La place des appariements au sein du service statistique public

Rencontre du Cnis du 28 mai 2025

C. Colin et C. Prost

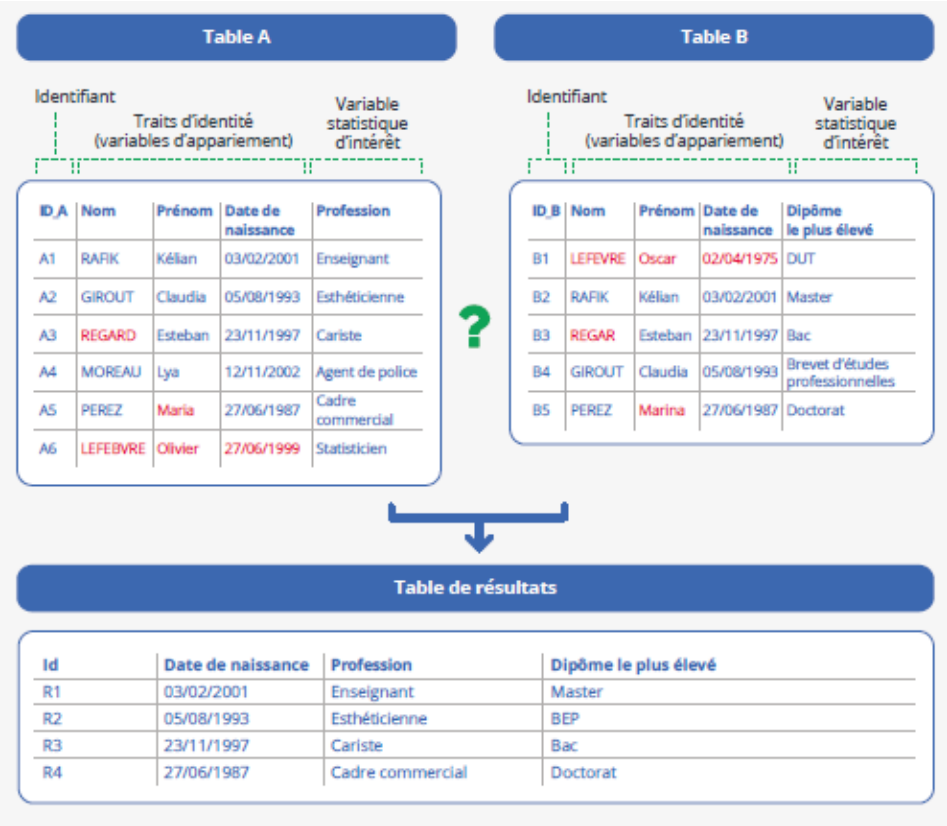
Le service statistique public (= l'Insee et les services statistiques ministériels) collecte des informations d'origines diverses, notamment par enquêtes et en réutilisant des données administratives. Elles peuvent être utilisées seules ou combinées pour fournir une information plus riche

Apparier des données relatives aux individus, c'est rapprocher pour une même personne les données la concernant et qui sont issues de différentes sources

Nouveau : Une définition juridique dans le décret fondant le répertoire statistique des individus et des logements, Résil (décret du 5 janvier 2024) :

[Les appariements facilités par le répertoire] 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. **Ces appariements donnent lieu à la création de nouveaux fichiers, lesquels constituent des traitements de données à caractère personnel** au sens du RGPD

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



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

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

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

Extrait de : « Les appariements : finalités, pratiques et enjeux de qualité », Courrier des statistiques n°11, juin 2024

LE SERVICE STATISTIQUE PUBLIC A UNE PRATIQUE ANCIENNE DES APPARIEMENTS DE DONNÉES INDIVIDUELLES :

- **Cette histoire, ainsi qu'un panorama des usages des appariements, avaient été retracés lors de la première rencontre du Cnis sur les appariements de janvier 2022**

LES PRÉCURSEURS :

