire du CASD[130].

[129] Les trajectoires des jeunes nés à l'étranger sont de moins bonne qualité, la possibilité de les retrouver au RNIPP étant plus difficile (l'Insee retrouve en moyenne 90 % des Nir pour les Français, et seulement 60 % pour les étrangers).
[130] Voir l'arrêté du 6 octobre 2016 relatif à la mise en œuvre de ce traitement (après délibération n° 2017-210 du 13 juillet 2017 de la Cnil) et l'arrêté modificatif du 17 août 2017 (après délibération n° 2017-210 du 13 juillet 2017 de la Cnil).

5. En guise de conclusion : quelques perspectives et enjeux

5.1. Deux opportunités majeures dans le champ des travaux du Céreq

Avec la mise en œuvre du règlement général sur la protection des données et la loi Numérique, l'évolution du cadre juridique offre un cadre théoriquement plus favorable au développement de l'utilisation des données administratives à des fins d'études et de recherche, même si ces opportunités sont dépendantes de la bonne volonté des responsables des données et des points de vue, pour une part subjectifs, des délégués à la protection des données des parties prenantes.
De fait, les projets se sont multipliés ces dernières années au sein de la statistique publique et, avec eux, des opportunités nouvelles. En particulier, dans les champs d'expertise du Céreq, deux projets apparaissent majeurs : le déploiement d'un INE unique certifié dans les bases de gestion des établissements scolaires et de l'enseignement supérieure d'une part, et le projet Agora, d'autre part.

Le déploiement de l'INE unique certifié permet d'envisager dans un horizon de moyen terme de mieux suivre l'essentiel du parcours scolaire des jeunes à partir des fichiers construits par la Depp et le Sies. Le recours à ces informations pourrait notamment permettre d'envisager des améliorations substantielles du dispositif d'enquêtes Génération, sur l'insertion professionnelle des jeunes à l'issue de leur formation initiale, en permettant d'enrichir les informations disponibles par rapport à celles collectées actuellement, de réduire la taille du questionnaire consacré au recueil des informations sur le parcours scolaire, voire d'améliorer aussi le traitement de la non-réponse. Sur ces sujets, les expérimentations envisagées dès cette année devraient permettre d'apporter des premiers enseignements.

Pour que cette opportunité soit pleinement effective et facilitée, une évolution de l'arrêté du 30 juillet 2018 portant création du traitement automatisé de données à caractère personnel dénommé « Système d'information sur le suivi des étudiants » (SISE), serait toutefois bienvenue afin d'ajouter un nouvel alinéa dans l'article 4 portant sur les destinataires possibles des données. Il s'agirait d'indiquer explicitement que le Céreq peut accéder aux données individuelles à des fins d'enquêtes effectuées dans le cadre de ses missions fixées par l'article R313-38 du code de l'Education.

De son côté, Agora est appelé à devenir un silo central de données relatives à la formation professionnelle. Il devrait donc devenir le pivot d'exploitations statistiques diverses à des fins d'études et d'évaluations nationales, sectorielles ou territoriales. Le Céreq devrait donc être conduit à terme à exploiter régulièrement ses données, dans le cadre d'une articulation nécessaire avec les travaux propres de la Dares et de France compétences, d'autant que dans le cadre strict des textes en vigueur, ces travaux nécessiteraient de conventionner avec la Dares ou s'effectuer en sous-traitance pour France compétences.

De leur côté, les dispositifs Inserjeunes, Force ou Trajam n'ouvrent pas des opportunités aussi structurantes pour les travaux du Céreq, leur intérêt principal se limitant aux objectifs ciblés qui ont conduit à leur mise en place :

- *Inserjeunes* couvre un champ très ciblé de la formation initiale. Il propose des photos ponctuelles de la situation « dans la DSN », qui ne rendent compte, ni de la trajectoire complète des élèves après leur sortie, ni de l'ensemble des situations possibles (emploi non salarié, chômage ou inactivité, etc.). Il ne comporte pas d'informations complètes sur le parcours scolaires antérieur et ne peut pas servir de base de sondage. Très classiquement pour une source administrative, son principal apport est de proposer des informations limitées mais sur des niveaux « géographiques » fins, là où l'enquête Génération permet de produire des informations plus riches et homogènes sur l'ensemble des niveaux de formation, sans permettre la même finesse territoriale.

- *Force* couvre un champ beaucoup plus large puiqu'il doit permettre de suivre la trajectoire professionnelle de tous les jeunes bénéficiaires d'une politique d'accompagnement vers l'emploi (Pôle emploi, formation continue, missions locales), là encore sous le seul prisme des données administratives, la DSN et Pôle emploi, avec leurs avantages et leurs inconvénients. La connaissance du parcours scolaire se limite à l'information sur le niveau de diplôme collectée à l'entrée de l'un ou l'autre dispositif.

- *Trajam* propose un dispositif de même type sur les jeunes bénéficiaires des politiques d'aides à l'insertion ou des emplois aidés (Epide, écoles de la deuxième chance, missions locale, Pôle emploi). Il ne couvre pas la formation professionnelle des demandeurs d'emploi non inscrits à Pôle emploi. Le choix de recourir au Panel DADS plutôt qu'à la DSN conduit à se restreindre à un échantillon au 12ème de bénéficiaires sélectionnés selon leur jour et mois de naissance. Là encore, la connaissance du parcours scolaire est réduite et les situations d'emploi et de non-emploi restreintes aux champs des sources administratives mobilisées. Construit pour faciliter les analyses transversales aux différents dispositifs, Trajam ne propose qu'une sélection des informations disponibles dans les bases de gestion de chacun des dispositifs considérés.

Du fait de leur champ ciblé, ces dispositifs ne sont pas structurants pour les travaux du Céreq car ces sources ont été construites pour répondre à des objectifs ciblés particuliers, mais il reste intéressant de les exploiter dans ce cadre pour se les approprier car les services producteurs peuvent en faire des bases de travail dans le cadre de certains appels à projets.

5.2. Des enjeux à prendre en compte collectivement

Les trois dispositifs cités précédemment, construits à partir d'appariements de fichiers administratifs, permettent d'illustrer quelques enjeux clés pour la statistique publique en général et pour le Céreq en particulier, qu'il serait pertinent de traiter collectivement pour que le déploiement de l'usage des données administratives s'effectue de façon ordonnée et à moindre coût pour la collectivité.

