UMR 8184
Thema
THéorie Économique, Modélisation et Applications

THEMA Working Paper n°2017-06
Université de Cergy-Pontoise, France

L'accès aux données très détaillées pour la recherche scientifique

Kamel Gadouche, Nathalie Picard

Février 2017

UNIVERSITÉ de Cergy-Pontoise
CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

L’accès aux données très détaillées pour la recherche scientifique

Kamel Gadouche[1], Nathalie Picard[2]

Résumé

L'accès aux micro-données confidentielles constitue un enjeu stratégique pour les chercheurs dans de nombreuses disciplines allant des sciences sociales, de l'économie ou du transport à l'épidémiologie ou la santé, en passant par l’éducation ou la justice, en particulier dans le cadre de l'évaluation des politiques publiques ciblées. Les pouvoirs publics en France ont pris pleinement conscience de ces enjeux scientifiques, et de nouveaux dispositifs législatifs ont été votés pour faciliter l’accès des chercheurs aux données. Les pouvoirs publics français ont récemment ouvert de nouvelles possibilités d’appariements entre des sources de données distinctes dans des conditions très encadrées de sécurité. Nous proposons un tour d’horizon des besoins des chercheurs, des enjeux en matière d’évaluation de politiques publiques et des possibilités offertes par de telles données.

Mots-clé : Big data, accès sécurisé, confidentialité, appariement, données administratives, évaluation, politiques publiques, panel, endogénéité

Classification JEL : C81, C55, C33, I28, J18, J68, R4

[1] CASD, Centre d’Accès Sécurisé Distant aux données, 6 rue Legrand, 92240 Malakoff
[2] Université Cergy Pontoise, 33 Boulevard du Port, 95000 Cergy-Pontoise, France

Obtenir l'accès aux micro-données confidentielles est un enjeu stratégique pour les communautés des chercheurs dans de nombreuses disciplines allant des sciences sociales, de l'économie, et du transport à l'épidémiologie. La demande de ces données a régulièrement augmenté grâce à l'effet conjugué de trois facteurs principaux : l'avènement de nouvelles méthodes pour traiter de très gros fichiers de données - également appelés *Big Data* comme, par exemple, les très grandes bases administratives, qui sont pour la plupart longitudinales, l'évolution des outils de modélisation informatique ainsi que les possibilités d'enrichissement des données envisageables grâce à l'appariement de différentes sources de données. Ces données très détaillées *Big Data* ne sont pas seulement la clé des progrès réalisés dans de nombreux domaines de recherche, mais elles contribuent également à l'évaluation des politiques publiques dans de nombreux domaines cruciaux pour la société : l'économie, la justice, l'éducation, la santé et le transport. De récentes études ont en effet révélé l'intérêt de l'exploitation de données médico-administratives sensibles pour les politiques de santé publique, mais aussi de données fiscales pour les questions socio-économiques. Par exemple, le secteur de la santé a longtemps été utilisateur de micro-données confidentielles notamment pour étudier les phénomènes épidémiologiques. Ce besoin croissant pour l'utilisation de micro-données confidentielles appariées aux bases de données médico-administratives a ouvert un nouveau champ de recherche pour la communauté scientifique (voir Guesdon *et al*., 2016). Les pouvoirs publics en France ont pris pleinement conscience de ces enjeux scientifiques, et ces dernières années, de nouveaux dispositifs législatifs ont été votés pour faciliter l'accès des chercheurs aux données. Tout récemment, les pouvoirs publics français ont ouvert de nouvelles possibilités de réalisation d'appariements entre des sources de données distinctes dans des conditions très encadrées de sécurité.

1. Besoins de données pour la recherche scientifique

1.1. Quelles données adaptées à quelle problématique ?

La recherche en sciences sociales nécessite des données adaptées pour répondre aux grandes questions sociétales, surtout lorsqu'il s'agit d'évaluer des politiques publiques, qu'elles soient déjà mises en œuvre ou seulement envisagées. Aussi, pour être efficaces, les mesures de politique publique doivent être ciblées, ce qui implique que les personnes susceptibles d'en bénéficier sont statistiquement différentes de celles qui n'y sont pas éligibles. Par exemple, une réforme visant à faciliter l'insertion sur le marché du travail ciblera les personnes ayant le plus de difficultés *a priori* à s'insérer sur le marché du travail. Une fois la réforme mise en place avec succès, les populations ciblées conservent généralement un désavantage sur le marché de l'emploi, par rapport aux populations non ciblées. L'efficacité de la réforme ne doit en aucun cas se mesurer en comparant la situation finale (après réforme) des deux populations sur le marché du travail, mais en mesurant la réduction de l'écart entre les deux populations, entre les situations avant et après réforme (voir, par exemple, Alibay *et al*., 2005).