- **Enquête revenus fiscaux puis revenus fiscaux et sociaux depuis 1956 (Insee)**
  - A l'origine par rapprochement de données du recensement et de données fiscales pour un échantillon de personnes recensées ; maintenant : enquête emploi + données fiscales + données sociales
  - Source de référence sur les niveaux de vie, les inégalités et la pauvreté monétaires
- **Échantillon Démographique permanent (EDP) depuis 1968 (Insee)**
  - A l'origine recensement de la population + données d'état civil pour un échantillon de personnes sélectionnées sur leur jour de naissance
  - Extensions progressives : inclusion des Dom, extension de l'échantillon des personnes sélectionnées, ajout progressif de nouvelles sources : fichier électoral, données sur salaires et périodes d'emploi, données socio-fiscales sur les revenus
  - Étude des trajectoires de vie familiale, résidentielle, évolution des revenus et déménagements en fonction d'évènements de vie
- **Panels divers depuis les années 70 : d'élèves (Depp), de salariés… complétés éventuellement d'enquêtes**

DEPUIS DES DÉCENNIES LES APPARIEMENTS ONT PERMIS DE METTRE EN PLACE DE NOUVELLES SOURCES, PLUS RICHES, PERMETTANT DE RÉPONDRE À DE NOUVELLES QUESTIONS

ILS PEUVENT CONCERNER :

- **Des rapprochements entre enquêtes** (différentes pour la même période - assez rare - ou la même enquête sur plusieurs années consécutives)
- **Des rapprochements entre enquêtes et données administratives** : enrichir les enquêtes de variables issues de données administratives ; compléter les données administratives en lançant des enquêtes sur un échantillon, ce qui permet d'allier la richesse des variables des enquêtes et l'exhaustivité des données administratives
- **Des rapprochements entre données administratives**, ce qui permet de caractériser des populations rares et d'étudier des phénomènes à un niveau géographique fin

LA FORCE DE CE MODE DE COLLECTE : DÉVELOPPER DE NOUVELLES SOURCES À UN COÛT RAISONNABLE EN TIRANT LE MEILLEUR PARTI DES QUALITÉS DE CHAQUE SOURCE EN ENTRÉE

LA MISE EN PLACE DU CODE STATISTIQUE NON SIGNIFIANT (CSNS) :

- **L'idéal pour réaliser un appariement entre deux fichiers : disposer d'un identifiant pour chaque individu, commun aux deux fichiers**
- **Le NIR, numéro d'inscription au répertoire national d'identification des personnes physiques, est un très bon exemple d'identifiant, l'INE, identifiant national étudiant, aussi**
- **Mais toutes les bases de données ne contiennent pas le NIR et l'usage du NIR est très encadré**
  - Avant 2018, les appariements de fichiers sur la base du NIR pour les besoins de la statistique publique nécessitaient, pour chaque traitement, un décret en Conseil d'État après avis publié et motivé de la Cnil
- **Pour permettre le développement des appariements à des fins statistiques tout en garantissant un niveau élevé de protection des données à caractère personnel, la loi pour une République numérique de 2016 (article 34) prévoit que les appariements peuvent se faire sans avis de la Cnil ni décret si on utilise un code statistique non signifiant à la place du NIR**
  - Le code statistique non signifiant (CSNS) est dérivé du NIR mais non signifiant
  - Il ne permet pas de revenir à l'identité des personnes, ni à leur NIR
  - Il garde les bonnes propriétés d'un identifiant : unique pour chaque individu

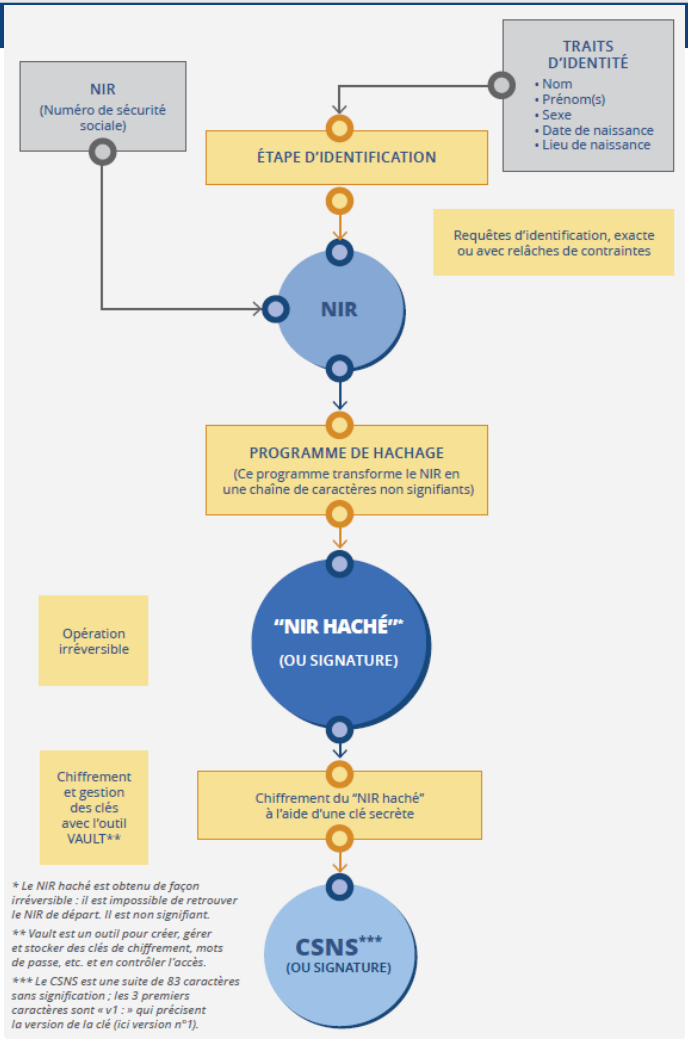
CONCRÈTEMENT, L'INSEE A MENÉ UN PROJET, PRÉSENTÉ LORS DE LA RENCONTRE DU CNIS DE JANVIER 2022, POUR METTRE EN PLACE LE SERVICE DE CALCUL ET DE FOURNITURE DU CSNS

