Conseil national
de l'information statistique

Montrouge, le 5 août 2025 – n° 121/H030

RENCONTRE SUR LES APPARIEMENTS DE LA STATISTIQUE PUBLIQUE

_____

Bercy, 28 mai 2025

ACTES DE LA RENCONTRE
« LES APPARIEMENTS DE LA STATISTIQUE PUBLIQUE »

28 mai 2025

Président : Bertrand DU MARAIS, Président du Cnis

RAPPEL DE L'ORDRE DU JOUR

Liste des participants

NOM	PRÉNOM	ORGANISME
ARNETON	Mélissa	Institut national d'études démographiques (INED)
AVOUAC	Romain	Institut national de la statistique et des études économiques (INSEE)
BALAVOINE	Angélique	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
BALLET	Bertrand	Service de la statistique et de la prospective (SSP)
BARLET	Muriel	Institut national de la statistique et des études économiques (INSEE)
BEAUFILS	Hugo	Insee Bretagne
BECHICHI	Nagui	Paris School of Economics – Université Paris 1
BENABDALLAH	Saïd	Rectorat de Versailles
BERNILLON	Pascale	Santé Publique France
BERTIN	Sévrine	
BESSONE	Anne-Juliette	Direction de l'animation de la recherche, des études et des statistiques (DARES)
BIAU	Olivier	Institut national de la statistique et des études économiques (INSEE)
BILLAUT	Anne	Service des données et des études statistiques (SDES)
BLANC	Sylvie	Secrétariat général du CNIS
BOUR	Romain	Direction de l'évaluation, de la prospective et de la performance (DEPP)
BRAJON	Delphine	L'Institut Paris Région
BREUIL	Pascale	Institut national de la statistique et des études économiques (INSEE)
BRIAND	Antonin	Service statistique ministériel de la sécurité intérieure (SSMSI)
BRIERE	Luc	Autorité de la statistique publique (ASP)
BRIZARD	Agnès	Département des statistiques, des études et de la documentation (DSED), Ministère de l'Intérieur (SSM Immigration)
BROCHARD	Anne-Sophie	Observatoire régional de la santé (ORS) – Pays de la Loire
CAILLET	Julie	Mission interministérielle pour la protection des femmes contre les violences et la lutte contre la traite des êtres humains (Miprof)
CANCEL	Sébastien	Secrétariat général du CNIS
CARAY	Jérôme	Service de la statistique et de la prospective (SSP)
CARON	Nathalie	Insee Bretagne
CARRASCO	Valérie	Service statistique ministériel de la sécurité intérieure (SSMSI)
CASTAGNÉ	Marie	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
CHALEIX	Mylène	Institut national de la statistique et des études économiques (INSEE)
CHARRANCE	Géraldine	Institut national d'études démographiques (INED)
CHESNE	Lou	Électricité de France (EDF)
CHEVALIER	Pascal	Ministère de la Justice – Sous-direction de la statistique et des études
CLANCHÉ	François	Institut national d'études démographiques (INED)

CLAUDE	Nicolas	Ministère de l'Intérieur
CLOAREC	Nathalie	Direction de l'animation de la recherche, des études et des statistiques (DARES)
COCHET	Paul	Institut national d'études démographiques (INED)
COLIN	Christel	Institut national de la statistique et des études économiques (INSEE)
CORRE	Tifenn	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (Inrae)
COUHIN	Julie	Caisse nationale d'assurance vieillesse (CNAV)
CRAVATTE	Céline	Secrétariat général du CNIS
CRESPIN	Aurélien	Agence d'urbanisme de Bordeaux
DARRIAU	Valérie	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
DAUVIN	Magali	Observatoire français des conjonctures économiques (OFCE)
DE PERETTI	Gaël	Direction générale de l'administration et de la fonction publique (DGAFP)
DE PIERO	Lorane	Service des données et des études statistiques (SDES)
DEDIEU	Marie-Sophie	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE)
DEMOLY	Elvire	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
DESCLODURE	Julien	Direction générale des collectivités locales – Département des études et des statistiques locales
DEUIL	Philippe	Insee Bretagne
DIOGO	François	Région Grand Est
DOMENS	Jérôme	Insee Provence-Alpes-Côte d'Azur
DU MARAIS	Bertrand	Conseil national de l'information statistique (CNIS)
DUC	Cindy	Insee Réunion-Mayotte
DUÉE	Michel	Institut national de la statistique et des études économiques (INSEE)
DUMOULIN	Céline	Université de Versailles Saint-Quentin-en-Yvelines
DUPONT	Françoise	Particulier
DURAN	Patrice	École Normale Supérieure de Cachan
DURRUTY	Bruno	Confédération générale du travail (CGT)
DUSSART	Josy	Institut national de la statistique et des études économiques (INSEE)
EL KHOURY	Carole	Caisse nationale d'assurance vieillesse (CNAV)
EVAIN	Manon	
EVEN	Karl	Sous-direction des systèmes d'information et des études statistiques (SIES)
FABRE	Marianne	Institut national de la statistique et des études économiques (INSEE)
FARGES	Audrey	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
FERRARI	Giulia	Institut national d'études démographiques (INED)
FESSEAU	Depp	Direction de l'évaluation, de la prospective et de la performance (DEPP)

FRECHOU	Hélène	Insee Info Service
GACHARD	Mathilde	DDT du Bas-Rhin
GADOUCHE	Kamel	Centre d'accès sécurise distant aux données (CASD)
GAINI	Mathilde	Direction de l'animation de la recherche, des études et des statistiques (DARES)
GARCIA	Kevin	Direction de l'animation de la recherche, des études et des statistiques (DARES)
GAUBERT	Émilie	Centre d'études et de recherches sur les qualifications (CEREQ)
GAUVIN	Charlotte	Direction générale de l'enseignement et de la recherche
GÉDOR	Elsa	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE)
GEEROLF	François	Observatoire français des conjonctures économiques (OFCE)
GILLES	Séverine	Institut national de la statistique et des études économiques (INSEE)
GODARD	Mathilde	Université Paris Dauphine
GOETZ	Emmanuel	Réseau des Agences régionales de l'énergie et de l'environnement (RARE)
GONZALEZ-DEMICHEL	Christine	Service statistique ministériel de la sécurité intérieure (SSMSI)
GOURDOL	Albane	Institut national d'études démographiques (INED)
GUÉRIN	Vivien	Institut national de la statistique et des études économiques (INSEE)
GUÉROUT	Anthony	Association des maires de France et des présidents d'intercommunalité (AMF)
GUILLAUMAT-TAILLIET	François	Secrétariat général du CNIS
HAAG	Olivier	Institut national de la statistique et des études économiques (INSEE)
HALLÉPÉE	Sébastien	Institut national de la statistique et des études économiques (INSEE)
HAMÉON	Vincent	Observatoire Régional de la Santé du Centre-Val de Loire
HEYDEMANN	Pascale	Institut français du cheval et de l'équitation (IFCE)
ISNARD	Michel	Particulier
JALUZOT	Laurence	
JAUNEAU	Yves	Institut national de la statistique et des études économiques (INSEE)
JUGNOT	Stéphane	Centre d'études et de recherches sur les qualifications (CEREQ)
JULLIOT	Mylène	Caisse nationale d'assurance vieillesse (CNAV)
KOUBI	Malik	Direction de l'animation de la recherche, des études et des statistiques (DARES)
LAGARENNE	Christine	Secrétariat général du CNIS
LAPINTE	Aude	Direction de l'animation de la recherche, des études et des statistiques (DARES)
LAVERGNE	Aurélien	Institut national de la statistique et des études économiques (INSEE)
LE GRAND	Hervé	Service de la statistique et de la prospective (SSP)
LEFEBVRE	Olivier	Institut national de la statistique et des études économiques (INSEE)
LEGLISE	Delphine	Insee Hauts-de-France
LEQUIEN	Matthieu	Institut national de la statistique et des études économiques (INSEE)

LEQUIEN	Laurent	Service des données et des études statistiques (SDES)
LOTH	André	Particulier
MAKDESSI	Yara	Ministère de la Justice – Sous-direction de la statistique et des études
MALGOUYRES	Clément	Centre de recherche en économie et statistique (CREST)
MARACINEANU	Roxana	Miprof-Observatoire national des violences faites aux femmes
MARMION	Violette	Direction de l'évaluation, de la prospective et de la performance (DEPP)
MARQUIER	Rémy	Centre d'accès sécurise distant aux données (CASD)
MAURICE	Léopold	Direction de l'animation de la recherche, des études et des statistiques (DARES)
MEURS	Dominique	Economix université Paris X
MIKOL	Fanny	Institut national de la statistique et des études économiques (INSEE)
MINODIER	Frédéric	Institut national de la statistique et des études économiques (INSEE)
MINODIER	Christelle	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
MISSÈGUE	Nathalie	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
MONSERAND	Alma	Agence de l'environnement et de la maîtrise de l'énergie (ADEME)
MONZIOLS	Martin	Institut national de la statistique et des études économiques (INSEE)
MORDANT	Guillaume	Département des statistiques, des études et de la documentation (DSED), Ministère de l'Intérieur, SSM Immigration
MOREL	Claire	CM Conseil
MOREZ	Claire	
MOUNIER	Lise	Particulier
MUTRICY	Philippe	Bpifrance
OKHAM	Elmostafa	Insee Bretagne
OURLIAC	Jean-Paul	SGP – conseil scientifique
PARRIAUD	Jean-François	Insee Auvergne – Rhône-Alpes
PERBEN	Margot	Sous-direction des systèmes d'information et des études statistiques (SIES)
PHAM	Trong-hien	Institut national de la statistique et des études économiques (INSEE)
PICARD	Tristan	Insee Bretagne
PICARD	Hugues	Particulier
PIFFETEAU	Hervé	
PIGUET	Virginie	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE)
PINELLI	Florence	Agence nationale de la cohésion des territoires (ANCT)
PLACE	Dominique	Observatoire économique de la défense (OED)
PLANCHON	Julie	Centre d'études et de recherches économiques sur l'énergie (CEREN)
POISSON	Claire	FranceAgrimer
PROKOVAS	Nicolas	Confédération générale du travail (CGT)

PROST	Corinne	Institut national de la statistique et des études économiques (INSEE)
QUANTIN	Catherine	Centre Hospitalier Universitaire
RAZAFINDRANOVONA	Tiaray	Service statistique ministériel de la sécurité intérieure (SSMSI)
RAZAFINDRATSIMA	Nicolas	Ministère de la Justice – Sous-direction de la statistique et des études
ROGER	Muriel	Université Paris 1 Panthéon-Sorbonne
SALEMBIER	Laurianne	Service statistique ministériel de la sécurité intérieure (SSMSI)
SAOUDI	Abdessattar	Santé publique France
SARRON	Clotilde	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
SCHUHL	Pierrette	Sous-direction des systèmes d'information et des études statistiques (SIES)
SCIBERRAS	Jean-Christophe	Newbridges et Cnis
SÉDILLOT	Béatrice	Service des données et des études statistiques (SDES)
SEKOURI	Mohamed	Secrétariat général du CNIS
SILBERMAN	Roxane	Centre national de la recherche scientifique (CNRS)
SIMONNET	Florian	Agence nationale de la cohésion des territoires (ANCT)
SONNETTE CHICH	Catherine	Service des données et des études statistiques (SDES)
STÉRIN	Anne-Laure	Université Paris 1 Panthéon-Sorbonne
SUESSER	Jan Robert	Ligue des droits de l'Homme
SUJOBERT	Bernard	Confédération générale du travail (CGT)
TARAYOUN	Tedjani	Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
TARTESSE	Sylvie	Association pour l'emploi des cadres (APEC)
TAVERNIER	Jean-Luc	Institut national de la statistique et des études économiques (INSEE)
TCHA	Stéphanie	Service statistique ministériel de la sécurité intérieure (SSMSI)
THÉODOSE	Teddy	Université Paris 13
TORTOSA	Thomas	Institut national de la statistique et des études économiques (INSEE)
TREGARO	Yves	Conseil général de l'alimentation, de l'agriculture et des espaces ruraux
TREYENS	Pierre-Éric	Insee Bretagne
VALICON	Astrid	Caisse nationale d'assurance vieillesse (CNAV)
VALLET	Louis-André	Centre national de la recherche scientifique (CNRS)
VESSILLIER	Delphine	Fédération française du bâtiment
VIAUD	Marie	Institut national de la santé et de la recherche médicale (INSERM)
WILLAUME	Benoît	Agence nationale pour la formation professionnelle des adultes (AFPA)
ZOBEL	Thibaud	Santé publique France

La séance est ouverte à 9 heures 05, sous la présidence de Bertrand du Marais.

INTRODUCTION

Jean-Luc TAVERNIER, directeur général de l'Insee, souhaite la bienvenue aux participants à rencontre sur les appariements de la statistique publique. Il s'agit de la deuxième rencontre consacrée à ce sujet, la première ayant eu lieu en janvier 2022.

L'importance de ce sujet pour le Cnis est mise en évidence par le développement croissant du recours aux appariements. En effet, la demande d'informations plus détaillées sur les politiques publiques est constante, comme l'illustre pour ne prendre que l'exemple le plus récent une réponse la veille à un relevé d'observations provisoires de la Cour des comptes concernant les politiques de soutien à la transition environnementale pour les ménages. Cette exigence d'informations plus granulaires sur les phénomènes économiques et sociaux se heurte cependant à la nécessité de ne pas surcharger les ménages et les entreprises de sollicitations.

Le contexte européen actuel renforce cette problématique. Certains pays, notamment ceux ayant décidé de coupes budgétaires importantes pour financer des efforts de défense, souhaitent réduire le recours aux enquêtes. D'autres font face à une baisse significative des taux de réponse aux enquêtes auprès des ménages qu'ils ne parviennent pas à enrayer. Cette situation pousse à un recours accru aux données administratives et aux appariements.

Les opérations d'appariement apparaissent ainsi comme une solution efficace pour produire davantage d'informations sans augmenter les moyens ni solliciter directement les personnes concernées. Cependant, cette approche est porteuse d'enjeux cruciaux :

- Elle nécessite une expertise statistique et une compétence pointue en systèmes d'information.
- Elle implique de garantir la confidentialité des données pour préserver la confiance du public.
- Elle exige d'assurer la transparence du traitement de données initialement non destinées à des fins statistiques ou à être combinées.

Ces enjeux avaient déjà été abordés lors de la rencontre de janvier 2022, qui avait mis en avant la nécessité d'un cadre juridique, d'outils améliorés, et d'un mandat social. Les traitements doivent présenter des garanties appropriées pour le respect des droits et libertés des personnes concernées, conformément au Règlement général sur la protection des données (RGPD), et démontrer leur nécessité vis-à-vis de l'intérêt général, avec des moyens proportionnés aux objectifs.

Depuis cette première rencontre, des avancées significatives ont été réalisées :

- Un groupe de concertation sous l'égide du Conseil national de l'information statistique (Cnis) a précisé les outils envisagés et débattu des questions éthiques.
- Des appariements ont été développés pour enrichir les enquêtes, ou se substituer à certaines questions, vérifier la fiabilité des enquêtes, et créer des parcours (entrée/sortie de la pauvreté, marché du travail, trajectoires des jeunes de l'école à l'insertion sur le marché du travail, etc.).
- Des phénomènes localisés ou ciblés sur des petites populations ont été étudiés, et des politiques publiques évaluées.

Le Répertoire statistique des individus et des logements (Résil) a été créé dans un cadre juridique validé par la Commission nationale de l'informatique et des libertés (CNIL) et le Conseil d'État, notamment pour pallier la disparition de la taxe d'habitation pour la plupart des logements.

Jean-Luc TAVERNIER conclut en soulignant l'importance de cette rencontre pour faire le point sur les avancées, réfléchir à la conformité des avancées avec le mandat défini en 2022, et envisager un cadre de référence formalisé pour l'ensemble des appariements de la statistique publique.

LA PLACE DES APPARIEMENTS DANS LA STATISTIQUE PUBLIQUE

Un document est projeté à l'ensemble des participants[1].

Christine LAGARENNE, secrétaire générale du Cnis, présente les deux intervenantes suivantes : Christel Colin, directrice des statistiques démographiques et sociales à l'Institut national de la statistique et des études économiques (Insee), et Corinne Prost, directrice de la méthodologie et de la coordination statistique et internationale à l'Insee.

Elles vont toutes deux approfondir le sujet en expliquant les outils statistiques et juridiques qui ont permis le développement des appariements dans la statistique publique, ainsi que les conditions de leur mise en œuvre.

Christel COLIN, directrice des statistiques démographiques et sociales de l'Insee, aborde la place des appariements dans les travaux du service statistique public, comprenant l'Insee et les services statistiques ministériels.

Les appariements consistent en la combinaison de données individuelles provenant de différentes sources pour fournir une information plus riche. Cette pratique consiste à rapprocher, pour une même personne, des données la concernant issues de différentes sources ou dossiers.

Depuis début 2024, une définition juridique des appariements existe dans le décret fondant Résil. Ce répertoire a vocation à faciliter les appariements de données administratives avec d'autres sources de données. L'article 1 de ce décret précise que ces appariements constituent des mises en relation, au sens de la loi Informatique et libertés, entre les données à caractère personnel enregistrées sur le répertoire statistique des individus et des logements et des sources de données statistiques tierces, donnant lieu à la création de nouveaux fichiers, lesquels constituent des traitements de données à caractère personnel au sens du RGPD.

Dans l'exemple du rapprochement entre deux fichiers, l'un contenant le diplôme le plus élevé par personne, l'autre la profession exercée, l'appariement de ces données permet d'analyser le lien entre profession et niveau de diplôme, en utilisant des traits d'identité comme le nom, le prénom et la date de naissance.

Le service statistique public réalise des appariements depuis les années 1960, avec des précurseurs tels que l'enquête Revenus fiscaux (devenue enquête Revenus fiscaux et sociaux, ERFS) de l'Insee, qui combine des données d'enquête avec des données fiscales et sociales. Un autre exemple historique est l'échantillon démographique permanent (EDP), créé par l'Insee en 1968, qui associe des données de recensement, d'état civil, et plus récemment des informations sur l'emploi et les revenus.

Dans les années 1970, divers panels ont été mis en place, notamment des panels d'élèves par la Direction de l'évaluation, de la prospective et de la performance (Depp) et de salariés par l'Insee, permettant de suivre des trajectoires scolaires et professionnelles. Ces appariements ont permis de créer de nouvelles sources plus riches, capables de répondre à de nouvelles questions, comme l'évaluation des politiques publiques.

Christel COLIN souligne l'intérêt des appariements comme mode de collecte, permettant de développer de nouvelles sources plus riches, permettant de répondre à de nouvelles questions, comme documenter les trajectoires scolaires, professionnelles, de vie, ou évaluer des politiques publiques en comparant les trajectoires de bénéficiaires et de non-bénéficiaires, en tirant parti des qualités de chaque source utilisée. Les enquêtes apportent généralement une richesse en termes de variables, tandis que les données administratives offrent souvent une exhaustivité sur leur champ de couverture, ce qui permet de caractériser des populations rares et d'étudier des phénomènes à un niveau géographique fin.

La mise en place du code statistique non signifiant (CSNS), issu de la loi pour une République numérique de 2016 facilite et fiabilise le développement des appariements, marquant ainsi une évolution significative dans les pratiques du service statistique public.

[1] Les diapositives présentées lors de la rencontre sont disponibles sur le site du Cnis : https://www.cnis.fr/evenements/rencontre-les-appariements/?category=1067

L'appariement de données nécessite des informations identifiantes suffisantes dans les fichiers à rapprocher. L'idéal est de disposer d'un identifiant pour chaque individu commun aux deux fichiers, comme le numéro d'inscription au répertoire national d'identification des personnes physiques (NIR), communément appelé numéro de sécurité sociale. Ce dernier permet de caractériser chaque individu de manière unique et sans ambiguïté, étant conservé dans un référentiel. Le numéro d'identification national des étudiants (INE) constitue un autre exemple d'identifiant efficace.