1° S'organiser pour bien connaître les données administratives utilisées

Le bon usage des données administratives nécessite, comme pour les données d'enquête, de bien connaitre leur processus de production, de la collecte initiale de la donnée à la construction du fichier final proposé aux statisticiens et aux chercheurs. À défaut, les données peuvent être mal comprises et mal utilisées, conduisant à des analyses erronées[131]. Quel est le champ couvert par les données ? Qui fournit l'information initiale et dans quel cadre réglementaire et opérationnel ? Il y a-t-il des référentiels utilisés ou des formats à respecter et lesquels ? Le processus de gestion prévoit-il des contrôles sur les informations collectées ? Si oui, lesquels ? Sont-ils bloquants ? Le gestionnaire peut-il les contourner ? Quelles informations sont les plus importantes dans le processus de gestion et le plus susceptibles de sanction en cas de mauvaise déclaration ? Les informations peuvent-elles être modifiées ou corrigées au cours du temps ? À partir de quel délai, les informations peuvent-elles être considérées comme stabilisées ? Comment s'effectue le passage des données de gestion aux données proposées pour des exploitations à des fins de recherche ou de production statitique ? En particulier, il y a-t-il des contrôles de formats et de cohérence, un filtrage des observations, des redressements éventuels de valeurs jugées anormales, des imputations de données manquantes, la construction de variables synthétiques et, dans tous ces cas, quels sont précisement les traitements opérés ? Ou encore, ces fichiers peuvent-il faire l'objet de révisions ultérieures ?

Les questions sont nombreuses. Pour les producteurs de données statistiques, la première exigence d'un processus qualité est de disposer des réponses à ces questions afin de connaître les données qu'ils traitent et pour fournir ces métadonnées à leurs utilisateurs. Cette exigence, qui relève d'une bonne pratique professionnelle, est rappelée par les politiques *d'Open* data, dont le plan national pour la science ouverte annoncé en juillet 2018, pour les données de la recherche. Elle va aussi de pair avec

[131] Le cas de l'évolution des statistiques du chômage et des demandeurs d'emploi en 2006, avec les polémiques qu'elle a engendrées, l'illustre par exemple- les controverses ont cet intérêt de rendre les réalités plus saillantes. Voir Collectif Lorraine Data (2009), *Le grand trucage*, Editions La Découverte // Debauche E., Deroyon T. & Mikol F. (2008), « Retour sur l'évolution du nombre de demandeurs d'emploi inscrits à l'ANPE en 2005 et 2006 », *Document* d'études, n°142, DARES, novembre.

certains des principes posés par le code des bonnes pratiques de la statistique européenne d'Eurostat[132], principes qui s'imposent à la statistique publique. Bien connaître les processus de production des données et les processus de gestion qui les génèrent nécessite cependant des investissements parfois importants, notamment pour les processus qui évoluent régulièrement, comme c'est le cas pour les politiques d'emploi ou les processus de gestion de Pôle emploi. De ce fait, la phase de prise de connaissance est trop souvent négligée. Si le développement du recours aux données administratives se poursuit, la façon d'organiser la production et la mutualisation des métadonnées pertinentes et leur portée à connaissance des utilisateurs, devra donc être traité, notamment pour les fichiers pivots les plus utilisés.

2° Choisir la méthode d'identification des personnes dans les sources que l'on souhaite rapprocher

Dans la sphère scolaire, l'INE constitue un identifiant utilisé dans de nombreuses bases de gestion mais son caractère unique et certifié est récent et il n'est pas encore généralisé, d'autant que certains types d'établissements d'enseignement ne l'utilisent pas. Dans la sphère sociale, le Nir est disponible dans certaines sources mais pas systématiquement en raison des règles strictes encadrant son usage. Dès lors qu'un identifiant certifié unique et commun aux fichiers à rapprocher ne peut pas être utilisé, la bonne identification des personnes devient un enjeu important pour la qualité des rapprochements obtenus. À défaut, elle est susceptible de biaiser les résultats obtenus, notamment si la qualité des informations identifiantes est hétérogène selon les catégories de personnes ou de publics concernées. C'est par exemple le cas des données d'état civil qui sont moins précises pour les personnes nées à l'étranger.

Quand les données complètes de l'état civil sont disponibles, une identification au répertoire des personnes physiques de l'Insee est techniquement envisageable, dans un cadre juridique très strict. Il est alors possible de récupérer le Nir ou un identifiant non signifiant directement dérivé, avant de procéder à un rapprochement sur la base de cet identifiant. Un bémol demeure car l'identification par l'attribution d'un Nir ne renvoie pas toujours un écho unique, y compris pour des personnes nées en France, par exemple en raison d'erreurs ou d'incomplétudes dans les informations d'état civil saisies dans les fichiers de gestion. Des choix sont alors nécessaires pour décider de l'issue à donner au rapprochement des observations dans cette situation.

De façon plus générale, en l'absence d'identiants communs, les rapprochements de données doivent s'appuyer sur une sélection d'informations directement ou indirectement identifiantes, disponibles dans les fichiers à rapprocher. Le sexe, le nom, le prénom, la date de naissance, le lieu de naissance et l'adresse font partie des informations les plus fréquemment utilisées. Pour ces seules variables, outre les éventuelles erreurs de saisie, il faut faire avec les cas où tous les prénoms ne sont pas toujours saisis ; où le nom de naissance et le nom d'usage ne sont pas toujours tous les deux disponibles ; où la date de naissance n'est pas toujours complète ; où le lieu de naissance est plus ou moins précis ; où les façons de saisir les adresses peuvent varier selon les fichiers administratifs utilisés. De ce fait, quelle que soit la technique utilisée[133], des arbitrages sont nécessaires pour se positionner entre des règles trop strictes qui conduiraient à ne pas rapprocher de nombreuses données qui auraient pu l'être, et des règles trop souples, qui rapprocheraient beaucoup de données qui ne devraient pas l'être.