- La mise en service complète est intervenue en 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 doit être unique pour chaque individu, quelle que soit la source de départ

LE CSNS NE PEUT ÊTRE TRANSMIS EN DEHORS DE L'INSEE ET DES SERVICES STATISTIQUES MINISTÉRIELS, IL NE PEUT ÊTRE UTILISÉ QU'À DES FINALITÉS STATISTIQUES

- Les chercheurs peuvent utiliser les nouvelles bases de données issues des appariements mises à disposition au CASD (sans accéder au CSNS lui-même)

CALCUL DU CODE STATISTIQUE NON SIGNIFIANT



Extrait de : « Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers », Courrier des statistiques n°9, juin 2023

Exemple de CSNS :

v1:3zUY-VUpJFX9g7z3oi3zKSIKXB0yLIE4oGmxsQi1Z2atWI0n8lfnmVrxWUiJqeKHHTvcH+i0PvuO2991M

9 SSM (BIENTÔT 10) ET 6 SERVICES DE L'INSEE UTILISENT LE SERVICE CSNS

CES UTILISATEURS DOIVENT SIGNER AVEC L'INSEE UNE CONVENTION DE SOUS-TRAITANCE, QUI COMPORTE UNE CHARTE D'USAGE

LE SERVICE CSNS PRODUIT UNE MESURE DE LA QUALITÉ DE L'IDENTIFICATION DES PERSONNES, INDISPENSABLE AFIN QUE LES UTILISATEURS PUISSENT APPRÉCIER LA FIABILITÉ DE L'APPARIEMENT

- Qualité du fichier en entrée : qualité des traits d'identité (valeurs manquantes, anomalies)
- Qualité de l'identification pour chaque enregistrement sur le critère des « faux positifs » (on attribue un NIR et donc un CSNS mais pas le bon) : une note qualité de chaque CSNS (de « parfaitement fiable » à « non fiable ») est calculée

  => on attribue un CSNS pour chaque individu, en documentant la qualité, ce qui permet ensuite d'apparier directement

DANS LES BILANS ANNUELS, LES SSM ET LES DÉPARTEMENTS DE L'INSEE DÉCLARENT 34 APPARIEMENTS RÉALISÉS EN 2023 ET 51 EN 2024

DES USAGES VARIÉS, PROCHES DES USAGES HISTORIQUES OU PLUS NOUVEAUX

DES ENRICHISSEMENTS DE DISPOSITIFS EXISTANTS QUI OUVRENT DE NOUVELLES POSSIBILITÉS D'ÉTUDES ; QUELQUES EXEMPLES :

- **Étudier le non-recours au minimum vieillesse** en appariant les données de l'échantillon interrégimes de retraités (EIR) de la Drees avec les données issues des déclarations fiscales => « Une personne seule sur deux, éligibles au minimum vieillesse, n'y recourt pas »
- **Étudier la vulnérabilité énergétique des ménages** en appariant le répertoire statistique des véhicules routiers du Sdes et le fichier démographique sur les logements et les individus (Fidéli, à partir des données fiscales) (présenté ensuite)
- **Étudier la multipropriété et qui détient quoi** en appariant les données du cadastre, Fidéli, le registre des entreprises et le registre des bénéficiaires effectifs : un appariement qui avait été développé sous forme d'un prototype sur une année donnée, et qui devient une nouvelle base de données annuelle (Base de la propriété foncière)