Cependant, toutes les bases de données ne contiennent pas le NIR et l'usage du NIR est strictement encadré pour des raisons de protection de la vie privée et des données personnelles. Avant 2016, tout appariement de fichiers basé sur le NIR à des fins statistiques nécessitait un décret en Conseil d'État après avis de la Cnil, processus long et complexe.

Pour faciliter ces appariements tout en garantissant la protection des données personnelles, la loi pour une République numérique de 2016 a introduit un nouveau dispositif. Désormais, les appariements peuvent se faire sans avis préalable de la Cnil ni décret en Conseil d'État, à condition d'utiliser un identifiant non signifiant. Ce CSNS, dérivé du NIR, conserve les propriétés techniques d'un identifiant tout en rendant impossible l'identification des personnes.

L'Insee a développé un service de calcul et de fourniture de ce CSNS, opérationnel dans sa version complète depuis octobre 2022. Ce service attribue à chaque individu qui se trouve dans un fichier administratif ou d'enquête un code, calculé pour chaque source, et dont le résultat est unique pour chaque individu, quelle que soit la nature de l'opération statistique originelle. Ce CSNS ne peut être transmis qu'à l'Insee et aux services statistiques ministériels (SSM), et uniquement pour des finalités de statistique publique. Néanmoins, les chercheurs peuvent accéder aux fichiers issus des appariements développés par l'Insee ou les SSM, notamment via le Centre d'accès sécurisé aux données (CASD).

Le calcul du CSNS s'effectue de deux manières. Pour les sources statistiques comportant le NIR, le code est calculé directement par hachage et cryptage du NIR avec une clé secrète. Pour les sources sans NIR, le calcul s'appuie sur les traits d'identité (nom, prénom, sexe, date et lieu de naissance) qui permettent de remonter au NIR, puis de le hacher et crypter pour obtenir le CSNS.

A partir du CSNS, qui est un nouvel identifiant pour chaque personne, il est ensuite possible et facile d'apparier les informations de fichiers différents.

Depuis fin 2022, neuf SSM et six services de l'Insee recourent au service de fourniture d'un CSNS. Son utilisation est encadrée par une convention de sous-traitance comprenant une charte d'usage avec des engagements spécifiques.

Le service CSNS fournit également une mesure de la qualité de l'identification des personnes, permettant aux utilisateurs d'évaluer la fiabilité de l'identification. Cette indication de qualité est notée de 1 (parfaitement fiable) à 7 (non fiable) pour chaque individu. Cette note qualité est calculée pour minimiser les risques de faux positifs, c'est-à-dire l'attribution d'un mauvais CSNS à une personne.

L'utilisation du CSNS a considérablement facilité les appariements. Les bilans 2023 et 2024 des travaux de l'Insee et des SSM font état respectivement de 34 et 51 appariements déclarés utilisant le CSNS. Ces appariements couvrent des usages variés, tant traditionnels que novateurs.

Parmi les exemples d'utilisation, on trouve :

- L'étude du non-recours au minimum vieillesse, montrant qu'une personne éligible sur deux n'y recourt pas.
- L'analyse de la vulnérabilité énergétique des ménages en termes de déplacement.
- L'étude de la multipropriété des logements, révélant qu'un tiers des propriétaires possède deux logements ou plus.

De nombreux appariements visent également à étudier les trajectoires d'insertion, professionnelle ou de fin de carrière, avec ou sans dimension d'évaluation des politiques publiques. Par exemple :

- L'analyse du passage entre les structures d'insertion par l'activité économique et les contrats aidés.
- L'étude des trajectoires professionnelles des sortants de l'enseignement supérieur.
- L'examen du recours aux minima sociaux en fin de carrière et au moment du passage à la retraite.

Christel COLIN souligne enfin le développement accru des appariements méthodologiques ou des tests pour envisager des évolutions de processus ou des allègements d'enquête. Cette tendance, facilitée par la simplification des appariements, ouvre de nouvelles perspectives d'innovation dans le domaine de la statistique publique.

Si des appariements méthodologiques continuent de se développer pour comprendre les écarts entre sources pour des concepts voisins, se développent également les appariements visant à alléger les questionnaires d'enquête, à remplacer certaines enquêtes ou à tester la faisabilité de tels allègements, ainsi que pour mesurer la qualité et la couverture de différentes sources de données.

Plusieurs exemples concrets illustrent cette tendance. La Direction de l'animation de la recherche, des études et des statistiques (Dares) étudie actuellement des scénarios d'évolution pour ses enquêtes auprès des personnes sortant de dispositifs d'insertion. Cette réflexion s'appuie sur des appariements entre les données sur les bénéficiaires de ces dispositifs et les déclarations sociales nominatives, permettant ainsi de suivre l'insertion sur le marché du travail.

Un autre cas significatif a été présenté à la commission Emploi, qualification et revenus du travail du Cnis. La décision a été prise de ne pas reconduire l'enquête Formation et qualification professionnelle (FQP), auparavant réalisée par l'Insee. En remplacement, deux approches sont envisagées : d'une part, l'utilisation de sources existantes pour couvrir une partie des besoins, et d'autre part, la mise en place d'appariements. Ces derniers, actuellement en développement, visent à remplacer les données auparavant collectées par voie d'enquête. Par exemple, pour permettre l'étude des mobilités et des trajectoires professionnelles, un enrichissement de l'enquête Emploi par des données administratives sur l'emploi et les salaires est en cours.

Ces exemples démontrent comment les appariements peuvent à la fois enrichir la connaissance et réduire la charge liée aux enquêtes directes, tant pour les enquêtés que pour les producteurs de statistiques. Néanmoins, il est important de souligner que les appariements ne peuvent pas se substituer à toutes les enquêtes, celles-ci restant nécessaires dans de nombreux cas.

Parallèlement à ces avancées techniques, des réflexions approfondies ont été menées ces dernières années sur les cadres méthodologique et déontologique de ces pratiques. La présentation suivante abordera ces aspects en détail.

Corinne PROST, directrice de la méthodologie et de la coordination statistique et internationale de l'Insee, poursuit l'exposé en se concentrant sur le cadre juridique et technique assurant la protection des données dans le contexte des appariements statistiques, ainsi que sur le renforcement du cadre déontologique.

Ces dernières années, le développement des appariements au sein du service statistique public a été facilité par plusieurs facteurs. Un partage de méthodes a été effectué, notamment au sein de la direction chargée de la méthodologie, avec la diffusion de documents techniques pour partager les connaissances avec les Services statistiques ministériels. Le CSNS a également joué un rôle important dans le développement des appariements de données personnelles, avec une restriction toutefois de champ liée au fait qu'il ne peut pas être utilisé pour les données du Système national des données de santé.

Le cadre juridique encadrant les appariements n'est pas spécifique à cette pratique, mais s'inscrit dans le cadre général de la production statistique. Il repose sur plusieurs textes fondamentaux :

- La loi de 1951 sur la statistique, qui définit notamment le secret statistique.
- La loi Informatique et libertés de 1978, définissant la protection des données personnelles.
- Le RGPD (Règlement général pour la protection des données) de 2018, qui renforce les droits des personnes concernées et les obligations en matière de transparence.

Dans ce cadre, le responsable de traitement doit justifier la finalité et la légitimité des traitements, garantir que les données traitées sont adaptées à ces finalités, et limiter la conservation des données à la durée strictement nécessaire.

Les appariements, considérés comme des traitements de données, doivent respecter ces principes. Un responsable de traitement doit être identifié pour chaque appariement, vérifier le respect des principes de nécessité et de minimisation, inscrire l'appariement au registre des traitements, et réaliser une étude d'impact si nécessaire.

Des mesures techniques et organisationnelles sont mises en place pour assurer la protection des données, notamment en sécurisant l'accès et en désignant des agents autorisés. Dans le cas du CSNS, des mesures supplémentaires sont appliquées, comme le chiffrement et l'isolement des données.

En complément du cadre juridique, un cadre déontologique s'applique, basé sur le Code de bonnes pratiques de la statistique européenne. Ce Code comprend 16 principes couvrant la qualité, la pertinence, l'exactitude, la cohérence et la clarté des productions statistiques. Deux principes en particulier encouragent l'utilisation des appariements : la limitation de la charge des déclarants et la recherche d'un bon rapport coût-efficacité.

Le code des bonnes pratiques a été renforcé par un mandat social spécifique aux appariements, s'inspirant des principes du RGPD tels que la nécessité, la proportionnalité et la transparence. Le service statistique public a déjà mis en place plusieurs initiatives de transparence, notamment une rencontre du Cnis en 2022, un billet de blog et la description des appariements dans les programmes de travail.

Cependant, les appariements présentent certaines limites. La qualité des données, en particulier l'identification des individus ou des entreprises, est cruciale pour un bon appariement. Certaines données administratives manquent de variables d'identification fiables, ce qui empêche leur utilisation. De plus, la pertinence des informations est essentielle. Par exemple, les données administratives sont considérées comme plus fiables pour mesurer le revenu des ménages que les enquêtes, mais certains concepts, comme le chômage au sens du Bureau international du Travail (BIT), ne peuvent être correctement mesurés que par des enquêtes.

Un autre défi majeur est la complexité croissante de l'analyse résultant de l'enrichissement des données. Bien que le coût de construction d'un appariement soit raisonnable, le coût d'analyse peut être élevé en raison de l'augmentation des dimensions et de la création de trajectoires complexes. **Corinne PROST** illustre ce point avec un graphique montrant des carrières salariales par secteur, âge et diplôme, soulignant la richesse, mais aussi la complexité des données obtenues.

Elle cite également l'exemple du panel Trajam de la Dares, qui a apparié différentes bases de données sur les dispositifs d'aide à l'insertion des jeunes sur le marché du travail. Ce projet, bien que fructueux, a été complexe à réaliser en raison de la multiplicité des bases de données et de la nécessité de comprendre la richesse du fichier final.

Ces expériences incitent à cibler les usages et à minimiser les appariements, en se concentrant sur des questions spécifiques plutôt que de chercher à créer un appariement trop large. Les projets utilisant le CSNS suivent cette approche, se limitant souvent à l'appariement de deux bases de données avec des objectifs bien définis.

En conclusion, bien que les appariements individuels soient une pratique ancienne bénéficiant d'avancées techniques et juridiques, ils nécessitent une réflexion approfondie sur leurs usages et une évolution de la concertation. Le cadre de référence, qui sera discuté lors de la table ronde à venir, vise à adapter cette concertation aux nouveaux usages des appariements.

Christine LAGARENNE remercie Christel Colin et Corinne Prost pour leurs présentations. La prochaine séquence portera sur des exemples concrets d'appariements de fichiers. Cette session mettra en lumière à la fois les enjeux méthodologiques rencontrés par les producteurs de données statistiques et les besoins de connaissance auxquels ces appariements répondent. Laurent Lequien, sous-directeur des statistiques des transports du service des données et études statistiques (Sdes) des ministères de l'Aménagement du territoire et de la Transition écologique, présentera des travaux sur les utilisateurs des véhicules routiers, qui seront discutés par Clément Malgouyres, chercheur au Centre de recherche en économie et statistique (CREST) et à l'Institut des politiques publiques (IPP).

QUELQUES EXEMPLES D'APPARIEMENTS ET LEURS ENJEUX CONCRETS

.1 Que savons-nous des utilisateurs de véhicules routiers ?

Un document est projeté à l'ensemble des participants.

Laurent LEQUIEN, sous-directeur des statistiques des transports du Sdes, présente l'objectif de son intervention, qui est d'expliquer comment la statistique publique a construit un système d'information autour des véhicules routiers, de leur utilisation et de leurs utilisateurs.

Il décrit d'abord l'état actuel de l'information disponible. Le répertoire statistique des véhicules routiers (RSVERO) contient des données sur tous les véhicules circulant en France et possédant une plaque d'immatriculation. Ce répertoire rassemble diverses informations, notamment :

- les caractéristiques techniques des véhicules (motorisation, puissance fiscale, types de carrosserie) ;
- les éléments de la vie du véhicule (date d'achat, reventes successives) ;
- des informations sur l'utilisateur actuel (identité partielle, localisation à la commune, revenus, informations sur le ménage) ;
- le kilométrage réel du véhicule.

Ces données permettent d'établir diverses statistiques, notamment sur :

- les immatriculations mensuelles ou annuelles (véhicules neufs et d'occasion) ;
- le parc en circulation (ensemble des véhicules circulant en France) ;
- les kilométrages parcourus, analysés par caractéristiques du véhicule ou profil de l'utilisateur.

L'exhaustivité de ce répertoire permet des analyses très fines, du niveau national jusqu'au niveau communal.

Laurent LEQUIEN explique ensuite l'évolution du système d'information. Avant 2013, le système d'immatriculation des véhicules (SIV) du ministère de l'Intérieur constituait la base de données principale. Ce système, géré par l'Agence nationale des titres sécurisés (ANTS), contient des informations techniques sur les véhicules, des données sur leurs utilisateurs, et enregistre toutes les opérations administratives liées à la vie du véhicule.

Le SIV permettait de suivre efficacement les immatriculations, mais présentait des limites pour l'estimation du parc en circulation. En effet, certains véhicules pouvaient rester enregistrés dans le système alors qu'ils n'étaient plus en circulation (exportation, abandon). Pour pallier ce problème, le service statistique limitait ses analyses aux véhicules de moins de 15 ans pour les voitures.

En 2013, une évolution majeure a eu lieu. Un décret a autorisé l'appariement des données du SIV avec deux nouveaux fichiers :

- les résultats des contrôles techniques ;
- le répertoire SIRENE.

L'intégration des données de contrôle technique a permis d'améliorer considérablement l'estimation du parc de véhicules en circulation. En effet, le passage au contrôle technique est considéré comme un « signe de vie » du véhicule. De plus, les relevés de kilométrage effectués lors de ces contrôles permettent d'estimer les distances annuelles parcourues par les véhicules.

Cependant, la mise en œuvre de cet appariement a présenté des défis techniques et méthodologiques. Les principaux obstacles étaient :

- le volume important de données à traiter (plus de 100 millions de véhicules, environ 100 000 opérations quotidiennes dans le SIV, et autant de contrôles techniques) ;
- la nécessité de définir des critères pour gérer les retards de contrôle technique sans exclure prématurément les véhicules du parc.

Ces défis ont nécessité plusieurs années de travail pour aboutir à un appariement opérationnel et efficace. Le Sdes a continué à publier un parc tronqué par âge pendant les premières années suivant 2013. Une fois par an, un appariement était réalisé entre le SIV et les données de contrôle technique pour publier le Bilan de la circulation, fournissant des informations sur les kilométrages parcourus en France.

Les avancées méthodologiques et informatiques ont conduit à la création d'un nouveau produit, appelé RSVERO 2 en interne, qui a été finalisé en 2021. Dès 2020, ce système a permis de publier pour la première fois un parc complet de véhicules, sans limites d'âge, en exploitant pleinement les données des contrôles techniques.

L'avancée de 2021 a permis de disposer d'un répertoire totalement opérationnel et intégré. Un serveur informatique centralise les données du SIV, des contrôles techniques et celles récupérées via SIRENE pour les utilisateurs professionnels. La mise à jour quotidienne de ce répertoire permet une grande réactivité dans la production d'informations.

Une méthodologie stable a été développée pour extraire les données relatives au parc automobile en circulation. Cette approche permet notamment de publier mensuellement un suivi des immatriculations des mois précédents. Par exemple, le 2 juin, les chiffres du mois de mai seront disponibles, suivis le lendemain par des analyses sur les émissions de gaz à effet de serre théoriques des véhicules immatriculés.

Pour illustrer ces propos, deux graphiques sont présentés. Le premier montre l'évolution des immatriculations par année et par type de motorisation. En 2010, les véhicules diesel dominaient largement le marché. Progressivement, leur part a diminué pour ne représenter que 6 % des immatriculations actuellement. Les véhicules essence ont d'abord pris le relais jusqu'en 2019-2020, avant de céder du terrain aux véhicules électriques et aux hybrides essence non rechargeables. Cette recomposition du parc automobile est particulièrement visible dans les immatriculations annuelles, bien que son impact sur le stock total de véhicules en circulation soit plus lent.

Le second graphique illustre le type d'informations désormais disponibles sur le parc en circulation, telles que l'âge moyen des véhicules ou la proportion de véhicules électriques. Ces données, auparavant difficiles à obtenir, sont maintenant facilement accessibles grâce au nouveau système.

Malgré ces avancées, une demande croissante émerge pour obtenir des informations sur les profils des utilisateurs de véhicules. Il s'agit notamment d'identifier les grands rouleurs, les propriétaires de véhicules anciens, ou encore les personnes susceptibles d'être les plus impactées par d'éventuelles mesures de renouvellement du parc automobile. Cependant, les données disponibles se limitent actuellement au nom, prénom et commune de résidence, ce qui est insuffisant pour établir des profils détaillés.

Pour pallier ce manque, RSVERO a été apparié avec le Fichier démographique sur les logements et les individus (Fideli). Cette démarche permet d'accéder à des informations plus complètes sur les revenus, la composition des ménages et la localisation des utilisateurs de véhicules. L'appariement de ces deux fichiers exhaustifs offre la possibilité de réaliser des analyses territoriales fines et d'étudier des profils très spécifiques.

Le processus d'appariement a bénéficié du soutien de l'équipe de l'Insee en charge du CSNS. Grâce à cette collaboration, il est désormais possible de retrouver les revenus pour près de 90 % des utilisateurs de véhicules. Les 10 % restants s'expliquent principalement par une qualité insuffisante du CSNS ou par l'absence de données de revenus pour certaines catégories de population (étudiants en cité universitaire, personnes âgées en Ehpad).

Laurent LEQUIEN souligne que cette marge d'erreur de 10 % peut parfois poser problème pour certaines statistiques spécifiques, comme l'analyse du nombre de véhicules par ménage. Néanmoins, ces cas restent marginaux et, dans l'ensemble, le système fournit des données fiables et précieuses pour l'analyse du parc automobile français.

Il présente ensuite les avancées récentes dans l'analyse des données liées à la transition écologique dans le secteur automobile. L'appariement entre RSVERO et Fideli a permis d'obtenir des informations éclairantes sur les effets individuels de la transition écologique. Ainsi, les personnes les plus modestes possèdent généralement des voitures plus anciennes et plus souvent Diesel. Concernant les acquéreurs de voitures électriques, il s'agit en moyenne de personnes plus âgées, mais aussi de jeunes et d'habitants

des zones rurales. Ces données continuent d'être analysées pour exploiter pleinement la richesse de cet appariement.

En 2023, un travail d'affinement des émissions de gaz à effet de serre produites à partir du RSVERO a été entrepris. Cette estimation se base sur le produit des kilométrages observés et de la consommation moyenne des véhicules. Pour aller au-delà de la consommation théorique fournie par les constructeurs, les données d'un site collaboratif allemand (SpritMonitor.de) ont été utilisées. Ce site recueille les informations précises sur les modèles de véhicules, les pleins d'essence ou de diesel effectués, ainsi que les kilomètres parcourus. Cette méthode permet d'obtenir une « consommation réelle » pour chaque type de véhicule, affinant ainsi l'estimation des émissions de gaz à effet de serre.

Actuellement, un écosystème complet de données renseigne sur les véhicules et leurs utilisateurs grâce à deux appariements : l'un avec Fideli pour les particuliers, l'autre avec Sirene pour les personnes morales. Les informations sur l'utilisation des véhicules proviennent des contrôles techniques.