[132] En particulier, le principe 7, relatif au recours à des méthodologies solides pour la production de statistiques de qualité, appelle les autorités statistiques à renforcer la coopération avec la communauté scientifique afin d'améliorer la méthodologie et l'efficacité des méthodes employées. Or la science avance d'autant mieux, que le partage des connaissances se fait largement, plutôt que dans un cercle restreint d'experts choisis. Le principe 8, relatif à la mise en place de procédures statistiques adaptées, mentionne, parmi ses items, la nécessité d'une bonne gestion des métadonnées tout au long des processus statistiques. Le principe 15, relatif à la clarté et l'accessibilité des données, demande explicitement la mise à disposition des métadonnées, y compris sur les méthodologies employées et s'inscrit dans la logique des données ouvertes.

[133] Très schématiquement, deux grandes approches sont possibles. Une première approche consiste à construire une clef d'identification par agrégation des variables identifiantes retenues (par exemple la combinaison du nom, du prénom et de la date de naissance). Les fichiers sont ensuite appariés sur cette clef, comme ils le seraient sur le NIR ou l'INE. Après un premier passage qui permet de rapprocher une partie des données, d'autres peuvent se succéder en assouplissant les variables composant la clef d'identification pour tenir compte des éventuelles erreurs ou trous de saisies (par exemple, au premier tour, la date de naissance complète est prise, puis seulement l'année et le mois). La seconde approche consiste à calculer une mesure globale de proximité entre chacun des individus contenus dans une table et chacun de ceux contenus dans l'autre table sur la base des variables choisies pour l'identification. Pour chaque paire, une distance est calculée. Il faut ensuite fixer un seuil en deçà duquel on considère que les individus sont identiques. Si les ressources sont disponibles pour un traitement manuel, une fourchette peut être retenue pour identifier des cas litigieux à examiner à dires d'experts.

Ces choix sont subjectifs. Ils doivent tenir compte des variables mobilisables et de la qualité de renseignement de ces variables dans les sources à rapprocher. Ils doivent tenir compte aussi des objectifs du rapprochement : veut-on minimiser le risque d'un rapprochement erroné ou est-on moins sensible à ce risque en raison d'un objectif avant tout statistique ? La possibilité d'opérer ou non un traitement manuel sur certains cas tendancieux, joue aussi. En fonction de la taille des fichiers, des contraintes techniques liées au temps de calcul peuvent aussi intervenir. Dans ces conditions, même si certains organismes développent des outils mutualisés[134], ils proposent toujours une palette de méthodes car il n'existe pas de solution unique disponible sur étagère, ni de méthode automatique qui permettrait de définir la méthode optimale à utiliser.

Dans ce contexte, il y aurait intérêt à développer les échanges de pratiques, pour mutualiser les enseignements des opérations de rapprochements déjà effectués. Cet intérêt serait d'autant plus judicieux que certaines sources sont appelées à avoir un rôle pivot et sont déjà très sollicitées dans les projets de rapprochements de données administratives, notamment la DSN et le fichier historique des demandeurs d'emploi pour les études et les recherches s'intéressant aux trajectoires professsionnelles. En toute logique, cette mise en commun des pratiques devraient relever de l'Insee[135]. Le Céreq pourrait aussi organiser, le cas échéant, des rencontres sur le sujet sur des projets susceptibles d'apporter des éclairages utiles aux travaux qu'ils pourraient mettre lui-même en oeuvre.

Si la question du choix de la méthode est importante, selon les praticiens, c'est cependant la phase de préparation des données identifiantes, pour qu'elles se conforment à des référentiels communs dans les sources à rapprocher, qui est la plus cruciale et chronophage.

3° Tirer les conséquences du rôle central de la DADS/DSN et du fichier historique des demandeurs d'emploi dans le suivi des trajectoires professionnelles

Sur les champs de compétences du Céreq, les DADS/DSN et les fichiers historiques des demandeurs d'emploi (FHS ou FHA) seraient les deux sources administratives susceptibles d'être mobilisées le plus souvent. Même si elles ne permettent pas d'identifier l'ensemble des situations d'emploi et de non-emploi possibles, l'importance du salariat dans l'emploi et du rôle de Pôle emploi dans le service public de l'emploi suffisent à en faire des sources pivots pour étudier les différences de trajectoires professionnelles selon les caractéristiques des personnes ou selon leur passage par tel ou tel dispositif. Inserjeunes, Trajam et Force sont autant d'illustrations du rôle pivot de ces sources.

Ces trois dispositifs, ajoutés à l'échantillon démographique permanent et le panel « Tous salariés », montrent aussi qu'une même source administrative, la déclaration sociale nominative, peut être mobilisée à des fins statistiques à différentes étapes de traitement et avec des informations brutes plus ou moins retravaillées : les informations mensuelles récupérées par la Dares ; les informations retraitées et corrigées par l'Insee pour produire un fichier annuel « Tous salariés » ; les informations de ce fichier annuel synthétisées pour le panel « tous salariés » ; les informations de ce panel synthétisé davantage dans l'échantillon démographique permanent. Une information apparemment de même origine est donc plus ou moins contrôlée, retravaillée et synthétisée selon la base de données utilisée pour effectuer les analyses. Ainsi, des taux d'emploi à *n* mois s'affichant comme issus des déclaration sociales nominatives pourraient se multiplier, sans qu'ils soient comparables. La complexité des données de Pole emploi pourrait conduire au même constat si leur utilisation se répend. Là encore, il pourrait être utile d'harmoniser les pratiques.

[134] Par exemple, StatCanada a mis en place un centre de ressources en couplage d'enregistrements (CRCE) : https://www150.statcan.gc.ca/n1/pub/12-206-x/2018001/02-fra.htm. Au niveau européen, Eurostat propose l'outil RELAIS (Record Linkage At IStat), développé par Istat, l'institut national de statistique italien : https://www.istat.it/it/files/2015/04/Insights-on-Data-Integration-Methodologies.pdf
https://www.istat.it/it/metodi-e-strumenti/metodi-e-strumenti-it/elaborazione/strumenti-di-elaborazione/relais

[135] Selon le 2e de l''article 1er du décret n° 46-1432 du 14 juin 1946, l'Insee doit « coordonner les méthodes, les moyens et les travaux statistiques des administrations publiques et des organismes privés subventionnés ou contrôlés par l'État […] ». Ce rôle de coordination des instituts nationaux de statistique est aussi rappelé dans le code des bonnes pratiques de la statistique européenne (principe 1bis).