La littérature internationale a depuis longtemps recensé les difficultés et les écueils à éviter pour mesurer précisément l'efficacité de réformes ciblées. Des techniques de collecte et d'exploitation de données ont été mises au point pour résoudre ou contourner ces difficultés et écueils (voir Blundell et Macurdy, 1999 pour une revue des différentes techniques dans le cas de l'offre de travail). Chaque technique comporte ses avantages et inconvénients, et se trouve plus ou moins adaptée à l'évaluation d'une politique publique donnée.

Les données administratives (recensement, fichier de bénéficiaires de la caisse d'allocations familiales, déclarations administratives de données sociales) comportent des informations officielles de nature administrative, utilisées pour calculer les droits (impôts ou subventions) sur les comportements effectifs des individus, tels qu'enregistrés par l'Administration. Ces données sont donc en général fiables et portent sur des échantillons de grande taille, exhaustifs ou représentatifs, mais elles sont souvent pauvres en informations sur les caractéristiques socio-économiques (sexe, âge, éducation, composition familiale) susceptibles d'expliquer les comportements, et sont souvent difficiles d'accès (confidentialité). Les

difficultés d'accès aux données administratives et la lourdeur des démarches nécessaires pour y accéder expliquent que, jusqu'à un passé récent, leur utilisation ait souvent été limitée au cadre de projets de recherche ciblés telle que l'ACI (Action Concertée Incitative) Jeunes Chercheurs dirigée par Nathalie Picard en 2002-2005, qui a nécessité la mise en place d'une conventions de recherche avec la Caisse d'Allocations Familiales de la Réunion pour mesurer les effets des différentes réformes du RMI et de l'API sur l'offre de travail des bénéficiaires (voir Alibay *et al*, 2005 et 2006).

De plus, les données administratives concernent généralement une thématique précise (éducation ou offre de travail ou logement ou transport) et ne favorisent donc pas l'analyse des liens entre plusieurs décisions ou de l'ensemble des déterminants d'une décision.

Les données d'enquêtes sont généralement plus riches en termes de caractéristiques socio-économiques et couvrent souvent plusieurs thématiques, mais portent, en contrepartie, sur des échantillons plus petits. Leur fiabilité est parfois mise en cause car elles sont de nature déclarative et/ou rétrospective. Cependant, elles permettent d'appréhender en détail une ou plusieurs thématiques. Les enquêtes dites de préférences révélées analysent l'expérience des répondants, alors que les enquêtes de préférences déclarées s'intéressent aux choix hypothétiques du répondant face à des situations inédites. L'approche en termes de préférences déclarées permet de mettre le répondant face à des situations bien définies et parfaitement contrôlées par le concepteur de l'enquête, ce qui permet de mesurer précisément les préférences du répondant pour chacune des caractéristiques pertinentes ou chacun des aspects à analyser. En revanche, dans le cas des préférences révélées, les choix du répondant sont influencés par de nombreux paramètres qui varient simultanément et qui sont alors difficilement contrôlables et mesurables par le concepteur de l'enquête. Ceci rend très difficile la mesure des préférences du répondant pour chacun des paramètres. Par exemple, lorsqu'une personne décide d'utiliser les transports en commun plutôt que la voiture individuelle pour un trajet donné, ce choix peut être motivé simultanément par plusieurs paramètres. Cette décision peut être liée non seulement à une préférence intrinsèque pour les transports en commun ou à une longueur de trajet différente entre les deux modes, mais aussi à la variabilité du temps de trajet, au coût monétaire du trajet, au niveau de confort dans chacun des deux modes ou encore à la commodité des horaires proposés par les transports en commun pour le trajet considéré. La méthodologie des préférences déclarées permet de construire des scenarii faisant varier indépendamment chacun de ces paramètres. Voir, par exemple, l'enquête Mimettic décrite dans le rapport du projet MobMen sur la mobilité des ménages, conduit par Nathalie Picard pour le Predit (Picard *et al.*, 2012).