Deux nouveaux appariements sont prévus pour 2025. Le premier vise à lier RSVERO avec les données sur les aides à l'acquisition de véhicules peu polluants, gérées par l'Agence des services et des paiements. Cet appariement permettra une analyse plus complète de l'impact de ces aides, en comparant les ménages bénéficiaires et non bénéficiaires.

Le second appariement, plus expérimental, concerne le fichier du véhicule assuré géré par l'Association pour la gestion des informations sur le risque en assurance (Agira). Cette initiative pourrait améliorer l'estimation du parc en circulation, l'équipement en assurance des véhicules immatriculés fournissant un « signe de vie » supplémentaire des véhicules. Elle pourrait également aider à estimer le parc de deux-roues motorisés, actuellement mal connu.

Laurent LEQUIEN conclut en rappelant que les données issues de RSVERO sont disponibles en open data sur le site du Sdes, avec des informations sur les parcs et les immatriculations à différents niveaux géographiques. Des fichiers individuels sont également mis à disposition des chercheurs via le CASD.

Clément MALGOUYRES, chercheur au Crest et à l'IPP, dresse un bilan de la situation antérieure à 2013. À cette époque, la plupart des études s'appuyaient sur le SIV. Ces données étaient souvent retraitées par des sociétés commerciales ou des associations de consultants automobiles. Par conséquent, les évaluations de politiques publiques sur le marché automobile utilisaient des données commerciales plutôt que des données directement issues de la statistique publique.

Il cite plusieurs exemples d'études réalisées avec ces données, notamment sur le bonus-malus écologique, les préférences des ménages pour les véhicules environnementaux, et l'impact du prix du carburant sur le choix des véhicules. Ces études s'appuyaient généralement sur des données datant de 2008-2009.

Plus récemment, une étude menée par Kessler et ses collaborateurs a utilisé des données achetées par l'Agence de l'environnement et de la maîtrise de l'énergie (Ademe) à AAA DATA, puis transmises aux chercheurs. **Clément MALGOUYRES** souligne que RSVERO présente l'avantage d'offrir une source unique et plus claire de données. Il note également que toutes les études mentionnées ne portaient que sur les immatriculations de véhicules neufs, sans information sur le parc existant. Cette limitation nécessitait des hypothèses fortes pour estimer l'impact des politiques publiques sur l'ensemble du parc automobile.

Auparavant, les études s'intéressant aux effets redistributifs en matière d'équité verticale devaient s'appuyer uniquement sur le lieu de résidence et l'âge communiqués dans des données commerciales pour approximer le revenu. L'apport de RSVERO-Fideli est donc considérable, particulièrement dans la littérature d'organisation industrielle qui souligne l'importance de l'hétérogénéité des ménages pour déterminer la demande à laquelle répondent les constructeurs et comme facteur déterminant de leurs stratégies commerciales.

La prise en compte des aspects distributifs et des interactions avec le marché de l'occasion constitue ainsi une avancée majeure permise par RSVERO-Fideli, grâce à l'appariement avec les données d'immatriculation. Ce dispositif permet d'estimer la consommation de carburant via le kilométrage, mais permettra également, une fois apparié à Fideli, de mieux comprendre l'incidence de la fiscalité sur les carburants. De nombreuses études influentes, notamment dans le contexte du mouvement des Gilets jaunes, reposaient sur des données d'enquête de qualité, mais avec des échantillons limités, restreignant mécaniquement la prise en compte de la dimension géographique. RSVERO-Fideli apporte une amélioration significative en croisant les dimensions revenu et géographie avec une précision accrue.

Une publication récente sur l'impact redistributif des Zones à faibles émissions (ZFE) s'appuie déjà sur RSVERO-Fideli. Cette étude examine les ménages les plus susceptibles d'avoir acheté un véhicule qui sera impacté par les ZFE et réalise une analyse ZFE par véhicule, démontrant l'utilité concrète de ces données.

Clément MALGOUYRES formule ensuite une suggestion qui dépasse le cadre des appariements. Il propose de rendre directement accessibles aux chercheurs via le CASD l'ensemble des bases de données sur les dispositifs, comme le bonus écologique, la prime à la conversion ou l'aide sociale. Si l'appariement représente l'option idéale, la mise à disposition des bases brutes anonymisées permettrait également des analyses plus fines. Pour illustrer ce point, il évoque une mission reçue de France Stratégie dans le cadre de France Relance pour évaluer la prime à la conversion, pour laquelle l'accès direct aux données via le CASD n'était pas disponible. Son équipe a dû s'arranger avec l'aide de différents services pour obtenir des versions agrégées au niveau communal, ce qui a compliqué l'évaluation.

Dans un autre exemple de ses travaux actuels sur des données commerciales d'immatriculations, il s'intéresse aux véhicules ayant perdu l'accès au bonus écologique suite à l'entrée en vigueur de critères environnementaux. La connaissance précise de l'impact de cette perte d'accès aux aides serait facilitée par l'accès à des données microéconomiques, même non appariées.

L'appariement des bases de l'Agence de services et de paiement (ASP) avec RSVERO représentera une amélioration qualitative majeure pour l'analyse de dispositifs très débattus. Ces aides sont théoriquement redistributives avec des barèmes progressifs favorisant les ménages modestes, mais tendent paradoxalement à subventionner des biens de luxe, comme les véhicules neufs, particulièrement électriques. Ces forces contradictoires pourront être analysées avec plus de précision grâce aux données appariées, mais la simple mise à disposition des données ASP constituerait déjà une avancée.

En résumé, ces initiatives représentent une amélioration qualitative substantielle pour la compréhension du parc automobile et des politiques publiques affectant ses dynamiques. **Clément MALGOUYRES** souligne également l'utilité de la mise à disposition en open data pour l'enseignement et les projets étudiants. Les appariements sont particulièrement bienvenus pour évaluer les politiques publiques qui, jusqu'à présent, étaient analysées sous des hypothèses parfois simplificatrices, comme celle d'un recours systématique aux aides disponibles.

.2 Que nous ont appris les appariements sur la divergence récente des sources statistiques sur l'emploi ?

Christine LAGARENNE accueille désormais Yves Jauneau, chef de la division Synthèse et conjoncture du marché du travail de l'Insee, qui interviendra sur l'utilité des appariements pour comprendre la divergence des sources statistiques sur l'emploi. Sa présentation sera discutée par Magali Dauvin, économiste à l'Observatoire français des conjonctures économiques (OFCE).

Un document est projeté à l'ensemble des participants.

Yves JAUNEAU, Chef de la division Synthèse et conjoncture du marché du travail de l'Insee, présente les travaux réalisés sur l'appariement de bases de données relatives à l'emploi et au chômage.

L'objectif principal était de comparer deux sources d'information sur l'emploi : d'une part, les données administratives et, d'autre part, les données d'enquête. Les données administratives sur l'emploi se composent principalement de deux ensembles : pour les salariés, les informations proviennent désormais à 99 % de la Déclaration sociale nominative (DSN), tandis que, pour les non-salariés, l'Insee dispose d'une base spécifique qui compile différents fichiers administratifs sur cette population. Du côté des données d'enquête, l'enquête Emploi de l'Insee interroge environ 90 000 personnes chaque trimestre.

Cette enquête présente un intérêt particulier, car elle constitue le seul instrument permettant de mesurer le chômage au sens du BIT, tout en fournissant également des indicateurs sur l'emploi. La démarche visait donc à établir une comparaison méthodique entre ces deux sources.

D'un point de vue pratique, les premiers travaux d'appariement ont débuté avant l'arrivée du CSNS, nécessitant l'utilisation de méthodes traditionnelles basées sur la comparaison d'identifiants, tels que les prénoms et les dates de naissance dans les différentes sources. Progressivement, l'implémentation du CSNS dans l'enquête Emploi en continu (EEC) a facilité ces appariements, bien que la mise en œuvre du CSNS dans l'EEC reste complexe et continue d'être améliorée.

La difficulté principale réside dans les spécificités de l'enquête Emploi où tous les noms de famille des personnes vivant dans un même logement ne sont pas systématiquement collectés. Cette situation complique l'attribution du CSNS dans certaines configurations d'habitation, comme les colocations de jeunes adultes.

Les travaux d'appariement se sont déroulés entre 2023 et 2024. Après de nombreuses manipulations des données dans différentes configurations, la première analyse a porté sur le phénomène de sous-déclaration de l'emploi. Ce concept mérite précision : il s'agit de situations où une personne, bien que répondante dans l'enquête Emploi (éliminant ainsi le biais d'échantillonnage), se déclare sans emploi au sens du BIT (donc chômeuse ou inactive), alors que les données administratives indiquent le contraire.

Pour les salariés, la mesure est très précise grâce à la DSN, permettant un suivi quotidien de l'emploi. Cette précision permet de s'aligner exactement sur la semaine de référence de l'enquête Emploi. En revanche, la situation est plus complexe pour les non-salariés, car les données administratives ne fournissent pas de mesure intra-annuelle de l'emploi. La comparaison se fait donc entre l'emploi du quatrième trimestre de l'enquête Emploi et l'emploi administratif en fin d'année.

Dans ce contexte, la sous-déclaration peut masquer des différences conceptuelles et de mesure. Par exemple, un non-salarié considéré comme employé en fin d'année dans les données administratives pourrait ne pas travailler lors de la semaine de référence de l'enquête Emploi. Cette divergence relève davantage d'une différence de concept que d'une véritable sous-déclaration.

L'appariement des données a révélé que la grande majorité des emplois des données administratives sont effectivement déclarés dans l'enquête-emploi. Cependant, un taux global de sous-déclaration d'environ 4 % a été identifié pour l'ensemble des emplois, avec des variations significatives selon les catégories.

Trois types d'emplois présentent des taux de sous-déclaration particulièrement élevés :

- les alternants (apprentis et contrats de professionnalisation) : 12 % de sous-déclaration ;
- les microentrepreneurs dont c'est l'activité principale : 14 % de sous-déclaration ;
- les salariés non-microentrepreneurs de 60 ans ou plus : 19 % de sous-déclaration.

Ces trois catégories représentent environ 10 % de l'emploi en France. Pour les 90 % restants, le taux de sous-déclaration est plus faible, autour de 3 %, variant principalement en fonction de l'âge et du type de contrat (par exemple, légèrement plus élevée pour les CDD).

L'analyse approfondie de ces trois populations révèle des mécanismes de sous-déclaration distincts. Pour les alternants, deux facteurs principaux entrent en jeu. D'une part, les réponses par procuration (proxy) augmentent significativement le risque de sous-déclaration, celle-ci étant deux fois plus élevée dans ces cas. D'autre part, la sous-déclaration est plus fréquente pour les contrats d'apprentissage que pour les contrats de professionnalisation. Cette différence pourrait s'expliquer par une plus grande probabilité d'être interrogé pendant une période de formation pour les apprentis, du fait de leur temps de formation plus important par rapport aux contrats de professionnalisation.

Concernant les microentrepreneurs, le revenu et l'âge sont les principaux déterminants de la sous-déclaration. Ainsi, celle-ci atteint 29 % lorsqu'aucun chiffre d'affaires n'a été perçu au quatrième trimestre, mais descend à 4 % (comparable à la moyenne générale) pour un chiffre d'affaires trimestriel supérieur à 9 000 euros. L'âge joue également un rôle important, avec 27 % de sous-déclaration chez les microentrepreneurs de 55 ans ou plus, qui représentent environ 30 % de cette catégorie. Ce taux grimpe à 44 % en cas de cumul avec une pension de retraite, une situation de plus en plus fréquente concernant 14 % des microentrepreneurs.

Pour les seniors de 60 ans ou plus, la sous-déclaration globale de 18 % cache des disparités importantes. Elle atteint 56 % pour les 70 ans ou plus, bien que l'impact soit limité par le faible taux d'emploi dans cette tranche d'âge en France. Le cumul emploi-retraite augmente significativement la sous-déclaration, suggérant des difficultés à capturer certains types d'emplois de courte durée ou à temps partiel dans l'enquête.

Yves JAUNEAU évoque également des hypothèses concernant la fin de carrière, comme la liquidation de compte épargne temps ou les congés, qui pourraient être mal captés par l'enquête.

Il explique que cette analyse de la sous-déclaration visait initialement à comprendre la divergence entre les sources de données sur l'emploi entre 2019 et 2023. Cette période a connu une croissance de l'emploi sans précédent, tant dans l'enquête Emploi que dans les données administratives. Cependant, une différence significative a été observée : 1,3 million d'emplois créés selon l'enquête emploi contre 1,75 million dans les estimations administratives, soit un écart de 450 000 emplois.

L'étude a révélé que les trois catégories présentant le taux de sous-déclaration le plus élevé sont également celles où l'emploi a été particulièrement dynamique. Les chiffres sont particulièrement significatifs : ces trois catégories représentaient 10 % de l'emploi en 2019, mais ont concentré 70 % des créations d'emplois sur la période étudiée.

Ce phénomène a engendré un effet mécanique où l'augmentation des emplois dans des secteurs structurellement moins bien déclarés dans l'enquête Emploi a créé une divergence statistique. Cette contribution a été quantifiée à environ 180 000 emplois, soit presque la moitié de la divergence observée entre les différentes sources de données sur la période.

D'autres facteurs expliquent le reste de cette divergence. Premièrement, le champ de l'enquête Emploi exclut les communautés au sein desquelles l'emploi a été très dynamique entre 2019 et 2023, ce qui représente environ 45 000 emplois non couverts par l'EEC. Deuxièmement, l'échantillon de l'enquête Emploi présente une sous-couverture des logements neufs, plus fréquemment occupés par des personnes en emploi. Troisièmement, on constate une sous-représentation de l'emploi des personnes nées à l'étranger, qui constitue environ 14 % de l'emploi total, mais représente un tiers des créations d'emploi sur la période.

Yves JAUNEAU précise qu'une étude détaillée documentant ces phénomènes et leur mesure est en cours de réalisation. Il souhaite ensuite présenter d'autres travaux réalisés en matière d'appariements sur le marché du travail.

Ces appariements permettent également d'étudier la qualité de certaines variables de l'emploi. Par exemple, la comparaison entre le contrat collecté dans l'enquête Emploi et celui figurant dans la DSN montre une convergence dans 98 % des cas, résultat significatif pour l'évaluation de la qualité des données. Ces analyses constituent également un matériau précieux pour adapter le questionnaire en cas de refonte future de l'enquête Emploi.

Un autre exemple concerne les effets liés aux modes de collecte. Depuis 2021, il est possible de répondre à l'enquête Emploi par Internet. Contrairement aux inquiétudes initiales, le taux de sous-déclaration dans l'emploi s'avère plus faible pour les réponses en ligne que pour les interviews classiques. Cette différence s'explique en partie par des effets de structure, notamment le fait que les réponses par Internet comportent moins de réponses par proxy. Cependant, même en contrôlant ces variables structurelles, comme l'âge et le niveau de diplôme, ce résultat persiste.

Yves JAUNEAU évoque ensuite les travaux menés sur la mesure du Revenu de solidarité active (RSA) dans l'enquête Emploi, sujet devenu particulièrement pertinent avec la loi Pour le plein emploi prévoyant l'inscription automatique des bénéficiaires du RSA à France Travail début 2025. Pour évaluer l'impact de cette loi sur les indicateurs du marché du travail, une mesure fiable du RSA dans l'enquête était nécessaire. L'analyse a révélé que la variable existante sous-estimait significativement le nombre de bénéficiaires.

Au troisième trimestre 2024, des modifications du questionnaire concernant la formulation et le ciblage des questions ont permis d'augmenter la couverture du RSA de 20 %. Dans ce contexte, l'appariement avec les données administratives du RSA a joué un rôle crucial, permettant de mesurer rapidement la qualité de la variable dans l'enquête Emploi. Grâce à la nouvelle méthodologie, l'enquête couvre désormais 90 % des bénéficiaires du RSA.

Sur le graphique présenté, la courbe bleue représente le taux de bénéficiaires du RSA mesuré avant le changement de questionnaire, tandis que la partie verte illustre ce qui manque encore. Cette analyse démontre que les bénéficiaires non couverts sont relativement bien répartis entre les diverses populations, permettant ainsi une mesure fiable des indicateurs du marché du travail sans recourir systématiquement à l'appariement. Cette approche est d'autant plus utile que l'appariement ne peut pas être réalisé en temps réel. Par exemple, pour les données de l'enquête Emploi du premier trimestre 2025 publiées mi-mai 2025, le fichier nécessaire à l'appariement n'était pas encore disponible.

Yves JAUNEAU mentionne un dernier cas concernant le chômage. Un appariement réalisé entre 2013 et 2017 a comparé les données de l'enquête Emploi avec celles des inscrits à Pôle emploi (devenu France

Travail), dans le but de comprendre les divergences d'évolution entre le nombre de chômeurs au sens du BIT et les demandeurs d'emploi en catégorie A. Bien que ces indicateurs diffèrent par nature, leur évolution fait souvent l'objet de comparaisons dans le débat public, justifiant cette analyse comparative.

L'appariement exploité conjointement par l'Insee, la Dares et Pôle emploi, malgré un certain retard, a donné lieu à plusieurs publications. Il a permis de comprendre, par exemple, que, parmi les inscrits en catégorie A, 56 % se trouvent au cœur du chômage au sens du BIT. De plus, une proportion significative de personnes inactives a été identifiée au sein des demandeurs d'emploi en catégorie A.

Pour 2025, un nouveau projet d'appariement est envisagé. Celui-ci vise à reproduire les travaux précédents tout en analysant l'évolution des indicateurs entre 2019 et 2024. Cette période est particulièrement intéressante, car elle a vu une diminution du nombre d'inscrits à Pôle emploi. L'appariement pourrait ainsi permettre de déterminer si cette baisse est due à une diminution du taux d'inscription pour certaines populations spécifiques, comme les seniors proches de la retraite.

Un point crucial à souligner est l'amélioration significative de l'efficacité et de la qualité du processus d'appariement, grâce notamment à l'implémentation progressive du CSNS. Cette évolution positive ouvre de nouvelles perspectives. Il serait ainsi envisageable de reproduire cet appariement annuellement, ce qui permettrait d'obtenir des analyses plus fréquentes et avec moins de retard.

Cependant, certaines limites méritent d'être mentionnées. La masse de données générée par les appariements peut s'avérer impressionnante. Ce constat est particulièrement frappant dans l'appariement sur l'emploi. En conséquence, il est crucial de définir clairement les objectifs de l'étude dès le départ. En effet, la richesse des données appariées peut parfois conduire à des analyses non pertinentes si les objectifs ne sont pas bien définis en amont.

Magali DAUVIN, économiste à l'OFCE, partage son expérience en tant qu'utilisatrice des données issues d'appariements statistiques. Dans le cadre de son travail, qui implique des exercices de prévision bisannuels sur le marché du travail, ces appariements s'avèrent particulièrement utiles.

L'apport principal de ces appariements réside dans l'amélioration de la qualité d'utilisation des données de l'Insee. En effet, ils permettent d'objectiver des concepts qui, bien que censés être identiques, ne donnent pas toujours les mêmes résultats dans différentes bases de données.

Ces travaux d'appariement ont permis de répondre à des questions cruciales, notamment concernant les écarts observés ces dernières années entre l'enquête Emploi et les sources administratives. Par exemple, ils ont mis en lumière des phénomènes inattendus, tels que la sous-déclaration des travailleurs nés à l'étranger ou des seniors.

L'appariement méthodologique joue également un rôle d'audit, permettant d'améliorer la mesure sans nécessairement modifier la collecte des données. Cette démarche enrichit la réflexion des utilisateurs et leur permet de communiquer plus précisément sur ces sujets, que ce soit dans leurs études, dans le débat public ou auprès de leurs étudiants.