Une transparence sur les méthodes et une bonne communication sur les métadonnées, déjà évoquée, peuvent y contribuer. Privilégier le même fichier source pour une problématique donnée pourrait également y contribuer. Ainsi, sauf les cas particuliers nécessitant des résultats urgents[136], il semblerait plus raisonnable de privilégier les données spécifiquement contrôlées et retraitées à des fins statistiques. De ce point de vue, s'agissant du repérage des trajectoires salariés, les fichiers annuels exhaustifs « Tous salariés » produits par l'Insee devraient jouer le rôle principal, davantage que la DSN produite par la Dares ou le panel « Tous salariés ».

4° Articuler intelligemment les sources administratives et les données d'enquête

Nous ne redétaillerons pas ici les insuffisances qu'il y aurait à ne recourir qu'aux seules sources administratives pour suivre les trajectoires professionnelles ni, de façon plus générale, les intérêts et les limites des sources administratives par rapport aux données d'enquêtes. Rappelons simplement que les données d'enquêtes permettent au statisticien de maitriser les concepts, les nomenclatures et plus largement, les informations qu'il entend collecter mais que la qualité de ces informations dépend notamment d'un taux de réponse qu'il ne maîtrise pas, tandis que le recours à l'échantillonage limite la capacité à détailler les résultats. *A contrario*, les statisticiens n'ont aucune maîtrise sur les données administratives, même s'ils sont parfois associés à leur construction, mais ces données sont disponibles généralement plus fréquemment et sont exhaustives sur leur champ, ce qui permet de produire des résultats à des échelons très fins. Rappelons enfin, que la mise en avant de la contrainte budgétaire pour privilégier les données administratives sur les données d'enquêtes résulte de l'idée fausse que ces données seraient presque gratuites en négligeant les coûts importants d'appropriation et de retraitement qu'il faut dépenser pour que ces données puissent être utilisées à des fins statistiques.

En réalité, il faut plutôt réfléchir à une articulation intelligente entre les données d'enquêtes et les données administratives. À l'enquête, de collecter les informations les plus essentielles que l'on souhaite homogènes et comparables pour l'ensemble des répondants et d'une enquête à l'autre. À l'enquête de collecter aussi les informations jugées essentielles qui sont absentes des données administratives mobilisées. Aux données administratives, d'enrichir les informations collectées et de permettre de mieux traiter la non-réponse ou de proposer des résultats détaillés avec des biais que l'enquête permet de documenter à un niveau moins fin.

Cette articulation peut se faire, soit en complétant par une enquête sur échantillon, des informations administratives préalablement compilées, soit de façon inverse, en enrichissant les données d'enquêtes par des données administratives. Dans ce deuxième cas, il est important d'anticiper les adossements à venir avec des sources administratives, non seulement pour bien informer les enquêtés de l'ensemble du dispositif envisagé, mais aussi pour s'assurer de bien collecter les variables d'identification qui serviront ensuite à rapprocher les données. Au Céreq, l'enquête Génération est la première concernée par cette réflexion. L'encadré 10 en présente quelques pistes, dont la faisabilité dépendra des conclusions des travaux expérimentaux amorcés en 2021 et de l'accord des partenaires potentiellement concernés.

[136] Ces situations sont rares puisque peu de décisions de politiques publiques sont déterminées sur la base d'indicateurs de pilotage quantitatifs issus de la statistique publique ou de travaux d'évaluations issue de la recherche.

Encadré 10 • Quelques pistes de réflexion pour une articulation entre l'enquête Génération et les données administratives

1. *Enrichir les données de l'enquête Génération par les informations du parcours scolaire et de la diplomation disponibles dans les fichiers de la Depp et du Sies.* Cet enrichissement devrait privilégier les fichiers synthétiques que les services statistiques ministériels produisent à des fins de recherche et de production statistique, plutôt que sur des fichiers amonts, afin de profiter des traitements effectués par les services statistiques, *a priori* mieux placés pour bien connaître les données.

2. *Maintenir dans l'enquête un nombre limité de questions sur les étapes jugées clefs du parcours scolaire quand cela est nécessaire pour garantir l'homogénéité de ces informations pour l'ensemble des répondants*, ceux qui relèvent de champs bien couverts par les sources administratives mobilisées et les autres. Le questionnaire ne doit pas être vu comme un moyen de compléter les trous d'observation des sources administratives. Une telle conception compliquerait le processus de production ainsi que les spécifications du questionnaire, sans permettre d'assurer l'homogénéité nécessaire de l'information.

3. *Maintenir dans l'enquête les questions nécessaires à la fluidité de la passation du questionnaire pour les enquêteurs et pour les enquêtés*. Par exemple, la mise en place d'un questionnement subjectif sur le déroulement de l'orientation en fin de troisième ou après le bac implique de poser des questions sur les choix d'orientation des élèves même si ces informations seraient mobilisables par ailleurs.

4. *Collecter dans l'enquête les informations détaillées sur la situation d'emploi ou de non-emploi à la date d'enquête*, ainsi que les informations détaillées sur le premier emploi (voire du premier CDI) car les sources administratives ne permettent pas d'identifier parfaitement l'ensemble des situations d'emploi et de non-emploi et parce que ces informations sont les informations majeures de l'enquête, utilisées pour la production de ses principaux indicateurs.

5. *Corrélativement, ne pas chercher à multiplier les appariements pour essayer de couvrir toutes les situations possibles par des fichiers administratifs*, en cherchant aussi à récupérer les informations des missions locales, des contrats aidés, etc. Une telle approche alourdirait considérablement les travaux de production sans effets garanties, d'autant que chaque nouvelle source aura ses propres échecs d'appariements.

6. *Réexaminer l'intérêt de collecter l'intégralité du calendrier professionnel.* Il complexifie le questionnaire et la collecte. Apparier l'enquête Génération avec les fichiers « Tous salariés » annuels de l'Insee et le fichier historique des demandeurs d'emploi de Pôle emploi permettrait de reconstruire des parcours. Certes, certaines périodes d'emploi et de non-emploi ne seraient plus identifiées mais les périodes couvertes par les sources administratives seraient sans doute mieux connues, notamment pour les parcours les plus heurtés. De plus, le plus important dans le cas d'une telle substitution n'est pas de récupérer les mêmes informations qu'auparavant mais de pouvoir continuer à montrer les différences de parcours en termes de vitesse d'accès à l'emploi ou, au contraire, d'éloignement durable du marché du travail, selon la formation initiale suivie et les caractéristiques sociales des jeunes.