Toutefois, les détracteurs de la méthode des préférences déclarées lui reprochent un manque potentiel de fiabilité des réponses, soit par manque d'incitation, soit par manque de réalisme des scenarii présentés s'ils sont trop éloignés des situations effectivement vécues par le répondant. Pour résoudre le problème d'incitation, l'économie expérimentale préconise, dans l'élaboration du protocole d'enquête, de rémunérer les répondants en fonction de leurs réponses, afin de les inciter à révéler leurs préférences, comme dans de Palma *et al*. (2011). Toutefois, la psychologie expérimentale considère que de telles incitations sont contre-productives et qu'une bonne explication des principes et de la finalité de l'enquête constitue une meilleure incitation à fournir des réponses sincères (voir de Palma *et al*., 2014).

Lorsqu'un chercheur ne trouve pas dans les sources disponibles en France (données administratives ou données d'enquêtes, de préférences révélées ou déclarées), même en les appariant, les données pour analyser la politique qui l'intéresse, il peut soit travailler sur un autre pays disposant de données mieux adaptées ou plus facilement disponibles, soit travailler sur des données moins bien adaptées qui risquent de dégrader la pertinence de ses recherches, soit constituer lui-même ses propres bases de données, en élaborant ses propres enquêtes. Cette dernière solution peut impliquer des coûts et des délais très importants, mais c'est en général la seule solution pour analyser des politiques ou des produits nouveaux qui ne sont pas encore disponibles. La méthodologie est alors nécessairement celle des enquêtes de préférences déclarées.

1.2. L'apport des données de panel

Dans les données administratives comme dans les données d'enquêtes, les paramètres influençant les choix du répondant sont généralement endogènes, dans le sens où ils dépendent des préférences du répondant. Par exemple, on peut estimer la valeur du temps de trajet domicile-travail en interprétant le mode choisi en le comparant aux temps de trajet moyen calculé pour chacun des modes disponibles. Cependant, la valeur même du temps de trajet pour chaque mode dépend de la valeur du temps intrinsèque du répondant : les individus ayant une valeur du temps élevée choisissent préférentiellement leurs localisations résidentielle et professionnelle proches l'une de l'autre de façon à minimiser le temps de trajet domicile-travail, alors que les individus ayant une valeur du temps plus faible accordent plus de poids à d'autres critères tels que le prix de l'immobilier dans leur choix de localisation résidentielle. L'endogénéité du temps de trajet génère des biais dont l'ampleur peut être très importante, comme cela est mis en évidence dans de Palma et Picard (2005) ou dans Picard *et al*. (2013).

L'une des méthodes les plus efficaces pour résoudre les problèmes d'endogénéité, en particulier dans le cas de l'évaluation des politiques publiques ciblées, s'appuie sur l'approche de différences en différences, telle que celle discutée dans Blundell et Macurdy (1999) : différences dans le temps des différences entre individus. Cette approche nécessite des données de panel qui permettent de suivre le comportement du même individu dans le temps et ainsi de calculer ces différences. S'il existe un effet individuel non observable qui influence les décisions individuelles de la même façon à chaque date, alors cet effet disparaît par différence. Cette approche permet donc de purger les effets individuels dans la mesure de l'impact des réformes et autres politiques publiques et ainsi mesurer leur véritable effet. L'approche de différences en différences reste néanmoins plus difficile à mettre en œuvre avec des variables discrètes - telle que l'activité – qu'avec des variables continues. Le modèle Logit à effet fixe, proposé par Chamberlain (1984) permet de contourner les difficultés dans le cas de l'analyse d'une variable binaire (2 valeurs possibles), comme l'illustre l'exemple ci-dessous.