Magali DAUVIN soulève ensuite une question d'ordre européen. Elle mentionne que l'institut statistique allemand a également constaté des écarts entre sources administratives et enquêtes Emploi. Elle s'interroge donc sur la possibilité de mutualiser les diagnostics au niveau européen, notamment dans le contexte d'une évolution des formes d'emploi que les enquêtes traditionnelles peinent parfois à capter.

Enfin, elle pose une question sur le potentiel de Résil pour améliorer la mesure de l'emploi dans l'enquête Emploi, notamment en ce qui concerne la prise en compte des logements non ordinaires.

Concernant la dimension européenne, **Yves JAUNEAU** confirme que plusieurs pays s'intéressent à cette problématique, bien que les enjeux diffèrent selon les pays. Le Lamas, groupe responsable des enquêtes *Labour Force Survey* (LFS) européennes, peut être un lieu de discussion sur ces sujets, mais ce n'est pas son objectif principal.

Il souligne cependant que les appariements et comparaisons de sources serviront davantage pour les évolutions futures. En effet, le règlement européen vise à éviter des changements fréquents de méthodologie pour garantir la continuité des séries statistiques. Ainsi, malgré la tentation de modifier immédiatement les questions suite à de nouvelles expertises, il est préférable de regrouper tous les changements sur une seule période.

Concernant Résil et la couverture des logements non ordinaires, **Yves JAUNEAU** confirme que des réflexions sont en cours à l'Insee pour réfléchir à l'extension des enquêtes aux ménages vivant en logements non ordinaires.

Pour l'enquête Emploi, il faut rappeler que le champ d'application des LFS au niveau européen s'arrête aux ménages vivant en logements ordinaires. Passé de grappes géographiques à des communautés plus larges semble également complexe. Concernant la méthodologie, une extension limitée à certains individus vivant en communauté de l'enquête Emploi a été réalisée. Celle-ci inclut les personnes vivant dans des logements ordinaires, mais passant une partie de l'année dans des logements non ordinaires, permettant ainsi une première analyse méthodologique sur cette population spécifique. **Yves JAUNEAU** note cependant qu'en temps normal, l'emploi des ménages vivant en logements ordinaires augmente généralement au même rythme.

Christine LAGARENNE remercie Yves Jauneau pour son intervention, soulignant l'intérêt de l'exemple méthodologique présenté ainsi que du premier exemple concernant RSVERO, qui offre des perspectives prometteuses pour la communauté des chercheurs.

Stéphane JUGNOT, Céreq, partage les préoccupations exprimées par des chercheurs au sein d'un groupe de travail initié par l'Institut national d'études démographiques (Ined) sur les fichiers de production et de recherche (FPR). Ces inquiétudes portent sur l'enrichissement des enquêtes par des données administratives et la substitution potentielle des données d'enquête par ces dernières.

La crainte principale concerne le risque que le FPR, disponible gratuitement sur la plateforme Progedo, ne contienne plus l'intégralité des informations précédemment accessibles. En effet, certaines données pourraient désormais relever du CASD, entraînant un morcellement des données et un accès potentiellement plus coûteux.

Par ailleurs, **Stéphane JUGNOT** souligne l'importance des appariements et rappelle la pertinence des limites rappelées par Corinne Prost en introduction, souvent négligées par certains responsables de services producteurs. Il insiste particulièrement sur l'importance des principes de minimisation et de transparence, craignant que le cadre de référence puisse ne pas être suffisamment approfondi sur ces aspects cruciaux.

Christel COLIN note cette préoccupation vis-à-vis des fichiers de production et de recherche ; garder une diversité de modes de mises à disposition des données est un sujet essentiel. Il s'agit donc d'un point de vigilance à garder à l'esprit.

Kamel GADOUCHE, directeur du CASD, confirme que ce point est effectivement un point de vigilance important. Un FPR doit être anonymisé, indépendamment des sources utilisées. Cette question n'a pas de rapport direct avec un accès par l'intermédiaire du CASD.

Roxane SILBERMAN, directrice de recherche émérite au CNRS et membre du Bureau du Cnis, rappelle le cas il y a quelques années de l'utilisation d'une source fiscale pour remplacer une question sur le revenu : à l'époque, les données fiscales n'étaient pas accessibles aux chercheurs et du coup l'information sur le revenu initialement disponible dans l'enquête ne pouvait plus être transmise aux chercheurs.

Elle souligne l'importance des appariements pour la recherche qui suppose cependant d'obtenir l'accord des administrations pour l'ouverture de leurs données. Elle rappelle à ce sujet qu'il existe deux procédures distinctes pour les appariements : une pour la statistique publique et une autre pour les chercheurs. Cette dernière ne peut fonctionner que si les données administratives sont ouvertes.

Roxane SILBERMAN se demande s'il ne serait pas plus simple de ce fait que les chercheurs puissent demander à l'Insee de réaliser ces appariements avec les données administratives auxquelles l'Insee peut accéder.

Elle souligne ensuite l'importance croissante de l'utilisation des données administratives et des appariements avec ces donnée au niveau européen, avec cependant des différences historiques entre pays, les pays dits à registres reposant initialement pour leurs statistiques sur ces données. Elle suggère qu'il pourrait y avoir une opportunité au moment où cette utilisation et les appariements se généralisent à une initiative pour aller vers plus d'homogénéité au niveau européen, l'utilisation de données administratives accroissant les risques en matière de comparabilité des données entre pays.

Maryse FESSEAU, Sous-directrice des statistiques et des synthèses, Ministère de l'Éducation nationale, intervient sur la question de la minimisation des données dans le processus d'appariement. Elle s'interroge sur le moment où la pertinence des variables mises à disposition est discutée, au-delà de la simple vérification de l'objectif de l'étude. Elle demande si, dans le processus actuel, il existe une étape spécifique où l'on discute de la réduction du nombre de variables, considérant que plus on apparie de bases, plus les risques mentionnés précédemment augmentent.

Nagui BÉCHICHI, adjoint à la cheffe du bureau de l'appui à l'évaluation des politiques publiques et de soutien à la recherche, s'adresse ensuite à Laurent Lequien, exprimant son intérêt pour l'utilisation du site internet allemand permettant d'obtenir la consommation réelle des véhicules. Il s'interroge sur les règles ou consignes appliquées permettant d'estampiller « statistiques publiques » les données issues de sources moins conventionnelles.

Laurent LEQUIEN reconnaît le caractère atypique de cette source pour la statistique publique. Son utilisation s'explique par l'absence d'autres informations disponibles. La méthodologie employée est la suivante : pour un modèle de véhicule spécifique, un nombre suffisant d'observations individuelles est nécessaire afin d'obtenir des données jugées fiables. Cette fiabilité se traduit par une relative stabilité des données d'une année à l'autre et une cohérence avec les données théoriques. En l'absence d'un échantillon suffisant, la définition du modèle est élargie. Par exemple, pour une Clio d'une génération et d'une motorisation données, si les données sont insuffisantes, les critères de motorisation peuvent être supprimés. Cette approche permet d'obtenir un échantillon plus large, certes moins précis, mais potentiellement plus pertinent.

Concernant la minimisation, **Corinne PROST** expose plusieurs situations et enjeux. Dans le cas du dispositif InserJeunes, les impératifs de production et les contraintes légales ont conduit à une application rapide du principe de minimisation. Le nombre de variables a été restreint pour assurer l'opérationnalité du dispositif dans tous les domaines.

Pour les études, la tendance est inverse : les demandes sont généralement plus expansives, avec peu d'accent sur la minimisation.

Des enjeux techniques de minimisation persistent néanmoins dans la production de fichiers, notamment pour les panels. L'identification des variables sur plusieurs années entraîne une complexité qui rend impossible la conservation de toutes les variables.

La minimisation intervient en amont du processus, lors de la constitution des fichiers, et non au moment de la diffusion aux chercheurs. L'objectif est de créer des fichiers ciblés répondant à des questions spécifiques, plutôt que des fichiers exhaustifs, mais difficiles à exploiter.

Corinne PROST souligne l'importance de sélectionner les informations les plus pertinentes pour des raisons pratiques et techniques.

Olivier LEFEBVRE, maître d'ouvrage du programme Résil, Insee, apporte des précisions sur l'évolution de la minimisation des données. Historiquement, la minimisation était principalement dictée par des contraintes techniques. Bien que ces contraintes persistent, elles sont aujourd'hui considérablement repoussées.

Désormais, la minimisation est davantage une question d'usage. Il s'agit de déterminer les données réellement nécessaires pour analyser un phénomène spécifique. Il illustre ce point par deux exemples.

Pour une enquête sur l'emploi nécessitant des informations sur le logement, quelques variables clés suffisent (statut de propriétaire, qualité du logement). À l'inverse, une étude centrée sur le logement requiert des informations détaillées sur ce sujet, mais peut se contenter de données plus générales sur l'emploi ou les revenus.

Cette approche s'inscrit par ailleurs dans le respect du principe de minimisation du RGPD, visant à ne pas collecter plus de données que nécessaire par rapport à la finalité poursuivie.

Olivier LEFEBVRE souligne qu'il n'existe pas de doctrine universelle pour la minimisation. Chaque traitement nécessite une appréciation spécifique. Il conclut en notant que cette réflexion sur la minimisation est salutaire, obligeant à mieux considérer les besoins en données. Les avancées techniques facilitent désormais une sélection plus fine des données nécessaires.

Jan Robert SUESSER, Ligue des droits de l'Homme, intervient pour apporter une perspective plus large sur les évolutions en cours dans la statistique publique. Il souligne l'importance des réflexions menées au niveau international, notamment dans le cadre de la Commission économique pour l'Europe des Nations unies (Unece), concernant les bonnes pratiques et l'éthique professionnelle dans la statistique publique.

Il met en évidence le contraste entre l'approche traditionnelle de la statistique publique, axée sur la protection de la vie privée et le fait de « ne pas nuire » aux personnes dont les données sont utilisées, et les discussions actuelles, dans le cadre d'un accès très diversifié aux données et à la puissance des outils de traitement qui considère qu'il faut étendre le champ des préoccupations éthiques pour la production des statistiques publiques.

Jan Robert SUESSER identifie trois principes fondamentaux guidant la réflexion dans la statistique publique :

- produire une information statistique pour le bien commun ;
- l'engagement à ne pas nuire aux personnes et aux communautés auxquelles les classifications statistiques les rattachent ;
- la confiance en la production de la statistique publique qui engage les institutions au-delà de la déontologie des personnes qui y travaillent.

Il note que ces réflexions sont en pleine évolution et qu'elles s'intéressent en particulier aux pratiques d'appariement de données et de leurs traitements du fait des changements des deux dernières décennies.

Il suggère que, du fait du changement d'échelle de ce que les appariements peuvent couvrir, il serait nécessaire d'avoir des pratiques transitoires adaptées à cette période de changement, avec des process de validation de demandes d'appariements renforcés en particulier lorsque sont incluses certaines variables dont l'utilisation dans les débats publics, voire la décision publique, indiquent des risques d'usages qui peuvent nuire.

Jan Robert SUESSER attire l'attention sur une phrase particulièrement frappante dans un document qui devrait être prochainement adopté par l'UNECE : « Ce n'est pas parce qu'on peut faire qu'on doit faire ou qu'on va faire », représente selon lui un changement radical dans le milieu de la statistique publique. En effet, on n'aurait pas trouvé une telle affirmation il y a encore cinq de cela.

Cette nouvelle approche met en question de nombreux raisonnements établis quant à la légitimité de ce qu'est la construction d'une connaissance statistique. Il note que bien que le document ait été rédigé avant l'arrivée de Donald Trump au pouvoir, à l'évidence il prend une résonance particulière dans le contexte étatsunien actuel. Les collègues américains sont actuellement contraints de réfléchir dans l'urgence à deux aspects cruciaux : comment protéger les données contre un détournement d'usage par un pouvoir qui s'affranchi des règles, comment assurer leur disponibilité pour les usages légitimes liés aux décisions publiques et débats publiques.

Mathilde GAINI, sous-directrice du suivi et de l'évaluation des politiques d'emploi et de formation professionnelle de la Dares, Ministère du Travail, soulève une problématique concernant l'utilisation des données administratives. Ces données sont particulièrement riches, mais généralement, un seul fichier est produit pour la diffusion. Or, les besoins de la recherche sont multiples et variés. Idéalement, il serait préférable de disposer d'un fichier distinct pour chaque projet de recherche ou usage spécifique.

Cependant, cette approche idéale se heurte à des contraintes pratiques. Deux options se présentent alors : soit minimiser dès le départ en limitant les usages, soit produire des fichiers par projet. Néanmoins, cette dernière option soulève la question des ressources nécessaires, car il n'est pas toujours possible de générer un fichier spécifique pour chaque projet.

La séance est suspendue de 11 heures 10 à 11 heures 30.

Christine LAGARENNE annonce que Pierrette Schuhl, sous-directrice des systèmes d'information et des études statistiques (Sies) du Ministère de l'Enseignement supérieur, présentera InserSup. Dans un second temps, Nagui Béchichi, créateur du site Suptracker, exposera les apports de ce dispositif.

.3 InserSup, orienter les jeunes dans l'enseignement supérieur

Un document est projeté à l'ensemble des participants.

Pierrette SCHUHL, sous-directrice du Sies du Ministère de l'Enseignement supérieur, explique qu'InserSup vise à mesurer et qualifier l'insertion professionnelle des diplômés de l'enseignement supérieur. Ce dispositif poursuit deux objectifs principaux : d'une part, aider les jeunes et leurs familles dans les processus d'orientation, tant à l'entrée de l'enseignement supérieur qu'au cours des études ; d'autre part, contribuer au pilotage de l'offre de formation. Ce second objectif s'est révélé au cours de la mise en place du système.

La conception d'InserSup repose sur l'appariement de fichiers administratifs exhaustifs. Ce modèle s'inspire d'InserJeunes, développé par la Depp en collaboration avec la Dares, pour le suivi des sortants des formations scolaires jusqu'au BTS.

Le contexte législatif ayant conduit à la création d'InserSup est multiple. Les lois LRU de 2007 et 2013 ont imposé aux établissements universitaires de publier des statistiques sur l'insertion professionnelle de leurs formations. Cette obligation a entraîné la mise en place d'enquêtes statistiques exhaustives, d'abord menées par les établissements eux-mêmes, puis coordonnées par le Sies pour harmoniser la méthodologie. Ces enquêtes se concentraient principalement sur les formations à visée d'insertion professionnelle immédiate, telles que les licences professionnelles, les masters, les DUT et les doctorats.

Bien que ces enquêtes ne fussent pas réalisées par des services de la statistique publique, le Sies était chargé du redressement des données et de la publication des résultats au niveau national. Les informations étaient également mises à disposition en open data pour chaque établissement. Cependant, les taux de réponse variaient considérablement, de 30 % à 100 %, ce qui compliquait l'obtention de taux d'insertion professionnelle fiables pour chaque formation.

La loi de 2018 sur la liberté de choisir son avenir professionnel a ensuite exigé la publication des taux d'insertion pour chaque centre de formation en apprentissage. En réponse, la Depp et la Dares ont créé InserJeunes, ouvrant ainsi la voie au développement d'InserSup.

Par ailleurs, la loi de programmation de la recherche de 2020 a imposé aux universités de présenter un rapport sur la situation professionnelle des titulaires d'un doctorat à un, trois et cinq ans après l'obtention du diplôme.

Une demande interministérielle a également joué un rôle crucial dans le développement d'InserSup. Initialement prévue pour fin 2024, la mise en production du système a été avancée d'un an sous une forte pression gouvernementale. L'objectif était de mesurer l'insertion professionnelle des sortants de l'enseignement supérieur à partir de fichiers administratifs, en s'inspirant d'InserJeunes, et d'afficher ces informations sur la plateforme Parcoursup.

InserSup vise à calculer les proportions de sortants du supérieur occupant un emploi au niveau global et au niveau des mentions. Grâce aux fichiers de la DSN, le système permet d'observer la situation professionnelle à 6, 12, 18, 24 et 30 mois après l'obtention du diplôme, offrant ainsi une vision plus détaillée et actualisée que les enquêtes précédentes.

Le dispositif permet également de mieux qualifier les emplois occupés en fournissant des informations sur la nature du contrat, les professions, les catégories sociales, la rémunération et les secteurs d'activité. Ces données, plus précises que celles obtenues par les enquêtes traditionnelles, sont destinées à être publiées sur diverses plateformes d'orientation, telles que Parcoursup, Mon Master, et en open data.

Pour mener à bien ce projet, le Sies a mis en place une structure dédiée en avril 2022. Initialement composée de deux personnes, un directeur de projet et un chef de projet statistique, l'équipe a bénéficié de moyens financiers issus des fonds de transformation de l'action publique pour développer le système d'information. Face à l'avancement de la deadline, l'équipe a été renforcée par trois data scientists pour assurer une mise en production à la fin 2023. Désormais, le système produit ces données en routine deux fois par an, à l'instar d'InserJeunes.

Les grands principes d'InserSup s'appuient sur l'expérience de la DEPP et de la Dares acquise lors de la mise en place d'InserJeunes. Le développement d'InserSup a été grandement facilité par l'implémentation du CSNS de l'Insee, qui a permis d'éviter le développement d'une méthode d'appariement sur les traits d'identité. Parmi ses premiers utilisateurs, le Sies a contribué au paramétrage du CSNS, ce qui a également aidé à calibrer InserSup.

Le CSNS a permis un rapprochement plus efficace des fichiers du monde étudiant. Le système d'information de suivi des étudiants (SISE) offre une vue exhaustive des trois millions d'étudiants. Il fournit des informations détaillées sur :

- les établissements d'inscription (publics, privés, universités, écoles) ;
- les formations suivies et leur niveau ;
- les caractéristiques d'état civil via l'identifiant national étudiant (INE).

Ces données sont complétées par des informations provenant de :

- la Depp, pour les formations en apprentissage (SIFA) ;
- la base BP-BAC pour l'enseignement supérieur dans les établissements secondaires (BTS et classes préparatoires).

Le système intègre également des données du monde du travail issues :

- des fichiers de la DSN de la Dares sur les mouvements de main-d'œuvre ;
- des bases Tous salariés de l'Insee pour les données de rémunération ;
- de la base des non-salariés (en cours de test) ;
- du système d'information de l'apprentissage.

Ce dispositif présente plusieurs avantages par rapport aux enquêtes traditionnelles :

- une exhaustivité des données sur tous les étudiants ;
- des informations individuelles plus précises et moins sujettes à erreur ;
- des points de mesure multiples (juin et décembre) ;
- une réduction de la charge de travail et des coûts pour les universités ;
- une mise à disposition plus rapide des données.

Cependant, le système comporte certaines limites. En effet, les fichiers administratifs français ne couvrent que l'emploi et les études en France, excluant ainsi les situations à l'étranger. Pour pallier ce manque, des enquêtes universitaires ont été maintenues pendant un certain temps afin de calibrer les données et obtenir des informations sur ces aspects.

Pierrette SCHUHL détaille ensuite le processus de traitement des données. Le système utilise des fichiers provenant de l'Insee, de la Dares et de la Depp. La Dares fournit spécifiquement des données sur les salaires des diplômés insérés dans le monde du travail. Le processus se déroule comme suit :

- identification des diplômés via les résultats de SISE ;
- vérification de la non-réinscription l'année suivante ;
- appariement des données via le CSNS ;
- obtention d'informations sur l'emploi et ses caractéristiques.

Elle souligne l'importance du CSNS pour créer un identifiant commun entre l'INE et le NIR.