7. *Procéder par étape*.

Annexe 1. Appel du collectif « STOP FICHAGE 13 » contre la base élèves (7 octobre 2008)

Appel repris sur le site internet du syndicat SUD éducation Val-de-Marne le 16 octobre 2008 https://www.sudeducation94.org/BASE-ELEVES-PREMIER-MAILLON-VERS.html.

BASE ÉLÈVES PREMIER MAILLON VERS UN FICHAGE GÉNÉRALISÉ !

Depuis plusieurs mois déjà, de nombreuses personnes (parents, enseignants, élus, syndicats, collectifs, associations ou citoyens) se mobilisent contre Base élèves. La mise en place de Base élèves se poursuit dans toute la France en contradiction avec les annonces du Ministre de l'Education Nationale : L'arrêté annonçant la modification de Base élèves n'est toujours pas paru.

Seuls 3 des 59 champs ont été retirés de l'application.

Le fichier reste partageable (en partie avec les mairies).

L'Inspection Académique continue d'envoyer injonctions et menaces aux directeurs qui ne remplissent pas la base.

Le Ministère de l'Education Nationale et l'Inspection Académique exigent des directeurs qu'ils alimentent Base élèves sans demander parallèlement que l'information des parents soit faite conformément aux exigences de la CNIL.

BNIE ATTENTION DANGER ! Au-delà de Base élèves, le danger est aussi la constitution d'un répertoire national (BNIE) des identifiants des élèves (INE). Le numéro d'identification des élèves, ainsi que les principales informations contenues dans Base élèves sont enregistrés dans La BNIE. Alors que les données de Base élèves devaient être détruites à la sortie de l'école primaire, ces mêmes données sont conservées dans la BNIE pendant 35 ans, avec un numéro unique de la maternelle jusqu'à la fin des études supérieures. Il est donc possible de suivre les élèves tout au long de leur parcours scolaire, et même au-delà. La CNIL n'autorise pas le ministère à utiliser le NIR (numéro de sécurité sociale), ce qui permettrait de suivre les anciens élèves dans le monde professionnel. Mais d'ici 35 ans tout est possible, il suffira qu'une autre administration demande à avoir accès à l'INE et l'interconnexion deviendra effective.

Base élèves, fichés... dès 3 ans !

La création du fichier Edvige, de Sconet au collège, la mise en place du dossier scolaire électronique et la loi de prévention de la délinquance confirment nos craintes de voir l'école utilisée comme premier maillon du fichage de la population.

Le développement de tous ces fichiers ne correspond ni aux missions fondamentales de l'Éducation nationale, ni à l'idéal d'une société respectueuse des libertés individuelles et des droits de l'enfant.

POUR TOUTES CES RAISONS, NOUS DEMANDONS LE RETRAIT DE BASE ELEVES ET LA DESTRUCTION DE TOUTES LES DONNEES ENREGISTREES.

Annexe 2. Variables des fichiers des apprenants de la Depp potentiellement d'intérêt pour les enquêtes Génération

Variable	Contenu	Observations
an_sco	Année scolaire	
source	Source de l'information	
etab	Etablissement de scolarisation	UAI
Caractéristiques individuelles		
ine_genere	Qualité de l'INE	Pour repérer les INE fictifs
type_doublon	Type de doublon	Information qualité sur les INE
flag_doublon	Sélection de l'enregistrement	Information qualité mobilisée pour l'identification des « viviers » construits pour récupérer les informations aux examens
sexe	Sexe	
date_nais	Date de naissance	
annee_nais	Année de naissance	
dept_nais	Département de naissance	
natio	Nationalité	Non détaillé dans SIFA
com_resid	Commune de résidence	Retenir la commune plutôt que le département permet d'envisager le recours aux zonages communaux
PCS PCS2	PCS du responsable 1 PCS du responsable 2	Non utilisable pour SIFA (une seule PCS demandée, sans repérage de la personne de référence et avec près de 50 % de non-réponse). Le redressement consiste à placer en premier responsable le père, la PCS du père étant privilégiée pour les statistiques sur l'origine sociale des élèves par la Depp.
PCSr PCS2r	PCS redressée du responsable 1 PCS redressée du responsable 2	
lien1, lien1r, lien2, lien2r	Lien du responsable 1 et 2, brut et redressé avec l'élève	
Scolarité de l'année N/N+1		
geostat	Situation géographique de l'établissement	Code permettant d'identifier le département et l'académie
secteur	Secteur d'enseignement	Information de la BCE, permettant d'identifier le secteur public et secteur privé
contrat	Contrat d'établissement	Information de la BCE, permettant d'isoler le secteur privé sous contrat
type_etab3	Type d'établissement	Regroupement du code « nature » de la BCE.
bourse	Boursier (oui /non)	Non présent dans SIFA
statut	Statut de l'apprenant	Scolaire ou apprenti
mefst11	Formation suivie	Nomenclature propre aux fichiers Scolarité
fortrams7	Formation suivie	Nomenclature transverse à Scolarité et SIFA
cursus	Durée de la formation et année dans la formation	Exemple : « 21 » = en 1ère année d'une formation de 2 ans
dipl_spec	Diplôme préparé	Nomenclature transverse à Scolarité et SIFA
motif_sortie	Motif de sortie	Pour les sorties en cours d'année. Absente de SIFA.
date_sortie	Date de sortie	Pour les sorties en cours d'année. Absente de SIFA.
Résultats aux examens de la session N+1 (2 examens possibles)		
exam_dipl exam_dipl2	Code diplôme de l'examen passé	
pres_dipl pres_dipl2	Présence à l'examen	
result_dipl	Décision du jury	

result_dipl2		
statut_c statut_c2	Statut du candidat	

Annexe 3. Variables de SISE potentiellement d'intérêt pour les enquêtes Génération