Le premier exemple concerne l'effet des politiques ciblées en matière de minima sociaux. Une collaboration avec la CAF (Caisse d'Allocations Familiales) de la Réunion a permis de constituer un panel de bénéficiaires des minima sociaux (RMI, API), allocations logement et autres prestations familiales. Ce panel nous a permis de mesurer l'effet différencié de la réforme Aubry de janvier 1999 concernant le cumul des minima sociaux et des revenus d'activité sur l'emploi des bénéficiaires des minima sociaux à l'île de la Réunion (Alibay *et al*, 2006). Cette loi aligne le régime d'intéressement des allocataires API sur celui des allocataires RMI. Les bénéficiaires du revenu minimum d'insertion (RMIstes) ne subissent qu'un changement léger dans les modalités de l'intéressement, alors que les bénéficiaires de l'allocation parent isolé (APIstes) bénéficient pour la première fois en janvier 1999 d'un cumul possible entre revenus d'activité et allocations. De ce fait, on peut s'attendre à ce que la reprise d'activité des APIstes (le groupe traitement) progresse relativement plus vite que celles des RMIstes (groupe contrôle) suite à la mise en vigueur de la loi Aubry, toutes choses égales par ailleurs. On peut toutefois craindre un fort biais de sélection endogène de l'échantillon des APIstes : en dehors du dispositif, ces derniers auraient en effet, après réforme, de plus faibles chances de travailler que les RMIstes, en raison de la présence d'enfant(s) en bas âge dans leur foyer, frein considérable à un retour à l'emploi. Les résultats montrent effectivement que, lorsque l'on corrige le biais de sélection endogène de l'échantillon par un modèle Logit à effet fixe, la réforme a un effet significativement plus important sur la population traitée (APIstes) que sur la population de référence (RMIstes). En revanche, les résultats d'un modèle de type Logit simple ou même Logit à effet aléatoire suggéreraient plutôt (à tort) que la réforme a un effet négatif sur l'activité des APIstes. Une analyse similaire (Alibay *et al*, 2005) a permis de mesurer l'effet réel de la mise en place de l'intéressement Aubry sur l'activité des bénéficiaires des minima sociaux à la Réunion.

Pour mesurer l'effet de politiques ciblées, il est donc indispensable de non seulement disposer de données appropriées (données de panel), mais aussi d'utiliser des outils économétriques adaptés (de type approche de différences en différences).

1.3. Le potentiel des données existantes

Il y a quelques années, Picard *et al.* (2010) écrivaient : « Les économètres peinent à la recherche de données longitudinales sur les ménages […], les panels sont rares en France […]. Aussi, les économètres sont-ils souvent conduits à utiliser des panels étrangers offrant de plus grands échantillons et de plus longues périodes de suivi ». La situation s'est améliorée depuis, bien que de façon contrastée selon les domaines.

Aujourd'hui, la mise à disposition de données françaises et étrangères pour la recherche en sciences sociales passe par le réseau Quetelet (http://www.reseau-quetelet.cnrs.fr). Ce réseau centralise les données collectées par de grands organismes producteurs d'enquêtes ou de bases de données administratives tels que l'INSEE (Institut National de la Statistique et des Études Économiques), l'INED (Institut National d'Etudes Démographiques), la DEP (Direction de l'Évaluation et de la Prospective), la DEPP (Direction de l'Évaluation, de la Prospective et de la Performance), l'OVE (Observatoire de la Vie Étudiante), le CEREQ (Centre d'Études et de Recherche sur les Qualifications), la DREES (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques), la DARES (Direction de l'Animation de la Recherche, des Études et des Statistiques), le CEREMA (Centre d'Études et d'expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement) et la DRIEA (Direction Régionale et Interdépartementale de l'Équipement et de l'Aménagement). Suivant la nomenclature de l'INSEE, les thématiques couvertes sont aussi variées que l'économie, la conjoncture, les comptes nationaux, la démographie, les revenus, le pouvoir d'achat, la consommation, les conditions de vie, la société, le marché du travail, les salaires, les entreprises, les secteurs d'activité, l'aménagement du territoire, les villes et quartiers, le développement durable ou l'environnement. La plupart des données agrégées sont en accès libre ou ne nécessitent que des formalités légères et rapides (voir http://www.reseau-quetelet.cnrs.fr/spip/article.php3?id_article=2). Cependant, pour pouvoir accéder à des données individuelles détaillées, il est nécessaire de demander des autorisations au Comité du Secret Statistique et l'accès se fait via le CASD (Centre D'accès Sécurisé Aux Données, voir section 5 de ce chapitre).