Actuellement, le système produit des indicateurs pour 6 900 formations et établissement. Le champ d'études s'est progressivement élargi : d'abord limité aux licences professionnelles et aux masters d'universités, il a été étendu aux instituts universitaires privés (instituts catholiques) puis aux licences générales. Depuis 2024, les écoles d'ingénieurs et de management ont été incluses (diplôme bac+5). **Pierrette SCHUHL** mentionne que la publication des niveaux de rémunération a suscité des réactions, notamment au sein des conférences des grandes écoles, en raison de positionnements différents de certains établissements. Les indicateurs produits comprenaient :

- le taux d'emploi salarié en France à 6, 12, 18, 24 et 30 mois ;
- des déclinaisons par genre et nationalité (Française et population totale) ;
- des données sur l'insertion professionnelle des étudiants ayant ou non obtenu leur diplôme.

Ces informations sont diffusées sur les différentes plateformes d'orientation (Parcoursup, Mon Master, Onisep) et par d'autres acteurs qui s'en emparent à partir de l'open data.

Pierrette SCHUHL présente ensuite les évolutions et perspectives d'InserSup. Des indicateurs seront déclinés selon le régime de scolarité, distinguant l'apprentissage. De nouvelles formations seront ajoutées, notamment les bachelors universitaires de technologie (BUT) et des diplômes du domaine culturel, couvrant ainsi ces formations d'ici la fin de l'année.

Une évolution majeure en décembre consistera en l'intégration de l'emploi non-salarié, grâce à la base de données de l'Insee. Cette avancée permettra d'obtenir un taux d'emploi total en France, ouvrant la voie à une évolution des enquêtes d'insertion professionnelle.

À partir du second semestre, des fichiers détaillés seront mis à disposition des chercheurs via le CASD, répondant ainsi à une attente de la communauté scientifique.

En 2026, l'extension des diplômes couverts se poursuivra, avec une attention particulière portée au doctorat. En effet, une part importante des doctorants, environ la moitié, sont étrangers et s'insèrent fortement à l'étranger. Même parmi les doctorants français, l'insertion à l'étranger est plus fréquente, bien que 60 % des doctorants étrangers s'insèrent en France.

Les données seront également déclinées par secteurs d'activité principaux et par professions et catégories socioprofessionnelles. Une étude ultérieure examinera si Résil peut améliorer l'identification des étudiants ne résidant pas réellement en France.

Pierrette SCHUHL souligne que les objectifs initiaux ont été atteints, malgré une montée en puissance continue. Ces objectifs comprennent l'information des jeunes et de leurs familles pour les choix d'orientation, ainsi que la présentation d'indicateurs au niveau le plus fin autorisé. Un seuil de 20 sortants par formation a été établi pour la diffusion des indicateurs, assurant ainsi la stabilité des taux d'évolution.

Le Ministère utilise désormais ces indicateurs d'insertion professionnelle pour le pilotage de l'offre de formation, en les croisant avec les taux de réussite et de poursuite d'études. Pour les licences, l'accent est mis sur les taux de poursuite d'études, la majorité des étudiants poursuivant leur formation.

La production de ces données repose sur une coopération exemplaire au sein du service statistique public, impliquant de nombreux fournisseurs de données et des travaux internes à la Depp, à la Dares, à la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees), à l'Insee et au Sies.

Les perspectives d'études sont multiples, tant pour le service statistique public que pour les chercheurs. Des fichiers de données individuelles seront mis à disposition, permettant d'étudier l'adéquation entre formation et emploi, l'efficience des formations, les liens entre parcours étudiants et insertion, ainsi que l'impact des jobs étudiants sur la réussite universitaire.

Nagui BÉCHICHI, doctorant en fin de thèse à l'École d'économie de Paris, se présente comme **cofondateur du site Suptracker**. Il expose l'importance de lier l'enseignement supérieur à l'insertion professionnelle, en mettant l'accent sur deux aspects principaux : l'orientation et les perspectives de recherche.

Concernant les sources d'information sur le lien entre études et insertion professionnelle, plusieurs enquêtes déclaratives existent, bien que certaines soient devenues obsolètes avec l'arrivée d'InserJeunes et d'InserSup. Ces dispositifs incluent les enquêtes IVA et IPA de la Depp), ainsi que l'enquête IP du Sies, maintenue en complément d'InserSup. D'autres dispositifs, tels que l'enquête Génération du Céreq et l'enquête Insertion de la Conférence des grandes écoles, ne sont pas assez exhaustifs.

L'importance de lier l'enseignement supérieur à l'insertion professionnelle est soulignée pour deux raisons principales. Concernant l'orientation, des recherches montrent que le manque d'informations fiables sur le rendement des études entraîne des conséquences négatives sur les choix d'orientation, particulièrement pour les élèves d'origine sociale défavorisée. La mise à disposition d'informations sur les perspectives d'insertion professionnelle est donc cruciale. Concernant les perspectives de recherche, l'appariement entre les données sur les études et l'insertion professionnelle ouvre de nombreuses possibilités pour la recherche et l'évaluation des politiques publiques.

Le contexte actuel de l'orientation en France est caractérisé par une grande diversité de formations. Parcoursup propose plus de 24 000 formations en 2024, tandis que la plateforme Mon Master compte déjà 8 100 parcours différents en 2025. Cette granularité rend nécessaire la fourniture de repères aux familles pour les aider dans leurs choix.

Le stress lié à l'orientation est particulièrement marqué en France. Dans ce contexte, certaines formations privées, de qualité variable, utilisent parfois un marketing agressif pour attirer les étudiants, sans garantie sur les perspectives d'emploi réelles.

La démocratisation des études supérieures a déplacé les différences sociales vers le choix des filières et des formations. C'est précisément à ce niveau qu'il faut fournir des repères pour améliorer l'égalité des chances dans l'enseignement supérieur.

Du point de vue de la recherche, le système d'information français sur les parcours des élèves est extrêmement complet, peut-être même le plus exhaustif au monde. L'enjeu est de tirer profit de ces données en les combinant avec celles sur l'insertion professionnelle.

Parmi les questions de recherche potentielles, **Nagui BÉCHICHI** mentionne :

- la rentabilité des diplômes professionnels par rapport à une entrée directe sur le marché du travail après le baccalauréat ;
- la valeur ajoutée d'une licence professionnelle après un BTS ;
- l'impact réel des classes préparatoires au-delà de l'effet de sélection ;
- le coût de l'orientation, notamment en termes de réorientation ;
- les reprises d'études.

Ces questions soulignent l'importance des données sur l'insertion professionnelle pour éclairer les choix d'orientation et évaluer l'efficacité des différents parcours de formation.

Concernant les développements envisageables, la couverture des projets est en cours d'amélioration. Le taux de recouvrement devrait progressivement s'accroître, ce qui sera bénéfique pour le perfectionnement d'InserSup. En matière d'indicateurs complémentaires, **Nagui BÉCHICHI** évoque la possibilité d'intégrer des indicateurs de valeur ajoutée, similaires à ceux présents dans InserJeunes. Par exemple, il s'interroge sur la faisabilité d'une logique de valeur ajoutée à l'entrée dans le supérieur pour des formations de type licence ou BTS, tout en reconnaissant la complexité analytique que cela pourrait impliquer.

Ensuite, il présente SupTracker, plateforme gratuite et transparente d'aide à l'orientation, développée bénévolement avec Antoine Prévotat. Cette plateforme s'appuie sur des données open data officielles, dont InserSup, et s'adresse aux lycéens, aux candidats à l'orientation, ainsi qu'à toutes les personnes intéressées par ces informations.

Le projet utilise actuellement 28 sources de données open data, avec un appariement relativement simple basé sur des codes d'établissement, de formation ou de diplôme. La plateforme a déjà attiré plus de 200 000 utilisateurs uniques, démontrant une forte demande pour ce type d'outil.

Nagui BÉCHICHI illustre ensuite le fonctionnement de la plateforme en présentant des exemples de visualisation de données, notamment l'évolution des candidatures Parcoursup depuis 2018. L'objectif est de rendre ces informations facilement accessibles et compréhensibles pour le grand public.

Il mentionne également l'intégration des données d'InserSup et InserJeunes dans les pages Parcoursup pour les BTS et dans la plateforme Mon Master. Ces informations permettent aux étudiants de visualiser le devenir des diplômés et les taux d'emploi à 6 et 12 mois après l'obtention du diplôme.

Stéphane JUGNOT rappelle que la charte des SSM disponible sur le site de l'Autorité de la statistique publique dénombre deux missions : produire des statistiques publiques et appuyer la mise en œuvre des politiques publiques. Il se demande donc si le CSNS peut être utilisé pour produire des indicateurs relevant davantage du pilotage des politiques publiques que de la production de statistiques publiques dans le cadre de la loi de 1951.

Concernant SupTracker, **Christel COLIN** souligne l'importance de la mise en œuvre du projet, qui mobilise de nombreuses informations. Elle s'interroge sur la stratégie adoptée pour la mise à jour future dles données, notamment pour obtenir des courbes d'évolution à 6 mois, 18 mois, etc.

Nagui BÉCHICHI reconnaît ne pas avoir de réponse définitive. Il explique que le nombre croissant de sources intégrées a complexifié le processus de mise à jour. Cependant, l'avantage est que les mises à jour se font souvent au même moment, généralement à l'approche de la période Parcoursup.

Face à la charge de travail croissante, il envisage la création d'une association pour obtenir de l'aide. Initialement, l'idée était de rendre le projet open source pour permettre des contributions externes. Toutefois, constatant que l'orientation est devenue un marché privé lucratif, l'équipe a décidé de garder le contrôle du projet pour le moment.

Nagui BÉCHICHI mentionne également des échanges avec le Cnis et exprime le souhait de pérenniser le projet.

Corinne PROST répond à la question sur la charte des SSM. Un travail de mise à jour sera prochainement lancé pour rendre cette charte plus claire et la mettre à jour sur les nouveaux enjeux. Selon son analyse, InserSup et InserJeunes s'inscrivent pleinement dans le travail de statistique publique, avec des résultats diffusés. Ces projets ne relèvent pas d'un travail interne ministériel non diffusé, exclu du champ de la charte.

Pour illustrer que ce type d'indicateurs ne concerne pas uniquement les SSM et ne répond pas seulement aux demandes gouvernementales, les indicateurs des objectifs de développement durable constituent un exemple pertinent. Ces indicateurs définis par l'ONU, comportant des cibles qui s'adressent aux gouvernements mondiaux, sont développés par les SSM et l'Insee, diffusés sur leur site, et servent à évaluer les politiques publiques dans une perspective élargie. Ce parallèle démontre que de nombreuses statistiques contribuent au pilotage global des politiques publiques.

Stéphane JUGNOT précise sa question : il souhaite savoir si le CSNS peut être utilisé par un SSM pour produire des indicateurs de pilotage de politique publique.

Corinne PROST confirme que cette question sera abordée lors de la table ronde sur le cadre de référence. Le CSNS peut effectivement être utilisé pour des indicateurs ponctuels.

Josy DUSSART relaye une question posée par un intervenant à distance, relative aux appariements InserSup : comment déterminer aujourd'hui si les jeunes obtiennent des emplois correspondant à leur formation, particulièrement ceux visés dans les fiches RNCP qui mentionnent les secteurs professionnels ?

Pierrette SCHUHL explique que des travaux spécifiques sur la corrélation emplois-formation seront menés ultérieurement, constituant ainsi un complément à venir aux données actuelles. L'approche envisagée ne visera pas nécessairement à établir une correspondance directe avec les fiches RNCP. L'analyse portera plutôt sur les objectifs des formations et l'insertion effective des jeunes dans les métiers décrits par les fiches des formations étudiées.

Christine LAGARENNE introduit Olivier Lefebvre, maître d'ouvrage du projet Résil de l'Insee.

ACTUALITÉ DU RÉPERTOIRE STATISTIQUE DES INDIVIDUS ET DES LOGEMENTS

Un document est projeté à l'ensemble des participants.

Olivier LEFEBVRE présente l'état d'avancement du projet Résil. Il propose d'exposer le chemin parcouru depuis sa première présentation lors de la rencontre de 2022, en rappelant brièvement le contexte et les enjeux, puis en détaillant la méthodologie employée et les résultats à venir.

Cette présentation abordera tant l'outillage développé que la réflexion menée au sein du Cnis dans le cadre du groupe de concertation. Les aspects relatifs à la construction, l'utilisation et la communication autour de ce projet représentent une dimension essentielle pour une initiative de cette nature.

Le programme Résil s'inscrit dans un contexte marqué par un choc conjoncturel : la disparition de la taxe d'habitation qui servait au recensement, aux contours des ménages pour les calculs de niveau de vie, et alimentait également la base de sondage. Il fallait donc construire un système alternatif offrant les mêmes services, mais construit de manière composite en mobilisant différentes sources pour éviter une dépendance excessive à une source unique. Ce choc conjoncturel s'accompagne d'une tendance de fond vers l'utilisation accrue de données administratives à apparier, d'où l'opportunité de moderniser les outils de valorisation de ces données.

Concernant les finalités de Résil, **Olivier LEFEBVRE** évoque le décret en Conseil d'État fondant ce traitement : Résil constitue un outil au service de la statistique publique pour remplir ses missions de contribution au débat public et d'aide à l'élaboration et l'évaluation des politiques publiques. Ce dispositif renforce la capacité de l'Insee et des SSM en s'appuyant sur un répertoire d'individus et de logements et sur des possibilités accrues d'appariement de données.

Le projet permettra d'aller plus loin que les dispositifs actuels, avec le CSNS, en facilitant davantage les appariements et en apportant des éléments supplémentaires sur leur fiabilité. Un point essentiel réside dans la capacité de Résil à réaliser des appariements sur mesure avec sélection précise du champ et des variables, appliquant ainsi le principe de minimisation et évitant la constitution de mégabases au profit de fichiers ciblés répondant à des besoins spécifiques.

Sur le plan du calendrier, le projet arrive à son terme après avoir comporté deux grandes dimensions : d'une part, la réflexion sur l'acceptation de Résil, incluant la concertation et la construction de l'environnement juridique (avec avis favorable de la CNIL et parution du décret en Conseil d'État fin 2023) ; d'autre part, l'ingénierie informatique et statistique classique, avec ses phases d'exploration et de développement. Actuellement en phase d'initialisation du répertoire, les premières utilisations effectives pour la production statistique s'échelonneront entre fin 2025 et le premier semestre 2026.

Trois grandes productions sont attendues de Résil : un répertoire statistique des individus et des logements à finalité exclusivement statistique ; une photo annuelle au 1er janvier servant de socle à de nombreuses productions statistiques ; et enfin un service de production de fichiers enrichis par appariement, offrant des fonctionnalités étendues par rapport au code statistique non signifiant actuel.

Olivier LEFEBVRE présente ensuite en détail les différentes composantes du système en cours de construction. Tout d'abord, il aborde le répertoire statistique des individus et des logements. Ce répertoire se compose de deux bases de données principales, alimentées par diverses sources. Pour les individus :

- le répertoire d'identification des personnes physiques, fournissant des informations sur les naissances, les décès et les changements d'identité ;
- des sources fiscales, incluant l'occupation des biens immobiliers (le nouveau dispositif Gérer mes biens immobiliers ou GMBI) et les déclarations de revenus ;
- des sources sociales, telles que la DSN (Déclaration sociale nominative) et le PASRAU (Passage des revenus autres) ;
- des données sur les prestations familiales et sociales servies par la CAF et la MSA ;
- des inscriptions dans l'enseignement supérieur, améliorant la couverture des étudiants.

Pour les logements :

- les bases foncières du cadastre sur le parc de logement ;
- le répertoire des communautés, utilisé pour le recensement de la population.

Ces sources permettent d'améliorer la couverture du parc de logement et de mieux distinguer les personnes vivant en logement ordinaire de celles vivant en communauté ou partageant leur temps entre ces deux situations.

Un aspect crucial du système Résil est la sélection rigoureuse des données. Seules les informations essentielles à l'alimentation du répertoire sont conservées, notamment la présence d'une personne ou d'un logement dans une source administrative, la période concernée et l'adresse. Cette approche permet de déterminer si une personne réside en France ou non.

Le répertoire contient principalement des données d'identification :

- pour éviter les doublons et faciliter les mises à jour, pour les individus comme pour les logements ;
- des adresses et des liens entre individus et logements ;
- d'autres variables, telles que la probabilité de présence des personnes sur le territoire national ou des variables de gestion.

Il est important de noter que le répertoire ne contient pas le NIR ni d'informations caractérisant les personnes telles que les revenus, les métiers, les catégories sociales ou les employeurs. Ces données sont stockées dans des systèmes d'information spécialisés et ne sont utilisées que sur demande pour des études

ponctuelles. **Olivier LEFEBVRE** souligne que Résil est conçu comme un « squelette » plutôt qu'une base de données multi-thèmes.

Les sources alimentant Résil sont précisées dans un arrêté, publié en même temps que le décret en Conseil d'État, et pris après avis de la Cnil et du Cnis. Toute modification de cette liste nécessite d'ailleurs leur avis ainsi qu'un nouvel arrêté, garantissant ainsi la transparence du processus.

Le répertoire est mis à jour en continu, au fur et à mesure de l'arrivée des données des différentes sources. L'univers de référence est une photo du répertoire donnant la situation au 01/01 de chaque année. Il se compose de trois listes :

- les individus présents sur le territoire français ;
- les logements ;
- les ménages (individus vivant dans le même logement).

Cet univers de référence sert de base pour la création d'échantillons, la fourniture d'éléments nécessaires à la préparation et à l'exploitation du recensement de la population, notamment le calcul des populations de référence, et d'autres usages statistiques. Les principales difficultés dans la création de cet univers de référence incluent :

- la sélection précise des résidents (ajustement du modèle de signes de vie, élimination des doublons) ;
- la détermination de l'adresse principale des individus.

La qualité de cet univers de référence est vérifiée en la comparant à des sources externes, telles que les enquêtes annuelles de recensement.

Le service d'enrichissement permet de produire par appariement des fichiers multi-sources, mixant plusieurs sources administratives, ou une source administrative et une enquête…. Ce service est réservé aux membres du service statistique public pour des finalités de statistique publique uniquement. Les utilisateurs doivent se conformer au RGPD et au cadre de référence sur les appariements.

Le processus de demande auprès du service Résil se déroule comme suit : le demandeur spécifie la population d'intérêt, les variables nécessaires et les éventuels descripteurs géographiques. Le gestionnaire Résil fabrique ensuite le fichier demandé, après s'être assuré du respect des obligations liées au RGPD et de la conformité au cadre de référence pour les appariements. Ce fichier sera ensuite utilisé par le demandeur dans son processus de production ou d'études.

Le guichet Résil reprendra les attributions de l'équipe en charge du CSNS, centralisant ainsi les finalités statistiques. L'identification des individus dans les fichiers pourra être plus précise que celle du CSNS, s'appuyant notamment sur l'indicateur d'adresse des personnes.

Ce service permettra également de construire des fichiers s'alimentant à différentes sources d'informations sans que le demandeur ait à contacter chaque producteur individuellement. De plus, il réduira le champ des appariements à la population résidente, ce qui facilitera l'étude des trajectoires et l'analyse du non-recours (car actuellement, si on ne retrouve pas une personne dans un fichier d'allocataires on ne sait pas si c'est dû à un non-recours ou à une sortie du territoire).

Concernant les chercheurs, bien qu'ils n'aient pas d'accès direct au guichet Résil, ils pourront bénéficier de ses apports. D'une part, dans le cadre de partenariats avec des entités du service statistique public, et d'autre part, grâce à l'enrichissement de la production statistique accessible via le Comité du secret statistique.

La concertation autour de Résil a débuté dès 2022, reconnaissant l'importance de confronter les opinions et convictions à divers points de vue, notamment juridiques et éthiques. Un groupe de concertation très ouvert a été créé, présidé par Jean-Marie Delarue, regroupant 15 personnes aux profils variés.