DE NOMBREUX APPARIEMENTS POUR ÉTUDIER LES TRAJECTOIRES INDIVIDUELLES, AVEC OU SANS DIMENSION D'ÉVALUATION DES POLITIQUES PUBLIQUES ; QUELQUES EXEMPLES :

- **Étudier le passage entre structures de l'insertion par l'activité économique (IAE) et contrats aidés** en appariant les données sur les bénéficiaires de l'IAE et de contrats aidés sur plusieurs millésimes (Dares)
- **Étudier la trajectoire professionnelle des sortants de l'enseignement supérieur grâce au dispositif Insersup (présenté ensuite)**
- **Étudier le recours aux minima sociaux en fin de carrière et au passage à la retraite** en rapprochant les données de l'échantillon interrégimes de retraités (EIR), de l'échantillon interrégimes de cotisants (EIC) et de l'échantillon national interrégimes d'allocataires de compléments de revenus d'activité et de minima sociaux (ENIACRAMS) de la Drees => nouvelle base de données, accessible aux chercheurs via le CASD. « Plus d'un bénéficiaire des minima sociaux en cours de carrière sur trois l'est encore après son départ à la retraite »

UNE DIMENSION BEAUCOUP PLUS DÉVELOPPÉE QUE PAR LE PASSÉ AVEC LA MISE EN PLACE DU CSNS : LES APPARIEMENTS À VOCATION MÉTHODOLOGIQUE, LES TESTS POUR ENVISAGER DES ÉVOLUTIONS DE PROCESSUS, LES ALLÈGEMENTS D'ENQUÊTES

- **Comprendre les écarts entre deux mesures de l'emploi en appariant les données de l'enquête Emploi avec les données administratives sur l'emploi (présenté ensuite)**
- **Instruire un scénario de non-réédition des enquêtes auprès des sortants de dispositifs d'insertion en emploi** en appariant les données sur les contrats de professionnalisation, les contrats aidés et l'IAE avec les déclarations sociales nominatives (Dares)
- **Alternatives à l'enquête Formation et qualification professionnelle (Insee) :**
  - Recours aux sources existantes + mise en place de nouveaux dispositifs :
  - Enrichissement de l'enquête Emploi avec des données issues de sources administratives (base Tous salariés, base Non salariés, Pasrau) pour permettre l'étude des mobilités et des trajectoires professionnelles
  - Enrichissement de la prochaine enquête sur la formation des adultes avec les mêmes données administratives sur l'emploi et les salaires, pour permettre d'étudier le lien entre formation et trajectoires professionnelles

UN DÉVELOPPEMENT DES APPARIEMENTS AU SEIN DU SERVICE STATISTIQUE PUBLIC, FACILITÉ

- Par un partage des méthodes
- Par le code statistique non signifiant pour les données personnelles (hors appariement avec le système national des données de santé)
- Bientôt par Résil qui fournira un service d'appariement de données qui permettra de gagner en qualité, en harmonisation et en sécurité

DANS UN CADRE JURIDIQUE ET TECHNIQUE QUI S'ASSURE DE LA PROTECTION DES DONNÉES

DANS UN CADRE DÉONTOLOGIQUE RENFORCÉ SUITE AUX RECOMMANDATIONS ÉMISES PAR LE GROUPE DE CONCERTATION DE RÉSIL PLACÉ SOUS L'ÉGIDE DU CNIS

LE CADRE JURIDIQUE GÉNÉRAL À LA PRODUCTION STATISTIQUE

- La loi de 1951, « loi statistique » qui définit le secret statistique, encadre l'ensemble des travaux statistiques et les autorise
- La loi Informatique et libertés de 1978, qui définit la protection des données personnelles, et s'inscrit à partir de 2018 dans le cadre juridique européen du règlement général sur la protection des données (RGPD)
  - Justifier la légitimité des finalités poursuivies, garantir que les données traitées sont adaptées à ces finalités et limiter leur conservation à la durée strictement nécessaire à ces finalités
  - Le RGPD renforce les droits des personnes concernées avec une obligation de transparence plus grande.
- La loi de 1951 comme le règlement européen 223 définissent les obligations qui s'attachent à l'usage de données confidentielles, qui ne peuvent être réutilisées qu'à des fins de statistiques ou de recherche. Des sanctions sont prévues en cas de violation du secret statistique.