SISE-Inscriptions

Variable	Contenu	Observations
Caractéristiques de l'établissement fréquenté		
ETABLI	Numéro d'identification de l'établissement dans la BCE	
UNIV	Indicatrice permettant d'isoler les universités	Certains établissements de SISE-Université ne sont pas des universités (ex :: IEP, INALCO, Dauphine)
COMETA	Code Insee de la commune d'implantation de l'établissement	Code Insee. Information issue de la BCE
COMINS	Code Insee de la commune d'implantation de la composante	Code Insee. Information issue de la BCE
Caractéristiques de l'étudiant		
IDETU	Numéro INE	
JONAIS	Jour de naissance	Non diffusée en dehors du Sies Absent de SISE-Ingénieurs
MONAIS	Mois de naissance	Absent de SISE-Ingénieurs
ANNAIS	Année de naissance	Si l'âge est inférieur à 14 ans ou supérieur à 99 ans, l'année de naissance est redressée en fonction du niveau et du type de formation suivie. Une variable AGE en âge atteint dans l'année de référence du fichier est aussi proposée.
SEXE	Sexe	En cas de non-réponse, imputation aléatoire
NATION	Nationalité de l'étudiant	La variable détaillée n'est pas diffusée en dehors du Sies. Les fichiers proposent aussi une variable agrégée par continent, avec distinction de l'UE au sein de l'Europe (NATRG), ainsi qu'une variable distinguant plus globalement français et étrangers (FR_ETR)
COMETU	Code Insee de la commune de la résidence de l'étudiant	Les établissements peuvent fournir le libellé de la commune ou le code Insee ou les deux. Le code postal est aussi demandé. Si le code commune n'est pas renseigné, il est codé à partir du libellé et du code postal. D'après le guide, « il est important d'utiliser cette variable avec précautions dans la mesure où parfois le code postal est parfois utilisé à la place du code commune ». Absent de SISE-Management et SISE-26bis
LCOMETU	Libellé de la commune de résidence de l'étudiant	
CP_ETU	Code postal de l'étudiant	
PCSPAR	PCS du premier parent	Pas de variable pour isoler père et mère.
PCSPAR2	PCS du deuxième parent	Pas de variable pour isoler père et mère.
COMREF	Code Insee de la commune des parents de l'étudiant	La documentation parle tantôt d'adresse des parents, tantôt d'adresse de référence. Il s'agit de l'adresse permanente, à laquelle l'étudiant peut être contacté à tout moment, à laquelle on peut par exemple lui expédier du courrier pendant les
LCOMREF	Libellé de l'adresse des parents	
CP_PAR	Code postal des parents	

		périodes de vacances scolaires ou après la fin de l'année universitaire. Même remarque sur le codage que pour la commune de l'étudiant Absent de SISE-Management et SISE-26bis
Scolarité antérieure de l'étudiant		
BAC	Série du bac ou équivalence	Des redressements sont effectués sur le nom du bac en fonction de l'année d'obtention pour se caler sur les dénominations en vigueur alors
BAC_RGPR	Série du bac regroupé	7 modalités (littéraires, économiques, scientifiques, technologiques STT, autres technologiques, professionnels, dispenses)
ANBAC	Année d'obtention du baccalauréat	Renseigné pour le bac, pas pour les équivalences
DEPBAC	Département d'obtention du bac	Département de l'établissement de scolarisation lors du passage du bac ou du centre d'examen pour les non scolarisés Absent de SISE-Ingénieurs
ANETAB	Année d'entrée dans l'établissement	Collecté uniquement sur SISE-Ingénieurs, SISE-Management et SISE-26bis
ANINSC	Année de première inscription dans le système universitaire français	Absent de SISE-Ingénieurs, SISE-Management et SISE-26bis
ANSUP	Année d'entrée dans l'enseignement supérieur	Absent de SISE-Management et SISE-26bis
SITUPRE	Situation de l'étudiant l'année précédente	Indique plutôt le type d'établissements de l'année précédente pour les scolarisés et la non-scolarisation
TYPREPA	Classe préparatoire fréquentée par l'étudiant	Concerne uniquement les nouveaux entrants ayant fréquenté l'année précédente une classe préparatoire aux grandes écoles (CPGE). Dans SISE-Université, n'est renseigné que pour les formations d'ingénieur.
Formation en cours de l'étudiant (rappel : en cas d'inscriptions multiples, une observation par inscription)		
DIPLOM	Code du diplôme d'inscription	La codification couvre les diplômes nationaux et les diplômes d'établissement (voir guide de l'utilisateur)
TYP_DIPL	Type de diplôme	Nomenclature fine en deux positions du diplôme préparé. Les variables CYCLE et CURSUS_LMD peuvent se déduire de cette variable.
SECTDIS	Secteur disciplinaire du diplôme d'inscription (52 modalités)	52 secteurs, codés à partir de DIPLOM
DISCIPLI	Discipline associée au diplôme d'inscription (16 modalités)	Regroupement en 16 modalités des 52 secteurs disciplinaires
GROUPE	Regroupement disciplinaire (8 modalités)	Regroupement des disciplines. Cette variable construite permet aussi de distinguer les formations secondaires et les formations tertiaires pour les IUT (selon une méthode présentée dans le guide utilisateur)
CURSUS_LMD	Cursus d'appartenance dans le schéma LMD	3 modalités : L, M, D
CYCLE	Cycle d'appartenance dans le diplôme préparé	Exemple pour la licence : 1er cycle pour les deux premières années, 2e cycle pour la troisième.