Les conclusions de ce groupe incluent l'approbation du principe du répertoire, la confiance dans la capacité du service statistique public, mais aussi la nécessité d'encadrer strictement la construction et l'usage de Résil. Des règles de fonctionnement ont été posées, s'appuyant sur des regards extérieurs, tels que le Cnis, l'Autorité de la statistique publique, l'ANSSI et la CNIL.

La transparence est un élément clé, avec des actions de communication prévues et la création d'un registre des appariements. Des garanties sont inscrites dans le décret, notamment sur les finalités exclusivement statistiques et les destinataires limités aux agents du service statistique public.

En conclusion, Résil offre de nouvelles possibilités de fabrication de données, mais implique également de nouvelles responsabilités éthiques, juridiques et techniques. **Olivier LEFEBVRE** souligne l'importance de rester vigilant sur la qualité et l'adéquation des fichiers construits, ainsi que sur la sécurisation du système d'information.

Bernard SUJOBERT, CGT, soulève deux points importants. Concernant la disparition de la taxe d'habitation, il rapporte les inquiétudes de ses collègues à la DGFiP. En effet, la base « Gérer mes biens immobiliers » (GMBI) n'a pas apporté les éléments utiles escomptés. La situation est jugée préoccupante, voire qualifiée de fiasco par certains. Face à ce constat, une amélioration est vivement souhaitée.

Par ailleurs, il évoque une récente heure mensuelle d'information organisée par des syndicats de l'Insee, en mars, sur Résil. Lors des discussions, une forte inquiétude s'est manifestée à son sujet. Des interrogations persistent quant à son utilisation dans le cadre d'un pouvoir politique autoritaire, et ce, malgré les garanties existantes.

Jean-Christophe SCIBERRAS, Président de la commission Emploi, qualification et revenus du travail du Cnis, soulève également deux questions. La première concerne la notion de communauté, dont la définition semble peu claire, notamment dans le contexte des ménages. Il sollicite des éclaircissements sur ce point.

La seconde question porte sur la notion de confiance, un thème récurrent dans les échanges. Il met en lumière une tension entre deux sources de confiance : la transparence et la confidentialité. La problématique soulevée est de trouver le juste équilibre entre ces deux aspects, contradictoires, dans le cadre des appariements de données.

Mathilde GACHARD, Direction départementale du travail du Bas-Rhin, s'interroge sur la capacité de Résil à mieux identifier les logements vacants. Cette préoccupation découle des difficultés rencontrées lors du recensement de la population, où la distinction entre logements temporairement inoccupés et véritablement vacants n'est pas toujours évidente.

Ensuite, elle demande si, dans un avenir proche, il sera possible d'identifier les logements utilisés pour des locations de courte durée (type Airbnb, Gîtes de France). Cette information est jugée cruciale pour évaluer l'impact de ces locations sur la disponibilité générale des logements dans certaines zones.

Dominique MEURS, professeure à l'Université Paris Nanterre, s'interroge sur la possibilité pour les chercheurs de proposer des initiatives en termes d'appariement. Cette question est particulièrement pertinente dans le contexte des recherches actuelles sur la mobilité géographique résidentielle, un domaine où les données font défaut.

Jan Rober SUESSER souligne l'implication de la Ligue des droits de l'Homme dans le développement de Résil, via la participation active de la LDH à la concertation, saluant la collaboration de l'Insee avec diverses communautés d'utilisateurs comme de vigies en matière de valeurs et d'éthique.

Il aborde ensuite la question de la protection des outils statistiques face à d'éventuels changements politiques. Il note que les questions posées par Jean-Marie Delarue dans l'introduction du rapport sur Résil n'ont pas reçu de réponses, au moins accessibles publiquement. Cette réflexion sur les garanties s'inscrit dans un contexte plus large de questionnements sur la solidité des protections qui impliquent de dépasser le cadre des bonnes pratiques habituelles. Même si des réponses satisfaisantes sont compliquées à établir, le débat sur ces réponses doit être poursuivi.

Pour la mise en place de Résil, il questionne l'adéquation des dispositifs actuels. Il suggère l'introduction de compétences juridiques et en droits humains lors de l'instruction de demandes ambitieuses impliquant des variables dont on sait qu'elles ont une sensibilité dans le débat public, alors qu'elles vont pouvoir être rapprochées de nombreuses autres données. L'objectif serait de renforcer les protections en vue des bonnes pratiques et d'améliorer la capacité collective à faire face aux défis éthiques actuels, considérant que la transparence seule et l'action a postériori pourraient ne pas être suffisantes dans le contexte actuel. Il demande qu'un tel dispositif soit introduit au moins pour une période transitoire de deux à trois ans, et qu'un bilan soit fait sur cette base.

Abdessattar SAOUDI, Santé publique France, s'enquiert de l'existence d'un lien entre Fideli et Résil. Il demande si, à terme, le second remplacera le premier, ou si ces bases de données cohabiteront.

Concernant GMBI, **Olivier LEFEBVRE** reconnaît les difficultés mentionnées, également relevées par la Cour des comptes. Des échanges avec la DGFiP ont eu lieu sur la fiabilisation de l'information. Bien que des progrès aient été réalisés en 2024, tous les propriétaires n'ont pas encore renseigné GMBI pour l'ensemble des occupants. Cette situation renforce l'importance pour Résil de s'appuyer sur plusieurs sources. Par exemple, l'utilisation conjointe des données fiscales et des caisses d'allocations familiales permet une vision plus complète de la composition des ménages. Malgré les imperfections potentielles, cette approche multisource représente la meilleure solution possible dans le cadre de Résil.

Concernant les logements vacants et les locations de courte durée, leur identification repose sur les déclarations des propriétaires, complétées par des informations sur la localisation des personnes issues d'autres sources. Cette méthode permet de caractériser les résidences principales, secondaires et les logements vacants. Cependant, la mobilité croissante des personnes rend cet exercice de plus en plus complexe, comme en témoignent les difficultés rencontrées lors du recensement et dans les sources administratives. Résil devrait permettre une meilleure appréhension de ces réalités mouvantes.

Pour les communautés, **Olivier LEFEBVRE** précise qu'il s'agit, au sens du recensement, de personnes partageant un lieu de vie commun sous l'autorité d'une même personne gestionnaire. Cela inclut notamment les Ehpad, les foyers de travailleurs, les résidences universitaires, les établissements pénitentiaires et les communautés religieuses. Ces formes d'hébergement ne sont pas toujours identifiées comme telles dans les fichiers du foncier, d'où l'intérêt de Résil pour mieux appréhender ces différentes formes de logement.

En ce qui concerne la transparence, la confidentialité et la confiance, le **Olivier LEFEBVRE** souligne l'importance de maintenir la confiance à deux niveaux : celle des utilisateurs de la statistique quant à sa pertinence et son utilité, et celle des citoyens et entreprises fournissant leurs données. Cette confiance repose sur des dispositions techniques assurant la qualité et la confidentialité des données, ainsi que sur la transparence concernant leur utilisation et la construction des statistiques. Un effort continu de pédagogie est nécessaire pour entretenir cette confiance.

Concernant Fideli, **Olivier LEFEBVRE** explique que Résil ne le remplacera pas directement, mais produira ou aidera à produire les services actuellement rendus par Fideli. Cela inclut la création de bases de sondage, la définition des contours des ménages, le calcul des niveaux de vie, et la production de fichiers d'études. Ces services seront intégrés dans l'écosystème construit autour de Résil.

Olivier LEFEBVRE aborde ensuite l'initiative des appariements et, plus largement, de la production de données statistiques. Il souligne que cette production répond aux besoins exprimés par les utilisateurs, notamment via le Cnis. Résil permettra de répondre à ces besoins de manière plus flexible, en facilitant la création de fichiers d'analyse sur des sujets spécifiques, comme l'immigration résidentielle. Cependant, pour des raisons de charge de travail, il n'est pas envisagé que des chercheurs isolés puissent directement solliciter la fabrication de données via Résil.

Enfin, concernant les risques liés à une potentielle sortie de l'Etat de droit ou à des pressions pour modifier l'utilisation des outils statistiques, **Olivier LEFEBVRE** rappelle que l'environnement juridique actuel protège contre les utilisations non conformes aux finalités statistiques. Il estime que, si ce cadre juridique venait à être remis en cause, cela nécessiterait une réflexion globale sur les pratiques de la statistique publique, au-delà de Résil. Il souligne que cette base, en tant que répertoire statistique sous la responsabilité du directeur général de l'Insee, offre une plus grande flexibilité pour en arrêter l'utilisation si les conditions ne permettaient plus son usage normal.

La séance est suspendue de 13 heures 10 à 14 heures 05.

LES APPORTS D'UN CADRE DE RÉFÉRENCE POUR LA RÉALISATION D'APPARIEMENTS

Bertrand du MARAIS, Président du Cnis, ouvre la session de l'après-midi en remerciant les participants pour leur présence à cette table ronde consacrée au cadre de référence pour la réalisation d'appariements. Cette rencontre du Cnis vise non seulement à faire le point sur des travaux d'appariements en cours, comme il en a été présenté le matin, mais également à développer et systématiser la pratique des appariements.

La technique des appariements enrichit les outils traditionnels de la statistique publique en combinant les enquêtes et les données d'échantillons avec des données administratives issues d'administrations publiques, d'entités chargées d'un service public, voire d'entreprises privées.

L'importance de cette pratique est particulièrement marquée en France, où les services publics sont très développés et centralisés. En effet, la France dispose d'un gisement considérable de données administratives, ce qui explique en partie son bon classement et sa reconnaissance en matière de politique d'open data. Ce potentiel représente un enjeu majeur, notamment pour l'évaluation des politiques publiques.

Cependant, la mise en œuvre de ces appariements doit se faire dans le respect d'un cadre juridique et déontologique strict. Il ne s'agit pas seulement par extension de leur étendre les règles et principes de la statistique publique, mais aussi de répondre aux inquiétudes, voire aux fantasmes, que peut susciter la mise en relation de traitements de données issues du secteur public. Le service de la statistique publique et le Cnis ont pour mission de réduire ces inquiétudes par la transparence et l'information.

Le cadre de référence pour les appariements, un document concis de quatre pages disponible sur le site du Cnis, a été présenté au bureau du Cnis le 12 mars 2025. Des commentaires ont été sollicités, et la CGT a fourni une contribution détaillée qui sera prise en compte pour une nouvelle version à présenter lors du prochain bureau du Cnis le 4 juin 2025.

La table ronde se déroulera en deux parties. Dans un premier temps, Corinne Prost, directrice de la méthodologie et de la coordination statistique et internationale à l'Insee, présentera le cadre de référence et les enjeux de coordination et de transparence. Ensuite, Anthony Guérout, maire de Saint-Aubin-Routot en Seine-Maritime et représentant de l'Association des maires de France (AMF) au Cnis, abordera les besoins des collectivités et les préoccupations des citoyens.

Dans un second temps, trois autres interventions sont prévues. Christelle Minodier, cheffe de service au sein de la Drees, partagera le point de vue des SSM sur l'utilité et l'utilisation du cadre de référence. François Clanché, directeur de l'Ined, apportera la perspective d'un utilisateur et producteur d'appariements. Enfin, Dominique Meurs, professeure d'économie à l'université Paris Nanterre et chercheuse associée à l'Ined, présentera le point de vue des chercheurs et évoquera potentiellement des comparaisons internationales.

Corinne PROST, directrice de la méthodologie et de la coordination statistique et internationale de l'Insee, présente le contenu du cadre de référence pour les appariements. Ce document complète le cadre déontologique existant en se focalisant spécifiquement sur les appariements et en définissant des rôles et des procédures, notamment au sein du Cnis. Le cadre de référence vise à rappeler et à rendre visibles les principes guidant les pratiques d'appariement, ainsi qu'à assurer leur mise en œuvre par le service statistique public. Il se compose de plusieurs parties :

- Un préambule qui contextualise les enjeux et définit les appariements.
- Une introduction expliquant la volonté du service statistique public de se doter d'un mandat social sur les enjeux d'appariement, allant au-delà des exigences déontologiques habituelles.
- Une première partie détaillant les engagements du service statistique public pour la pratique des appariements, rappelant les principes de nécessité, de proportionnalité, de minimisation et de transparence.
- Une dernière partie expliquant le rôle du Cnis dans la mise en œuvre de ce cadre, répondant ainsi à une demande du groupe de concertation Résil.

Le document précise les différents cas de figure pour lesquels des avis seront émis par le Cnis, établissant un parallèle avec la procédure existante pour les enquêtes. Par exemple, le premier cas concerne les enquêtes enrichies par des variables issues de données administratives. Les enrichissements prévus par appariement seront mentionnés lors de la présentation de l'avis d'opportunité en commission.

Dans le cadre de l'article 7bis, l'appariement prévu sera explicité et discuté lors de la présentation pour l'accès aux données administratives.

Pour les opérations où l'appariement n'est pas prévu initialement, un document de présentation et une consultation électronique seront mis en place. Les autorisations pour les opérations récurrentes seront valables plusieurs années. Les appariements à titre exploratoire, par exemple pour des fins méthodologiques, ne nécessiteront qu'une simple information au Cnis.

En cas de projets d'ampleur ou de sensibilité particulière, un débat approfondi en commission pourra être demandé par le Bureau, un président de commission ou un responsable de traitement. Enfin, les suivis statistiques de cohortes internes à un ministère ne nécessiteront pas d'avis spécifique du Cnis, mais seront généralement expliqués lors des commissions thématiques concernées.

Bertrand du MARAIS remercie Corinne Prost et donne la parole à Anthony Guérout, au sujet du point de vue des collectivités territoriales et des citoyens sur la notion d'appariement.

Anthony GUÉROUT, AMF, souligne l'importance des données dans les collectivités territoriales. Ces dernières en produisent et en utilisent beaucoup pour élaborer leurs politiques publiques. Les collectivités ont besoin de données riches, fiables, précises et à jour.

Parallèlement, il insiste sur l'importance de protéger les libertés individuelles. Les collectivités croisent régulièrement des données pour leurs politiques publiques, par exemple pour choisir l'emplacement d'équipements ou planifier des réseaux de transport.

Anthony GUÉROUT illustre l'utilité des données par un exemple concret : dans le cadre de son mandat de conseiller délégué à l'habitat au sein de la communauté urbaine Le Havre Seine Métropole, des données de thermographie et des compteurs Linky sont utilisés pour repérer les ménages en précarité énergétique. Cela permet de leur proposer des aides pour améliorer leur logement. Ce dispositif d'aide est doté d'un budget de 15 millions d'euros.

Anthony GUÉROUT aborde ensuite les inquiétudes des citoyens concernant la collecte et l'utilisation des données. Lors des recensements, certains habitants expriment leur méfiance quant à l'utilisation de leurs informations personnelles. Il souligne que, contrairement aux idées reçues, les collectivités n'ont pas accès à des listes exhaustives de leurs habitants.

La crainte des citoyens porte également sur les possibles mésusages des données, notamment en cas d'appariements ou de croisements. **Anthony GUÉROUT** reconnaît que ces préoccupations sont légitimes, car, entre de mauvaises mains, ces données pourraient être utilisées à des fins malveillantes.

Pour répondre à ces inquiétudes, il estime qu'un cadre garantissant la transparence est nécessaire. Cependant, il n'est pas certain que cela soit suffisant à l'avenir. Il suggère de mettre en place des cadres définissant l'usage des appariements et assurant la sécurité informatique.

Anthony GUÉROUT évoque également les défis en matière de cybersécurité, citant l'exemple des attaques subies lors des Jeux olympiques ou lors de visites officielles. Il souligne l'importance de protéger la vie privée et les données personnelles.

En conclusion, il met en garde contre les risques liés à l'utilisation abusive des statistiques, faisant référence à des événements récents aux États-Unis. Il appelle à réfléchir à un cadre permettant d'éviter de telles dérives.

Bertrand du MARAIS ouvre la discussion en sollicitant des réactions aux interventions précédentes. Il souligne l'importance du cadre de référence évoqué par Corinne Prost. Ce cadre attribue au Cnis un rôle crucial notamment en tenant un état des lieux des appariements. L'objectif est de fournir aux citoyens un accès simple à des informations standardisées sur les appariements réalisés, leurs auteurs et les ressources utilisées. La mise à jour régulière de ce registre constitue un élément essentiel de transparence, auquel **Bertrand du MARAIS** accorde une grande importance personnelle.

Anthony GUÉROUT établit un parallèle entre le rôle de référence du Cnis et celui de la Cnil, notamment depuis l'introduction du RGPD. En effet, la Cnil sert de référence commune en matière de protection des libertés individuelles liées aux données. Ainsi, le Cnis pourrait assumer une fonction similaire dans le domaine des appariements et des statistiques publiques.

Bertrand du MARAIS donne la parole à Christelle Minodier, qui présentera le point de vue des SSM.

Christelle MINODIER, Cheffe de service à la Drees, explique que le cadre de référence, bien que concis, est dense et riche en informations. Elle se propose d'en examiner plusieurs aspects qu'elle considère comme des points de vigilance essentiels.

Concernant les apports du cadre de référence pour la réalisation d'un appariement :

- L'existence même du cadre est primordiale. Sa formalisation par écrit facilite l'appropriation des bonnes pratiques.
- Le cadre rappelle les règles légales (loi de 1951, loi de 1978) et réglementaires (règlement 223 au niveau européen, RGPD) applicables.
- Il insiste sur les valeurs fondamentales du service statistique public : pertinence, impartialité, indépendance et transparence.
- Le cadre vise à harmoniser les pratiques au sein du Service statistique public et à les rendre visibles, renforçant ainsi la confiance du public.
- Il prend en compte les spécificités de la statistique publique, notamment en termes de droits et de devoirs.

Christelle MINODIER souligne que des travaux sont en cours pour décliner ce cadre de manière opérationnelle au sein des SSM. Cette démarche implique des échanges inter-SSM, entre les directions juridiques et les délégués à la protection des données, mais également avec l'Insee.

Deux points de vigilance sont particulièrement mis en avant :

- La capacité à arbitrer entre la sélection et la minimisation des données. Il est crucial de déterminer précisément les variables et populations nécessaires pour chaque projet et chaque appariement, sans collecter de données superflues.
- Les modalités pratiques relatives aux infrastructures informatiques, notamment la séparation des données identifiantes des autres variables pour garantir la confidentialité.

Christelle MINODIER insiste sur la transparence, qui passe par l'information du public et des personnes concernées par la collecte de données. Cette transparence se manifeste à travers :

- les avis d'opportunité et les lettres-avis pour les enquêtes ;
- la publication d'informations détaillées sur les sites internet des SSM et de l'Insee ;
- la mise à disposition d'une liste annuelle des appariements sur le site du Cnis ;
- la publication obligatoire de la liste des appariements issus de Résil sur le site internet de l'Insee.

Ces mesures visent à renforcer la confiance du public et à garantir une utilisation éthique et transparente des données statistiques.

Concernant les finalités des appariements de données, deux utilisations principales sont identifiées. La première, qualifiée d'utilisation primaire, concerne la production de statistiques publiques. La seconde, dite utilisation secondaire, implique l'accès des chercheurs à ces données. Cette mise à disposition est cruciale. Certains SSM jouent un rôle d'animation de la recherche, notamment la Dares et la Drees. Il existe une collaboration étroite entre les SSM et les acteurs de la recherche, notamment à travers des appels à projets.

La mise à disposition des données pour les chercheurs se fera, selon le cadre de référence, dans les conditions en vigueur pour l'accès aux fichiers détails produits par la statistique publique. Cette utilisation constitue ce que **Christelle MINODIER** qualifie d'utilisation « secondaire » des appariements. En effet, le résultat d'un appariement crée un nouveau fichier détail produit par la statistique publique, qui peut également être mis à disposition des chercheurs.