DANS LE CADRE DU RGPD :

- Un appariement de données personnelles est un traitement de données au sens juridique. Un responsable de traitement est identifié et doit : vérifier le respect des principes de nécessité, minimisation et proportionnalité, inscrire l'appariement au registre des traitements, réaliser une étude d'impact si ce traitement présente certaines caractéristiques (population nombreuse, variables sensibles,…).

PROTECTION DES DONNÉES

- Mesures techniques : protection et surveillance des accès et des réseaux, mots de passe,...
- Mesures organisationnelles : règles et procédures de désignations des agents habilités à accéder aux données confidentielles, définition d'une politique de sécurité, formation et sensibilisation des agents aux enjeux de la confidentialité,…
- Dans le cadre du code statistique non signifiant, des mesures techniques renforcées : hachage du NIR, puis chiffrement du « NIR haché » à l'aide d'une clé secrète ; après appariement, les CSNS doivent être conservés de façon isolée dans un fichier qui ne comprend aucune variable socio-démographique ni aucun trait d'identité ; les CSNS sont renouvelés tous les 10 ans.

LE CADRE GÉNÉRAL DU CODE DE BONNES PRATIQUES DE LA STATISTIQUE EUROPÉENNE

- Les 16 principes s'appliquent bien sûr également dans le cadre des appariements
- 2 principes conduisent à préconiser les appariements : le Principe 9 de charge non excessive pour les déclarants ; le Principe 10 de rapport coût-efficacité

AU-DELÀ DU CODE DE BONNES PRATIQUES

- Le cas spécifique des appariements a conduit à renforcer un « mandat social », issu de réflexions en France comme à l'étranger
- Principe de nécessité ; principe de proportionnalité et de minimisation ; principe de transparence
- C'est le cadre de référence, qui sera détaillé lors de la table ronde cet après-midi, qui pose ces principes, renforçant le code de bonnes pratiques
  - Pédagogie et transparence : Rencontre du Cnis de 2022, billet de blog de l'Insee, appariements décrits dans les programmes de travail et les bilans des services producteurs pour le Cnis

APPARIER DIFFÉRENTES BASES DE DONNÉES PEUT PERMETTRE DE :

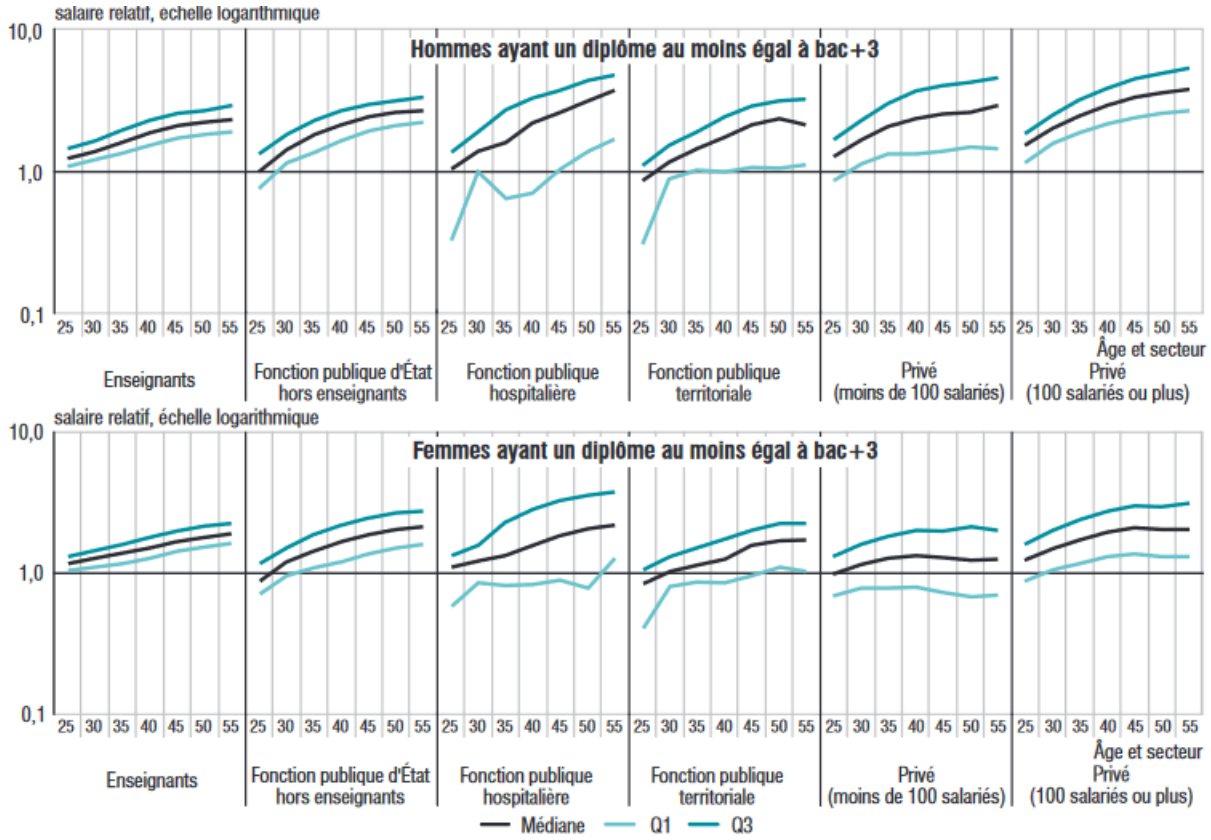
- éclairer certains phénomènes socio-économiques, compléter des champs d'analyse, mieux comprendre le contenu des sources analysées
- améliorer des processus de production : alléger les questionnaires d'enquête, mettre à jour un répertoire, analyser la couverture d'une source