		Exemple pour le master : 2e cycle pour la première année, 3e cycle pour la seconde.
VOIE	Voie du diplôme pour les diplômes LMD	3 modalités : générale, professionnelle, recherche
NIVEAU_D	Niveau théorique atteint au début de formation du diplôme d'inscription	En nombre d'année après le bac. Variable construite par le Sies
NIVEAU_F	Niveau théorique atteint en fin de formation du diplôme d'inscription	En nombre d'année après le bac. Variable construite par le Sies
DEGETU	Degré d'étude théorique atteint pendant l'année d'inscription en cours	En nombre d'année après le bac. Variable construite par le Sies
NATURE	Nature juridique du diplôme préparé	Permet, par exemple, de distinguer les diplômes homologués nationalement des diplômes d'université, les diplômes des Ecoles, les préparations à des concours administratifs, etc.
Information complémentaire sur la scolarité de l'étudiant		
AMENA	Indicateur de l'existence ou non d'un parcours aménagé	Cette variable permet d'isoler des cas qui ne se conforme pas au parcours théorique, sans pour autant qu'il n'y ait redoublement. Par exemple, licence en 4 ans, master en 3 ans, année de césure. Absent de SISE-Management et SISE-26bis
CONV	Indicateur signalant si les enseignements sont dispensés dans le cadre d'une convention	Permet d'isoler les inscriptions relevant d'une convention entre une université et un institut catholique. Depuis 2015, repère aussi les inscriptions liées à une convention entre une classe préparatoire aux grandes écoles et un EPSCP Absent de SISE-Ingénieur, SISE Management et SISE-26bis
CURPAR	Indicateur signalant l'existence d'un cursus parallèle dans un autre établissement supérieur	Par exemple, en cas de cursus dans une Ecole et une université, dans une classe préparatoire et une université (avec ou sans convention), etc. Absent de SISE-Management et de SISE-26bis
EXOINS	Type d'exonération des frais d'inscription (s'il y a lieu)	Absent de SISE-Management et de SISE-26bis
ECHANG	Indicateur signalant les inscriptions dans le cadre d'un programme d'échange international	Absent de SISE-Management et de SISE-26bis
INSPR	Indicateur de repérage des inscriptions principales	L'inscription principale est définie par le Sies selon différents critères détaillés dans le guide utilisateur

SISE-Résultats

Variable	Contenu	Observations
Rappel : en cas d'inscriptions multiples, une observation par inscription		
IDETU IDETU1	Identifiant INE	Certaines universités ayant modifié le numéro INE de leurs étudiants, IDETU propose le dernier INE disponible et IDETU1, l'INE contenu dans le fichier de SISE Inscriptions de l'année correspondante. Le Sies dispose en effet du numéro d'inscription de l'étudiant dans l'établissement, ce qui lui permet de repérer les changements d'INE.
NONAPPAR	Résultat de l'appariement entre SISE Résultats et SISE Inscriptions	Permet par exemple de repérer notamment les inscrits ayant un diplôme attendu (« appariement total »), ceux ne correspondant pas à celui attendu (« appariement partiel »), etc.
COMPOSR	Unité de rattachement	
DIPLOMR	Diplôme SISE	Diplôme obtenu (variable de référence pour exploiter SISE Résultats, notamment pour repérer les réussites et les échecs)
ETATINS	État de l'inscription	Permet de distinguer les différents types de situation au regard de l'inscription et de l'appartenance au champ SISE
RESDIP	Obtention du diplôme	
DIPINT	Code diplôme intermédiaire	Vise notamment les DEUG et maîtrises que les universités peuvent délivrer en niveau intermédiaire depuis la mise en place du LMD. Depuis 2013, la remontée des DEUG intermédiaires n'est cependant plus obligatoire.
RESINT	Indicatrice de réussite du diplôme	1=diplômé. 0 intègre les échecs, la non-présence, les réorientations. La variable reste à blanc quand elle n'a pas été renseignée.
NIVEAUR	Niveau dans le diplôme	Voir Guide.
PREXAM	Présence à l'examen	Demandé à toutes les universités mais pas renseigné par tous les établissements dans SISE. Non exploitée par le Sies

Annexe 4. Les systèmes d'information sur la formation des demandeurs d'emploi vus par la Cour des comptes

Extraits du rapport de la Cour des comptes de mai 2018 sur formation des demandeurs d'emploi. https://www.ccomptes.fr/sites/default/files/2018-07/20180704-formation-demandeurs-d-emploi.pdf

Des systèmes d'information inadaptés et trop peu interconnectés
[…]

1 -Des systèmes qui peinent à s'interconnecter

Certains financeurs, notamment les régions, se sont dotés de systèmes d'information exclusivement centrés sur la passation et l'exécution des marchés publics : ils manquent de ce fait d'outils et de données pour le pilotage, le suivi et l'évaluation, faute de données suffisamment individualisées. En particulier, l'information disponible sur l'offre de formation n'est pas fiable.

En outre, ces systèmes d'information qui ont été conçus de manière éclatée peinent à s'interfacer : les organismes de formation doivent saisir les mêmes informations à plusieurs reprises dans des applicatifs différents.

Par ailleurs, la difficulté pour nombre de régions de relier directement leur système d'information à celui qui a été mis en place par la Caisse des dépôts et consignations pour la gestion du compte personnel de formation (CPF) ne facilite pas le suivi en temps réel des dossiers correspondants. […]
Le cas de l'Île-de-France illustre les nombreux systèmes d'information de la formation professionnelle impliquant des échanges entre Pôle emploi, la région, le CARIF-OREF, l'État, la Caisse des dépôts et consignations, les organismes de formation, ainsi que l'agence de services et de paiement (ASP), qui est l'organisme payeur principal des rémunérations versées aux demandeurs d'emploi en formation, lorsqu'ils ne sont pas indemnisés par Pôle emploi. »

Comme le schéma suivant le montre, les mêmes informations sont parfois saisies à plusieurs reprises par les organismes de formation sur des applications différentes. À titre d'exemple, les organismes de formation prestataires de la région Île-de-France doivent saisir les données relatives aux stages sur les outils suivants :
- leurs propres outils internes (outils de gestion retraçant le déroulement des formations et outils comptable et financier) ;
- la base Dokelio tenue par le CARIF-OREF pour rendre compte de l'état de l'offre de formation en temps réel (saisie des dates prévisionnelles de stages et des places encore vacantes) ;
- le portail « Kairos » de Pôle emploi, prescripteur de formation ;
- l'outil « Safir » du conseil régional d'Île-de-France, conçu pour la passation et l'exécution des marchés publics (saisie du détail de l'offre de formation et des heures de formation à financer) ;
- l'application « Remunet » de l'organisme payeur (ASP), qui n'est pas interfacée avec l'outil Safir de la région (saisie des heures réalisées par chaque stagiaire).
Ce schéma d'organisation comporte plusieurs zones de risques :
- l'évaluation des politiques de formation professionnelle est impossible dans ce cadre : les indicateurs d'efficacité et de performance ne peuvent reposer sur des données homogènes et centralisées ;
- le suivi et l'accompagnement des stagiaires depuis le début du processus de formation jusqu'à leur embauche est malaisé, car les données sont dispersées et peu fiables ;
- des erreurs de saisies peuvent intervenir avec un effet direct sur la qualité des échanges entre les différents intervenants : certains organismes de formation ont dû supporter des dépenses supplémentaires en raison des multiples saisies et du reporting particulier nécessaires pour le suivi du Plan 500 000 ;
- une information exacte sur les places disponibles dans les sessions de formation n'est pas toujours accessible, ni sur les sites Internet pour les demandeurs d'emploi, ni sur l'outil interne de Pôle emploi (AUDE) pour les conseillers de cet opérateur ;

- en dépit des dispositions légales, les financeurs des formations ne sont pas systématiquement informés en temps réel sur les entrées et les sorties des sessions ; en outre, Pôle emploi ne peut pas mettre à jour automatiquement la liste des demandeurs d'emploi sur la base des entrées et des sorties de formation.