Sur la question de la minimisation des données, les SSM et l'ensemble du SSP français auront une responsabilité particulière. Actuellement, la tendance est de mettre à disposition une base complète. Or, avec la multiplication des appariements, les fichiers détails deviennent de plus en plus complets. Il faudra probablement envisager de séparer ces bases en différents blocs pour permettre un accès modulaire et veiller à la minimisation des données mises à disposition. Ce travail prend une importance croissante face aux enjeux actuels.

Christelle MINODIER aborde ensuite la spécificité des données de santé, un sujet particulièrement pertinent pour la Drees. Ces données sont considérées comme sensibles au sens de l'article 9 du RGPD et de l'article 6 de la loi de 1978. Le périmètre des données de santé n'est toutefois pas clairement défini, malgré une définition large prévue par le RGPD. Les frontières juridiques restent mouvantes, ce qui crée des difficultés d'interprétation.

Une donnée considérée isolément peut ne pas être qualifiée de donnée de santé, mais le devenir lorsqu'elle est couplée à d'autres informations. Cette ambiguïté explique pourquoi il n'existe pas de consensus strict entre les positions du Conseil d'État, de la Cnil et d'autres instances, nécessitant ainsi une vigilance accrue.

Par exemple, la collecte du poids et de la taille peut constituer un traitement de données de santé dès lors que ces informations pourraient révéler qu'une personne souffre d'obésité. De même, les questions du module européen (sur l'état de santé, les maladies chroniques, les limitations) peuvent également être qualifiées de données de santé dès lors que ces informations permettent de tirer des conclusions sur l'état de santé d'une personne, et plus précisément sur la gravité d'un problème de santé. En revanche, le handicap, sans précision sur sa nature, peut parfois ne pas être considéré comme une donnée de santé selon la jurisprudence du Conseil d'État. Toutefois, pour l'indicateur GALI dans le recensement, la Cnil a clairement établi qu'il ne s'agissait pas d'une donnée de santé dans ce cadre spécifique.

Les démarches pour le traitement des données de santé varient selon la personne qui les réalise. Les SSM ont le droit de traiter ces données moyennant une information au Cnis et éventuellement une Analyse d'impact relative à la protection des données (AIPD). Pour les chercheurs, en tant que réutilisateurs de données de santé, le processus diffère : ils doivent obtenir une autorisation de la Cnil ou s'engager à respecter une méthodologie de référence. L'autorisation requiert un avis préalable du Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (Cesrees).

Certains traitements peuvent également impliquer la levée du secret médical, qui englobe toute collecte de données réalisée par un professionnel de santé dans un établissement de santé ou médico-social. Pour ces cas, une autorisation spécifique avec avis préalable du Cesrees est nécessaire. L'avis du Cnis pour les appariements est crucial, car il fournit un avis d'opportunité et des arguments justifiant la levée du secret médical ou la réalisation d'appariements.

Christelle MINODIER rappelle que Résil fournit un univers de référence daté et la liste des personnes résidentes en France au 1er janvier d'une année *n*. Le service d'appariement développé dans ce cadre sera essentiel pour les SSM, permettant de calculer des taux, des prévalences et le non-recours avec un dénominateur robuste. Alors que les SSM disposent généralement du numérateur, le dénominateur est souvent moins fiable.

Enfin, pour les appariements avec des données du Système national des données de santé (SNDS), l'avis du Cnis reste un préalable. L'avis du Cesrees et l'autorisation de la Cnil sont également obligatoires, car il n'existe aucune méthodologie de référence pour ces appariements spécifiques.

Bertrand du MARAIS remercie Christelle Minodier et donne la parole à François Clanché pour présenter le point de vue de l'Ined et des chercheurs qui y sont affiliés.

François CLANCHÉ, directeur de l'Ined, rappelle tout d'abord que les données d'appariements ne sont pas uniquement utilisées par les statisticiens publics, mais également de nombreux chercheurs du secteur public.

Les chercheurs de l'Ined travaillent depuis longtemps à partir de l'échantillon démographique permanent (EDP) de l'Insee. Ils ont même développé, à partir de l'EDP, un échantillon démographique des couples permettant d'étudier leur formation et leur séparation. Cette approche a notamment permis de publier une étude démontrant que les probabilités de séparation augmentent significativement lorsque les revenus sont homogènes ou lorsque la femme gagne davantage que son conjoint. Des travaux sur l'impact des séparations sur les ressources des enfants ont également été publiés grâce à ces données.

L'Ined utilise par ailleurs les données du SNDS, qui ne relève pas strictement de la statistique publique, mais s'avère extrêmement précieux pour la recherche. L'Institut exploite même le couplage entre l'EDP et le SNDS, appelé « EDP-Santé », permettant d'établir des liens entre événements démographiques, événements de santé et facteurs économiques.

La recherche publique constitue donc une utilisatrice majeure des appariements réalisés par la statistique publique. Dans cette perspective, le cadre de référence s'avère crucial, d'abord pour réaffirmer les possibilités d'accès de la recherche publique à ces données, ensuite pour en améliorer la transparence. Tout comme les listes d'enquêtes sont disponibles depuis longtemps, disposer de listes d'appariements

représentera un atout important pour les chercheurs découvrant ou approfondissant un domaine, leur permettant d'identifier les données déjà traitées par les services statistiques.

Une question émergera progressivement concernant la localisation et l'accessibilité de ces données : existera-t-il des fichiers de production et de recherche (FPR) directement utilisables par les chercheurs après présentation de leur projet, ou faudra-t-il systématiquement passer par le CASD, avec les contraintes que cela implique ? Si ces questions sont désormais clarifiées pour les enquêtes, elles devront également l'être pour les données d'appariement.

François CLANCHÉ souligne également que l'Ined réalise lui-même des appariements, au-delà de son rôle d'utilisateur. L'Institut mène des enquêtes et effectue ses propres opérations d'appariement, notamment en amont des enquêtes à des fins méthodologiques, pour élaborer ou enrichir des échantillons. Il cite un exemple spécifique d'appariement réalisé en 2023 pour l'enquête sur l'entrée dans la vie affective et sexuelle des jeunes. Cette étude ciblant uniquement les moins de 30 ans s'appuyait sur une génération aléatoire de numéros de téléphone. Pour optimiser le processus, le service des enquêtes de l'Ined a soumis à l'administration une liste de plusieurs centaines de milliers de numéros, demandant d'identifier ceux correspondant à des ménages sans individu de moins de 30 ans. Cette démarche a considérablement réduit les coûts d'enquête et évité de déranger des personnes hors cible, tout en respectant le RGPD.

L'Institut enrichit également ses données d'enquête avec des données administratives. Par exemple, l'enquête Famille-Employeurs sur la conciliation entre vie familiale et professionnelle est actuellement enrichie par des données provenant de l'Insee, concernant tant les revenus et situations des personnes que les entreprises, puisqu'il s'agit d'une enquête à double niveau (individu et entreprise). Cette démarche fonctionne, car elle était prévue dès la conception de l'enquête.

D'autres situations sont moins prévisibles. **François CLANCHÉ** évoque l'étude longitudinale sur l'enfance (Elfe), suivant depuis 2011 une cohorte d'enfants qui seront accompagnés jusqu'à l'âge adulte. Actuellement, l'Ined et l'Institut national de la santé et de la recherche médicale (Inserm) envisagent d'enrichir cette cohorte avec des données administratives de l'Assurance maladie, de l'Éducation nationale pour le suivi des parcours scolaires, et ultérieurement des données sur l'emploi, le travail et les revenus. Cette évolution en cours de route nécessitera l'obtention de multiples autorisations.

La statistique publique n'est donc pas la seule à réaliser des appariements, la recherche publique procède également à ces opérations, dans le respect de l'éthique de la cherche et du RGPD, avec ses propres délégués à la protection des données. Ces acteurs ne doivent pas être oubliés dans les différentes versions du cadre réglementaire.

En conclusion, **François CLANCHÉ** plaide pour une nécessaire souplesse prenant en compte les aspects méthodologiques des appariements en recherche, car l'innovation commence souvent par l'expérimentation méthodologique. Il faudra adapter le cadre à ces démarches exploratoires. Il rappelle également la question des moyens humains, puisque certains outils, comme le CSNS ou Résil ne seront pas directement accessibles aux organismes hors SSP. La recherche publique devra passer par des partenaires, ce qui pose des questions de capacité et de priorisation. Il souhaiterait que la recherche publique ne soit pas systématiquement reléguée en fin de file d'attente. Si de nouveaux horizons s'ouvrent, il conviendra de trouver un juste équilibre entre finalités et moyens, en veillant à ne pas exclure la recherche publique et les opérateurs hors statistique publique de ces dispositifs.

Bertrand du MARAIS, soucieux de la place des chercheurs dans cette problématique, passe justement la parole à Dominique Meurs pour présenter un autre point de vue de la recherche publique.

Dominique MEURS, professeure à l'Université Paris Nanterre, explique s'être tournée vers des doctorants pour mieux comprendre leurs attentes vis-à-vis du cadre de référence.

Les jeunes économistes d'aujourd'hui se soucient principalement de l'accès à des bases de données pour répondre à des questions de recherche très concrètes. Leur approche vise à déterminer l'efficacité des politiques économiques, à identifier d'éventuels effets contraires ou à analyser l'évolution des carrières. La science économique s'est profondément transformée ces 15-20 dernières années, privilégiant désormais le travail approfondi sur les données, le dialogue avec les statisticiens publics et la production de connaissances pragmatiques.

Concernant les pratiques à l'étranger, plusieurs doctorants travaillent sur des bases de données néerlandaises ou scandinaves. Par exemple, une doctorante étudie la délinquance des jeunes Néerlandais

en lien avec une réforme des politiques d'aide sociale ayant réduit les aides aux familles tout en maintenant celles destinées aux jeunes adultes isolés. Sa recherche examine l'impact de ce changement de ressources sur la délinquance juvénile. Les Pays-Bas constituent un terrain de recherche privilégié, car ils disposent de systèmes d'appariement de données exceptionnels, fondés sur des registres de population avec identifiants individuels permettant de connecter différentes bases entre elles. Cette doctorante a ainsi pu obtenir un accès pour suivre les trajectoires des jeunes avant et après la réforme concernée.

Il convient toutefois de préciser que ces pays, comme la France, appliquent des procédures strictes pour protéger le secret statistique. Les chercheurs n'accèdent pas directement aux données brutes, mais travaillent via des systèmes sécurisés, en l'occurrence une clé VPN transformant l'ordinateur en environnement isolé. Ces dispositifs de sécurité élaborés représentent un coût important.

Cette richesse de données explique pourquoi tant d'articles scientifiques portent sur les Pays-Bas, la Suède ou d'autres pays scandinaves. Ces ressources permettent d'analyser les changements générationnels depuis le début du XXe siècle et de répondre à de nombreuses questions de recherche. Cependant, ces pays constituent des exceptions dans le paysage mondial. Au Canada ou en Allemagne, l'accès aux données reste beaucoup plus difficile qu'en France, avec une qualité souvent inférieure.

Malgré ses atouts, la France présente deux problèmes majeurs pour les jeunes chercheurs en matière d'accès aux données. Le premier concerne le coût prohibitif du CASD : 3 000 euros annuels pour une configuration de base souvent insuffisante, puis 3 800 euros pour la configuration supérieure généralement nécessaire pour un simple chapitre de thèse. Cette situation oblige les doctorants à soumettre des projets de recherche spécifiques pour financer l'accès aux données, créant ainsi un cercle vicieux puisqu'ils doivent parfois demander des financements aux mêmes services statistiques qui ont produit les données et qui les commercialisent via le CASD.

Ce problème de coût génère également des inégalités territoriales significatives. Si les laboratoires parisiens parviennent généralement à mobiliser les fonds nécessaires, les structures en région rencontrent des difficultés considérables. **Dominique MEURS** mentionne qu'un laboratoire d'Arras disposant seulement de moins de 5 000 euros annuels par chercheur pour l'ensemble des dépenses de recherche (colloques inclus) ne peut tout simplement pas envisager de projets impliquant le CASD.

Le second obstacle majeur concerne les délais d'accès. Les procédures administratives sont extrêmement chronophages, certains chercheurs mentionnant jusqu'à quatre ans d'attente pour accéder à des fichiers appareillés. Cette situation est particulièrement problématique pour les doctorants dont la thèse doit être réalisée en trois ans. Elle estime qu'une approche plus pragmatique s'impose. Si certaines données très sensibles justifient effectivement des processus longs et rigoureux, d'autres pourraient être rendues accessibles plus rapidement via des plateformes alternatives au CASD.

Dominique MEURS suggère que les services de statistique publique gagneraient à intensifier leur dialogue avec les chercheurs, particulièrement les jeunes doctorants, et à leur accorder davantage de confiance.

Un point essentiel concerne également la liberté fondamentale du chercheur. Elle s'inquiète des propositions visant à limiter certaines variables, car de nombreuses découvertes scientifiques majeures surviennent par hasard, lors de l'exploration des données. Les chercheurs sont parfaitement capables de manipuler des bases de données complexes, et restreindre prématurément leur champ d'investigation risquerait de compromettre des avancées potentielles.

Enfin, concernant la méfiance des citoyens évoquée lors d'une intervention précédente, elle approuve totalement la nécessité de transparence comme prérequis. Elle suggère également qu'une présence accrue des chercheurs, notamment des jeunes, dans le débat public contribuerait à renforcer la confiance. Ces derniers reçoivent d'ailleurs désormais une formation à la médiatisation de leurs travaux, tenant compte des spécificités de communication selon les publics et les médias.

Leur accorder une place plus importante dans ce débat public et favoriser davantage le dialogue constituerait un moyen efficace de réduire la méfiance légitime des citoyens vis-à-vis de la collecte des données. Cette ouverture permettrait de démontrer concrètement l'utilité de ces collectes, notamment pour l'évaluation des politiques publiques et la détermination de leur efficacité, même si celle-ci comporte certaines nuances et limites.

Le sens de la nuance, caractéristique du travail quotidien des chercheurs dans leurs publications scientifiques, représenterait un apport considérable au débat public. Cette approche nuancée pourrait

effectivement renforcer la confiance de l'opinion publique envers l'utilisation de leurs données et développer leur esprit critique face aux informations circulant sur les réseaux sociaux, favorisant ainsi une plus grande réflexivité.

Dominique MEURS plaide donc pour l'établissement de davantage de débats entre les statistiques publiques et les jeunes docteurs, ainsi que pour une meilleure reconnaissance de ces derniers qui font preuve d'une brillance remarquable et de compétences extraordinaires. Elle estime qu'il est nécessaire de leur accorder plus d'attention, de les écouter et de leur faciliter l'accès aux données à moindre coût.

Bertrand du MARAIS remercie les intervenants pour leurs contributions. Il souligne que la conclusion présentée par Dominique Meurs constitue en réalité une véritable « ode au Cnis » puisque le débat représente l'essence même de cette institution. Avant de donner la parole à la salle, il propose au panel de réagir immédiatement aux propos qui viennent d'être tenus.

Corinne PROST aborde plusieurs aspects concernant les relations entre statisticiens et chercheurs. Elle souligne tout d'abord les initiatives récentes prises par plusieurs SSM qui ont organisé des séminaires ouverts à la recherche sur le thème de l'accès aux données. Ces événements, distincts du Cnis, ont été spécifiquement dédiés à cette problématique. La Depp a initié cette démarche, suivie par le SSM Agriculture et alimentation, qui a attiré 200 participants, constituant un véritable succès. D'autres SSM ministériels prévoient d'organiser des événements similaires.

La prise de conscience de la nécessité d'aborder ces questions d'accès aux données et de diffuser l'information est désormais bien établie. En effet, certaines difficultés résultent parfois simplement d'un manque d'information sur les accès disponibles et les procédures à suivre.

L'engagement se poursuit également dans la fourniture de fichiers de production pour la recherche, car il n'est pas toujours nécessaire d'accéder au CASD pour réaliser de simples statistiques descriptives sur des sujets peu complexes. Cependant, la situation devient plus problématique concernant les données administratives exhaustives. **Corinne PROST** cite l'exemple du recensement agricole ; du fait de l'exhaustivité, il est plus facile d'identifier un agriculteur donné dans la base du recensement agricole. Face à cette difficulté, l'approche privilégiée consiste à créer des échantillons à partir des données exhaustives pour les mettre à disposition. Dans le cas du recensement agricole, cette solution a été facilitée par l'existence parallèle d'une enquête sur échantillon.

Par ailleurs, elle rappelle la position de Xavier Timbeau, président de la commission Environnement et développement durable du Cnis, qui a souligné l'importance du discernement dans l'utilisation des données. Toutes les données n'ont pas la même valeur et il convient d'éviter une approche tous azimuts qui risquerait de faire perdre du temps sur des données finalement peu pertinentes. Ce discernement doit s'accompagner de sobriété, concept qui complète celui de proportionnalité évoqué précédemment.

Cette approche de sobriété s'applique également dans le domaine de l'intelligence artificielle, où il convient de s'interroger sur la pertinence de modèles très consommateurs de ressources. Cibler les usages de manière précise apparaît comme une démarche plus efficace et judicieuse que la recherche d'une solution universelle.

François CLANCHÉ exprime son accord général sur la nécessité d'éviter une diffusion excessive et non ciblée des données. Il souligne cependant que l'identification des variables pertinentes nécessite souvent une phase de test portant sur de nombreuses variables pour déterminer celles qui présentent un réel intérêt. Cette démarche exploratoire, parfois impossible pour les statisticiens publics en raison de leurs autres obligations, pourrait bénéficier de la contribution des chercheurs, qu'ils soient jeunes ou expérimentés.

Il s'inquiète donc d'une restriction trop importante ou prématurée de l'accès aux données, qui serait motivée par des considérations inappropriées.

Corinne PROST acquiesce sur le caractère problématique des restrictions fondées sur de mauvaises raisons. Elle évoque la possibilité de différencier les appariements pour enjeux méthodologiques de la diffusion en routine, suggérant que la première catégorie pourrait faire l'objet d'une collaboration plus poussée pour identifier les variables pertinentes.

Bertrand du MARAIS indique tout d'abord que le cadre de référence actuel s'applique principalement aux services de la statistique publique. Tous les acteurs concernés ont cependant conscience de la nécessité d'engager un travail d'identification d'éléments harmonisés de méthodes pour les chercheurs eux-mêmes, ce qui constituera la prochaine étape.

Ensuite, il rappelle que la France se distingue par sa réticence à l'égard des registres de population, pour diverses raisons, notamment des aspects tragiques de notre histoire. Cette spécificité présente des avantages, mais également des inconvénients potentiels.

Enfin, **Bertrand du MARAIS** observe que le cadre de référence intègre plusieurs éléments de garantie qui figuraient dans l'avis de la Cnil. Il souligne ainsi la cohérence entre l'avis de la Cnil sur Résil, le décret créant Résil et le cadre de référence.

Kamel GADOUCHE, directeur du CASD, remercie Dominique Meurs pour sa présentation, qu'il juge très éclairante sur le besoin des chercheurs. Il s'associe pleinement à cette analyse et souhaite apporter des précisions sur la question du coût.

La matinée a été consacrée à des exposés sur la confiance et l'amélioration des appariements et de l'explicabilité des données. La confiance passe nécessairement par la protection des données et la sécurité quant à leur usage. La sécurité a un coût. Cette situation n'est pas propre à la France, mais se retrouve également dans d'autres pays, comme les Pays-Bas ou le Canada.