MAIS SOUS PLUSIEURS CONDITIONS :

- Il faut que les informations permettant d'identifier les individus ou les entreprises soient de qualité suffisante. Or il y a une forte variabilité dans la qualité de ces informations dans les données administratives.
- Il faut que les informations fournies par les données soient pertinentes.
- Certaines données sont de meilleure qualité lorsqu'elles proviennent de données administratives que lorsqu'elles sont collectées par voie d'enquête déclarative (ex. : revenu des ménages par exemple).
- Inversement, certains concepts ne peuvent être mesurés que par des enquêtes (ex. : chômage au sens du Bureau international du travail).

UNE COMPLEXITÉ QUI S'ACCROÎT TRÈS RAPIDEMENT

ON AUGMENTE LES DIMENSIONS : PHÉNOMÈNES À LA CROISÉE DE PLUSIEURS DOMAINES ; TRAJECTOIRES



UNE COMPLEXITÉ QUI S'ACCROÎT TRÈS RAPIDEMENT...

- On augmente les dimensions : phénomènes à la croisée de plusieurs domaines ; trajectoires
- Il faut connaître déjà bien les bases ; un précurseur : le panel Trajam (trajectoires des jeunes appariées aux mesures actives du marché du travail) 2010-2015

... CE QUI INCITE À CIBLER LES USAGES, ET DONC CONDUIT À LA MINIMISATION

LES PROJETS UTILISANT LE CSNS VONT DANS CE SENS :

- Le plus souvent appariements de 2 bases de données
- Des objectifs bien définis (thème d'analyse, enjeu méthodologique)

- **Une pratique très ancienne d'appariements individuels dans le service statistique public**
- **Des avancées techniques et juridiques qui favorisent davantage d'appariements et impliquent donc une réflexion sur les usages de ces appariements**
- **La concertation doit évoluer pour tenir compte de ces appariements : c'est l'objet du cadre de référence.**

- Quels types de sources l'Insee utilise-t-il pour construire ses statistiques ?, blog de l'Insee, 16 mai 2023
- Les appariements de données de la statistique publique : des analyses enrichies, un cadre juridique protecteur, blog de l'Insee, 1 septembre 2023
- Appariements de données individuelles : vers une meilleure qualité et plus de transparence, Chroniques du Cnis n°32, avril 2023
- Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers, Courrier des statistiques N9, juin 2023
- Confidentialité des données statistiques : un enjeu majeur pour le service statistique public, Courrier des statistiques N9, juin 2023
- Les appariements : finalités, pratiques et enjeux de qualité, Courrier des statistiques N11, juin 2024
- La concertation : une étape essentielle pour le projet Résil, Courrier des statistiques N11, juin 2024
- Comment l'Insee protège-t-il les données qu'il collecte ? Blog de l'Insee, 23 avril 2025

Retrouvez-nous sur

insee.fr

Insee Mesurer pour comprendre