La région Île-de-France a indiqué travailler actuellement à une simplification des interconnexions avec Pôle emploi, en lien avec les organismes de formation. En Auvergne-Rhône-Alpes, une difficulté supplémentaire résulte de la fusion des régions qui a imposé d'harmoniser les systèmes d'information sur l'ensemble du nouveau territoire régional. Les problèmes d'interface ne sont pas encore tous résolus […]

Dans ce contexte, Pôle emploi mène depuis quelques années des actions de sensibilisation vis-à-vis de ses partenaires, dans le but de mettre en place une interconnexion des systèmes. Un portail dénommé « Kairos » est en place pour faciliter les échanges d'information avec les prestataires sur le parcours des demandeurs d'emploi, de leur sélection à leur entrée en formation. Toutefois, cet outil n'est pas utilisé systématiquement par les organismes de formation (environ 7 000 y recourent actuellement).

Le tableau ci-dessous indique par région le taux de dématérialisation des entrées en stage (AES) fournies par Kairos, entre janvier et juin 2017. Il convient toutefois de préciser que la simplification introduite par le recours à Kairos a permis d'augmenter ce taux : dans la région Grand Est, il est ainsi passé de 5 % en juin 2017 à 39,5 % en octobre 2017. » [max =69% en Bourgogne Franche Comté]
Les organismes de formation peuvent utiliser le portail Kairos pour transmettre à Pôle emploi les attestations d'inscription (AIS) et d'entrée en stage (AES). En cas de non-utilisation de Kairos, un processus complexe s'applique […]

2-Un système de recueil de données qui reste à construire

Pour tenter de dépasser ces difficultés et mettre en place un système de recueil de données permettant de mieux évaluer les résultats de la formation professionnelle, la DGEFP a lancé en 2016 le projet Agora. Dans un premier temps, ce projet vise à normaliser et homogénéiser les données à travers les systèmes d'information des différents acteurs, notamment via le Nir. Dans un deuxième temps, l'ensemble de ces données sera déployé au sein d'un entrepôt unique, alimenté par les mêmes acteurs et accessible à tous.

A terme, Agora devrait ainsi permettre de réduire le nombre de saisies de données par les acteurs, de simplifier ou alléger les processus de collecte de données par des bases issues de référentiels gérés par des acteurs qui en ont la maîtrise et la gouvernance, et de disposer de statistiques plus complètes et plus fiables, y compris sur l'ensemble des parcours de formation.

Toutefois, le projet Agora repose sur l'hypothèse que chaque acteur de la formation professionnelle dispose de données individualisées sur les parcours de formation des demandeurs d'emploi. Or, tel n'est pas toujours le cas, ainsi que l'a montré l'impossibilité pour les régions sollicitées par la Cour de répondre à certaines demandes de données visant à associer les caractéristiques des stages à celles des stagiaires. Les systèmes d'information permettent en effet de s'assurer par exemple que le nombre de stages, les volumes horaires, et les montants financiers prévus ont bien été respectés par les organismes prestataires, mais pas de produire des données relatives à l'analyse des parcours individuels des stagiaires.

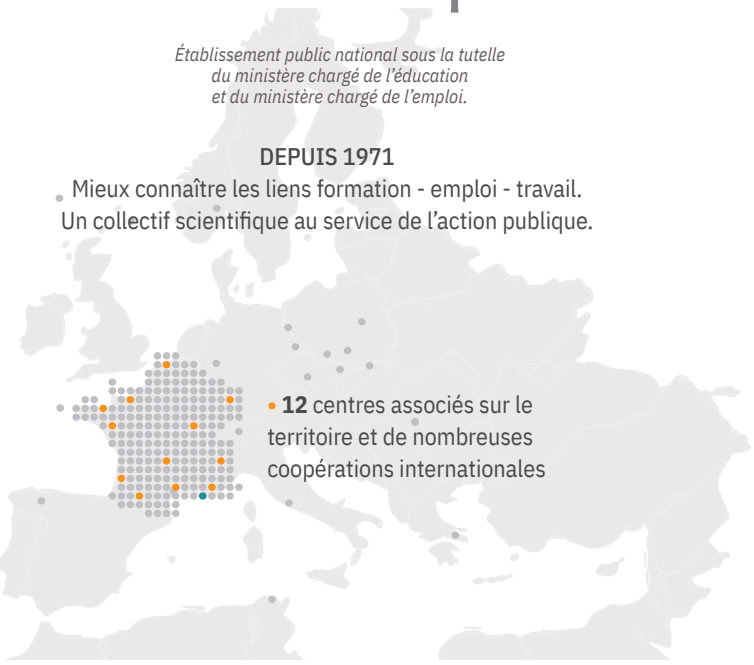
De plus, il conviendrait que ce projet associe l'ensemble des intervenants avec des experts en systèmes d'information au sein d'un groupe de travail intégré. Or, à titre d'exemple, le directeur des systèmes d'information de la région Île-de-France n'est pas membre du groupe de travail du projet Agora, bien que le système d'information de cette région soit représentatif des difficultés auxquelles ce projet est susceptible de se heurte.

Céreq

Établissement public national sous la tutelle du ministère chargé de l'éducation et du ministère chargé de l'emploi.

DEPUIS 1971

Mieux connaître les liens formation - emploi - travail.
Un collectif scientifique au service de l'action publique.



• **12** centres associés sur le territoire et de nombreuses coopérations internationales

+ d'infos
et tous les travaux

À explorer
www.cereq.fr

+ de 600 publications
Accessibles librement