Une correction importante s'impose concernant les tarifs mentionnés : pour un projet doctoral comprenant un serveur de traitement des calculs avec stockage des données, le coût est bien de 4 200 euros par an. En revanche, l'utilisateur supplémentaire ne coûte pas 4 200 euros, mais 420 euros par an. Ce coût marginal est relativement faible, car il n'implique pas l'installation d'un nouveau serveur.

Le CASD encourage depuis longtemps l'accès aux FPR, fournissant d'ailleurs des environnements pour la réalisation de FPR comme, par exemple, sur la base exhaustive du recensement agricole et d'autres sources de données. Le CASD est particulièrement attaché à l'existence de sources complémentaires FPR permettant d'éviter le recours à des procédures plus lourdes lorsque cela n'est pas nécessaire. Il est pleinement conscient que le processus d'enrôlement, avec sa dimension sécuritaire, représente une contrainte pour les chercheurs. Cette sécurité constitue cependant un mal absolument nécessaire.

En contrepartie de ces contraintes, il est essentiel que les chercheurs puissent accéder à d'autres sources de données moins sensibles quand cela est suffisant, notamment via les FPR ou l'open data. À ce titre, le CASD a développé sur ses ressources pendant plusieurs années un portail pour télécharger des données FPR pour le compte de Progedo. Néanmoins, lorsque l'anonymisation pose des problèmes techniques, un niveau de sécurité adapté devient indispensable.

La question du coût du financement de la recherche demeure entière : ces coûts existent, ne peuvent être supprimés, et nécessitent des solutions de financement que le CASD ne peut apporter seul.

Concernant la durée des procédures, **Kamel GADOUCHE** reconnaît que les délais étaient parfois longs, mais conteste l'existence de procédures durant quatre ans. En moyenne, les délais se sont nettement améliorés ces dernières années, atteignant désormais deux à trois mois. Des progrès significatifs ont été réalisés grâce aux investissements du Comité du secret statistique et des producteurs. Par comparaison internationale, si la France présente des délais légèrement plus longs, elle offre en contrepartie un avantage majeur : l'accès à diverses sources de données avec une seule demande. Cette particularité, rendue possible grâce à la gouvernance mise en place par le service statistique public français, permet d'accéder à plusieurs sources de données provenant de multiples producteurs dans un environnement unique. De plus, ce dispositif inclut également des données de la Banque de France et d'autres organismes, ce qui est remarquable compte tenu des délais d'instruction nécessaires pour l'accès à ces données multiples parfois fortement identifiantes.

Dominique MEURS intervient pour compléter son propos précédent en soulignant l'excellence des services rendus par le CASD, information qui lui a été rapportée par tous les doctorants interrogés. Elle précise également que sa remarque concernant la durée des procédures faisait référence à l'ensemble du processus, depuis le repérage initial des données jusqu'à l'obtention effective de l'accès au CASD, et non uniquement à la phase finale d'accession.

Jean-Luc TAVERNIER aborde la question du coût du CASD, dont le statut implique qu'il devait, après l'obtention d'une subvention initiale, trouver les moyens de son autofinancement.

Le service rendu, particulièrement lorsqu'il est de qualité, a nécessairement un coût qui doit être assumé par quelqu'un. Il souligne n'avoir constaté ni enrichissement personnel au sein du CASD, ni dépenses somptuaires, ni même une rémunération avantageuse des employés par rapport au marché, bien au contraire. Ce coût incompressible doit impérativement être pris en charge d'une façon ou d'une autre.

Il est essentiel d'abandonner l'idée selon laquelle un laboratoire de sciences humaines et sociales n'aurait pas besoin de financement, comme si un crayon et une feuille de papier suffisaient. La production de travaux quantitatifs avec des données dans un environnement sécurisé requiert un investissement financier. Les sommes en question restent relativement modestes au regard des budgets universitaires, particulièrement en comparaison avec les laboratoires de physique.

Jean-Luc TAVERNIER précise qu'il sera impossible de trouver des financements dans le budget de la statistique publique pour subventionner l'accès des chercheurs. La situation se complique davantage avec la réduction de la contribution du Centre national de la recherche scientifique (CNRS) au CASD, qui était déjà symbolique.

Pour conclure sur une note plus constructive, il évoque une piste explorée dans un rapport rédigé avec Nicolas Véron : la possibilité d'inclure dans les appels à projets une provision pour le coût du CASD sur une durée plus longue. En effet, le financement actuel s'arrête généralement à la remise du travail, alors que la recherche se poursuit bien au-delà avec les phases de finalisation, de publication et de réponse aux évaluations. Il serait donc judicieux de prévoir un financement sur une période plus étendue pour tenir compte de cette réalité.

Concernant le CASD, **Stéphane JUGNOT** rejoint la position de Jean-Luc TAVERNIER, en soulignant que les sciences dures financent leurs outils d'observation et qu'il est donc normal que les sciences sociales financent également les leurs. L'idée selon laquelle les données administratives seraient gratuites repose sur un malentendu.

Concernant le principe de minimisation évoqué précédemment, il considère que le rôle des producteurs de statistiques publiques ne se limite pas à mettre à disposition des données administratives brutes. Ces producteurs effectuent légitimement un travail de sélection et de reconstruction qui facilite le travail des chercheurs grâce à leur expertise des données. Cette phase préparatoire est donc parfaitement justifiée.

Le principe de minimisation et de proportionnalité devrait également conduire les chercheurs à préciser spécifiquement les données souhaitées plutôt que de demander l'accès à un fichier complet. Cette démarche implique de déterminer si un échantillon est suffisant ou si des données exhaustives sont nécessaires. Cette approche soulève cependant la question des moyens nécessaires pour réaliser ces extractions ciblées.

Ce problème se posera inévitablement à terme. **Stéphane JUGNOT** illustre son propos avec l'exemple des enquêtes Génération du Céreq qui comprennent le Siret, des données communales, etc. Ces données sont protégées par le secret statistique, mais tous les chercheurs n'ont pas nécessairement besoin de l'ensemble de ces dimensions. La mise à disposition uniquement des dimensions utiles devient donc une préoccupation majeure.

Il est ainsi nécessaire de travailler à la fois sur l'habitude des chercheurs à préciser leurs besoins et sur la facilitation des extractions des seules données nécessaires. C'est cette démarche qui caractérise véritablement le principe de minimisation.

Concernant la question du coût, **Roxane SILBERMAN**, souhaite rappeler que lorsqu'elle était en charge du Réseau Quetelet au moment de la mise en place de Progedo, elle a dû à l'époque, d'abord avec le CNRS puis avec le Ministère de l'Enseignement supérieur et de la Recherche, obtenir d'eux le versement à l'Insee d'un montant annuel de 20 000 € pour que ces FPR soient *« gratuits »* pour les utilisateurs finaux.

Les chercheurs pourraient se mobiliser collectivement, au niveau du CNRS comme au niveau ministériel, pour obtenir un système pérenne de financement pour les coûts à régler au CASD s'agissant du traitement sécurisé des données confidentielles Elle compare cette situation à celle de la généralisation des ordinateurs dans les laboratoires où il avait bien fallu prendre en compte finalement les sciences humaines

et sociales pour lesquelles le crayon ne suffisait pas. Le CNRS avait finalement mis en place une ligne spécifique permettant de soumettre des demandes pour ces équipements. Elle suggère qu'une approche similaire pourrait être envisagée pour les données.

Sa deuxième réflexion porte sur les appariements de données. Dans les commissions du CNIS, l'examen des demandes d'accès aux données administratives ne suscitent généralement aucune discussion. Cela risque d'être le cas pour l'examen ponctuel de telle ou telle demande d'appariement. Ne faudrait-il pas s'interroger sur ce qui inquiète le plus le citoyen ? N'est-ce pas plutôt la quantité ? et quelle quantité ? La quantité d'informations dans une megabase ou bien le nombre croissant d'appariements réalisés ? Cette question rejoint la problématique de minimisation et nécessite une réflexion sur l'équilibre à trouver et peut-être une coordination des demandes d'appariement pour éviter leur multiplication.

Roxane SILBERMAN distingue ensuite deux objectifs très différents des appariements : enrichir les données et minimiser la charge pour les répondants. L'enrichissement implique d'ajouter des données via des appariements multiples, tandis que la minimisation de la charge pourrait conduire à supprimer des enquêtes au profit d'appariements.

L'équilibre entre développement des appariements et enquêtes constitue un enjeu crucial. Davantage d'appariements au détriment des enquêtes ne serait pas nécessairement bénéfique. Les données administratives sont très riches mais ne contiennent que des informations nécessaires à des objectifs de gestion. Typiquement, par exemple, elles ne peuvent inclure une information telle que le pays de naissance des parents. Avec une enquête, il est facile d'ajouter un petit module complémentaire.

Elle appelle donc à engager une réflexion sur un nouvel équilibre à trouver autour du corpus d'enquêtes centrales permettant tant des appariements que des modules complémentaires.

Françoise DUPONT indique que les participants aux commissions du Cnis, c'est-à-dire les utilisateurs de données, supporteront l'analyse de la proportionnalité soumise à l'examen des commissions.

Hervé PIVETEAU note que les procédures d'accès aux fichiers détaillés pour la recherche sont des dispositions relatives au partage du secret permettant aux chercheurs d'accéder aux données protégées. Il souligne l'absence de procédure de levée du secret qui empêche l'accès aux données.

François CLANCHÉ revient sur la question du CASD en soulignant le principe que tout service ayant un coût doit être financé par l'utilisateur. Cette problématique a été remontée au Ministère de la Recherche et de l'Enseignement supérieur par ses soins dans le cadre de ses activités dans le monde de la recherche. Les chercheurs établis dans leur carrière trouvent généralement les quelques milliers d'euros nécessaires, mais la difficulté se pose principalement pour les jeunes chercheurs. La communauté scientifique doit donc se mobiliser pour maintenir ces financements, bien que le contexte actuel ne soit pas favorable.

Corinne PROST trouve intéressant le retour de **Roxane SILBERMAN** concernant la perception des utilisateurs : multiplier les appariements pourrait générer plus d'inquiétudes que la création d'un entrepôt centralisé de grande taille. Cette observation mérite réflexion, car elle remet en perspective les discussions de la journée.

Le cadre de référence en cours de mise en place est susceptible d'évoluer. Concernant la remarque de Françoise Dupont sur la minimisation et la proportionnalité, il ne s'agit pas de dresser une liste exhaustive des variables. L'objectif est plutôt d'obtenir un engagement des services producteurs à respecter ce principe de minimisation, engagement déjà existant vis-à-vis de la Cnil dans le cadre du RGPD, mais qui se trouve ainsi renforcé.

Les commissions du Cnis, historiquement habituées à travailler sur les enquêtes, vont désormais évoluer dans leur méthodologie pour traiter les appariements. Un bilan sera réalisé après une période d'expérimentation pour évaluer ce qui fonctionne et ce qui doit être amélioré.

Bertrand du MARAIS confirme que la période actuelle constitue une étape dans un processus évolutif. Le Cnis souhaite établir une clause de revoyure à deux ans pour faire le point sur la mise en œuvre du cadre de référence. Un autre chantier devra être ouvert spécifiquement pour l'utilisation des appariements réalisés à la demande ou par des chercheurs, sujet particulièrement complexe comme a commencé à le montrer cette discussion.

Concernant l'équilibre entre appariements et enquêtes, cette question est fondamentale. Le Cnis, dont la mission inclut d'assurer la pertinence et la proportionnalité des travaux effectués, gardera systématiquement cette préoccupation à l'esprit.

François CLANCHÉ s'interroge sur les critères d'évaluation de la pertinence. Pour les enquêtes traditionnellement examinées par le Cnis, la longueur du questionnaire sert généralement d'indicateur, permettant d'estimer la charge imposée aux répondants, qu'il s'agisse d'entreprises ou de ménages. Le producteur de l'enquête est censé avoir optimisé son questionnaire pour respecter les contraintes de temps.

Pour les appariements, il suggère de développer des indicateurs similaires, comme le nombre de variables ou leur répartition par thème. Ces métriques permettraient aux producteurs de démontrer qu'ils ont effectivement réfléchi à la question de la proportionnalité, et faciliteraient l'évaluation de l'équilibre général du dispositif. Il reconnaît néanmoins que la mise en œuvre ne sera jamais simple.

Mathilde GODARD, chargée de recherche CNRS à l'Université Paris Dauphine, prend note que le cadre de référence actuel pour les appariements ne concerne pas ceux initiés directement par les chercheurs, mais se réjouit que ce point soit identifié comme un chantier futur auquel elle sera heureuse d'être associée.

En attendant, certains aspects discutés présentent un intérêt immédiat pour les chercheurs, notamment la possibilité de collaborer avec une équipe du SSP pour initier un appariement. La proposition de systématiser les appels à projets faciliterait l'identification d'interlocuteurs dans le SSP sans nécessiter de relations préexistantes.

De même, l'idée d'établir une liste référençant les appariements sur le site de l'Insee serait extrêmement utile. Par analogie, elle utilise fréquemment le répertoire du CASD, qui recense toutes les sources disponibles pour alimenter sa réflexion et générer des idées de recherche.

Concernant les délais d'accès, elle reconnaît avoir peut-être exagéré en évoquant quatre ans, mais précise que son expérience portait sur un appariement avec des données de santé particulièrement complexes (EDP-Santé). Cette procédure a nécessité l'avis du Cesrees, de la Cnil et du Comité du secret, puis l'établissement d'une convention avec la Drees, dont la seule signature a pris un an. Au total, la procédure aura duré quatre ans.

Elle salue également la proposition de mettre à disposition sur Progedo des échantillons des appariements, éventuellement avec moins de variables identifiantes. Cette pratique, déjà en place pour certaines sources et dans d'autres pays, comme l'Allemagne, permet aux chercheurs de commencer leurs travaux pendant le traitement des demandes d'accès aux données complètes.

Louis-André VALLET, CNRS, s'interroge sur le fait que certains services producteurs ne mettent pas leurs données à disposition sur le CASD, citant spécifiquement le cas de la Depp. Il souhaite comprendre les raisons de cette situation et savoir si des évolutions sont envisageables pour soutenir la recherche en éducation.

Jean-Luc TAVERNIER répond que ce sujet fait débat depuis de nombreuses années, y compris dans le cadre de débats législatifs. Néanmoins, il annonce que la directrice de la Depp s'est engagée à transmettre ces informations au CASD. Un calendrier de déploiement a été établi pour la fin de l'année en cours.

Roxane SILBERMAN souligne à propos des possibilités pour les chercheurs de collaborer avec des SSM pour des appariements qu'il ne faut pas que cela conduise à retarder du même coup la mise à disposition de ces nouveaux jeux de données pour d'autres chercheurs. Cette situation s'étant produite à plusieurs reprises par le passé lors de collaborations par ailleurs tout à fait souhaitables, elle recommande que le processus soit correctement encadré.

Patrice DURAN, professeur émérite à l'École normale supérieure Paris Saclay, rappelle que la statistique constitue également un outil de gouvernement. Les enjeux actuels relèvent davantage du traitement des problèmes publics, par nature transversaux, que de simples logiques de production. Dans ce contexte, la coordination devient un aspect fondamentalement crucial. Les discussions portent souvent sur la nécessité de dépasser les logiques de silos. Les appariements représentent précisément un mode de traitement favorisant cette coordination.

Christelle MINODIER apporte un commentaire sur la diffusion des données, sujet largement abordé concernant l'accès des chercheurs. La Drees est très favorable à la mise à disposition des données, mais le processus nécessite un temps considérable.

La méthode de diffusion, que ce soit via FPR ou CASD, dépend directement de la sensibilité des données concernées. Le choix du CASD n'est pas une solution de facilité. Par ailleurs, une diffusion sur mesure des données nécessaires, projet par projet, bien qu'idéale pour respecter le principe de minimisation, représenterait un coût prohibitif. Cette approche personnalisée permettrait d'adapter précisément les composantes potentiellement identifiantes en fonction de chaque projet, mais n'est pas réalisable avec les ressources disponibles.

L'accompagnement des chercheurs, notamment pour l'exploitation des données complexes, comme celles de l'EDP-Santé ou du SSM, mobilise des ressources importantes. Malgré ces contraintes, les services statistiques demeurent très favorables au partage de leurs données avec les chercheurs, dont les travaux complètent utilement ceux réalisés au sein de la statistique publique.

CLÔTURE

Bertrand du MARAIS exprime sa reconnaissance envers tous les intervenants qui ont apporté des éclairages précis et concrets durant les sessions du matin et de l'après-midi.

Il tient également à remercier le directeur général de l'Insee, Jean-Luc Tavernier, qui soutient ce sujet depuis au moins 2022, ainsi que son prédécesseur à la présidence du Cnis, Patrice Duran, favorable à la mise en place du groupe de concertation sous l'égide du Cnis qui a permis d'aborder le projet Résil de manière apaisée. Ces efforts bénéficient aujourd'hui à l'ensemble des participants.

Il remercie par ailleurs l'audience pour sa participation active aux débats, conformément à la mission du Cnis de favoriser les échanges entre producteurs et utilisateurs de données statistiques. Sa gratitude s'étend également, pour la très bonne organisation de ces rencontres, à l'équipe du Secrétariat général, notamment à Christine Lagarenne, François Guillaumat-Tailliet, en bonne coordination avec Josy Dussart de l'équipe Résil.

Il se dit impressionné par l'intégration progressive de la notion d'appariement dans les pratiques courantes, au point qu'un néologisme comme « CSNSïsation » a émergé. Sur une note plus sérieuse, il reprend les propos de Jean-Christophe Sciberras concernant la tension entre transparence et confidentialité, estimant que ces deux notions doivent en réalité rester complémentaires plutôt qu'opposées, car la transparence s'applique aux procédures et la confidentialité, aux données.

Il évoque également une inquiétude exprimée le matin même concernant l'utilisation potentiellement malveillante de ces systèmes. À cet égard, il rappelle que la Cnil maintient une approche hypothético-déductive, anticipant les scénarios parfois les plus problématiques. Ceci constitue une garantie malgré les critiques parfois suscitées par cette méthode.

À ce titre, **Bertrand du MARAIS** attire l'attention sur l'avis de la Cnil concernant le projet Résil, dans lequel sont prises en compte diverses garanties proposées par l'Insee et approuvées par le groupe de concertation. Ces garanties sont particulièrement robustes, englobant des mesures de chiffrement, de sécurité et même de réactivité.

Dans cette problématique, la collégialité et le dialogue entre les différents acteurs sociétaux de la statistique publique constituent un rempart essentiel contre d'éventuels abus de pouvoir. C'est précisément le rôle du Cnis que de favoriser cette concertation. Par conséquent, les projets de loi circulant actuellement dans les institutions parlementaires, qui visent à simplifier le système en supprimant soit le Comité du secret, soit le Cnis lui-même, représenteraient une atteinte significative à cette délibération démocratique fondamentale pour le secteur.

Sur le plan plus concret, une nouvelle version du cadre de référence sera élaborée en s'appuyant sur les discussions menées au Bureau et lors de cette rencontre et sur la contribution reçue de membres du Cnis. Ce travail sera réalisé en étroite collaboration avec le groupe de travail constitué au sein de l'Insee,

comprenant des représentants du SSM. Cette version révisée sera présentée lors de la prochaine réunion du bureau prévue le 4 juin 2025.

Bertrand du MARAIS annonce également le lancement ultérieur d'un chantier spécifiquement dédié aux problématiques des chercheurs. Par ailleurs, une clause de revoyure est prévue dans deux ans afin d'évaluer la mise en œuvre effective de ce cadre de référence.

En conclusion, il remercie l'assistance pour la richesse des débats et donne rendez-vous aux participants pour les prochaines échéances du Cnis.

La séance est levée à 15 heures 50.