Même si la France ne dispose pas de panels aussi larges que le PSID américain (Panel Study of Income Dynamics), le BHPS britannique (British Household Panel Survey) ou le GSOEP allemand (German Socio-economic Panel), elle dispose toutefois de versions françaises de panels internationaux tels que les panels européens des ménages (ECHP 1994-2001 et EU-SILC à partir de 2001). Ces panels sont assez diversifiés, mais de taille beaucoup plus modeste que les panels américains, anglais ou allemands précédemment cités (en terme de nombre de répondants et de suivi dans le temps). Il existe en revanche en France des panels thématiques très intéressants (certains sont décrits à la section suivante) et un effort particulier a été consacré, au cours des dernières décennies, à la constitution et à la mise à disposition de données administratives, souvent couplées à des données d'enquête. Ces données sont suffisamment riches pour avoir alimenté de nombreuses études et publications internationales de très haut niveau, comme cela est illustré, par exemple, dans la section 1.4

1.4. Concernant le marché du travail

L'un des exemples le plus intéressants pour illustrer l'intérêt de l'utilisation de données individuelles sensibles dans l'étude du marché du travail est fourni par une série de travaux de Pierre-Philippe Combes, Gilles Duranton, Laurent Gobillon, Diego Puga et Sébastien Roux, dont l'un a été publié dans la très prestigieuse revue *Econometrica* (Combes *et al.*, 2012 ; voir aussi Combes *et al.*, 2011). Ces travaux concernent la mesure des effets d'agglomération, c'est-à-dire l'augmentation de productivité induite par le déplacement d'un salarié ou d'une entreprise d'une zone peu dense vers une zone dense. Le terme « induite » signifie que l'on considère l'effet de la migration sur la productivité. La difficulté

pour mesurer ce lien de causalité réside dans la causalité inverse, ou plus précisément dans la sélection endogène des migrants, que ce soit au niveau du salarié ou au niveau de l'entreprise. En effet, les salariés ou les entreprises qui décident de migrer vers une zone dense sont a priori ceux qui sont les plus susceptibles de bénéficier des effets d'agglomération. Une simple régression de la productivité sur la densité mesure alors autant les véritables effets d'agglomération, pertinents pour l'évaluation de politiques publiques (l'effet de la densité sur la productivité) que la sélection endogène de l'échantillon de migrants (l'effet de la densité sur la décision de migration ou de relocalisation).

Pour étudier ce biais de sélection endogène, Combes et ses collaborateurs utilisent des données de panels d'individus et/ou d'entreprises, issues d'un appariement entre l'enquête emploi (niveau individuel) et les DADS (niveau entreprises/établissements). Ceci leur permet de contrôler les effets individuels à la fois au niveau du salarié, de l'entreprise et de la localisation géographique. Ils montrent que, lorsque les effets fixes de l'individu, de l'entreprise et/ou de la localisation ne sont pas correctement pris en compte, cela conduit à une surestimation des effets d'agglomération dont l'ordre de grandeur varie de 20% à 50%.

2. Accès aux données

À la différence des sciences dites dures, l'expérimentation en sciences sociales et en économie est rarement possible et les chercheurs dans ces domaines utilisent des données recueillies au moyen d'enquêtes ou disponibles dans les dossiers administratifs. L'accumulation et la réplication dans le temps et/ou dans l'espace de ces données permettent aux chercheurs de valider expérimentalement leurs théories. Dans ce contexte, le *Big Data*, de grandes bases de données contenant des informations individuelles, s'impose comme une source de données incontournable pour l'étude des comportements des individus dans leur environnement social.

Globalement, ces grandes bases de données *Big Data* ne sont utilisées que partiellement car l'accès complet est strictement limité pour des raisons de confidentialité.

Ces données, souvent produites à partir de sources privées ou administratives, offrent de nombreuses possibilités en termes d'analyse, mais restent difficiles à anonymiser. Produites en France par plusieurs organisations ou institutions, ces données très détaillées offrent un extraordinaire potentiel pour l'étude statistique, mais restent néanmoins largement inexploitées et peu accessibles à la communauté scientifique. Ces bases de données proviennent de plusieurs sources : les statistiques officielles recueillies par les organismes statistiques (par exemple l'INSEE en France), les ministères ou les données collectées par des acteurs privés (localisation, téléphones mobiles, énergie..). D'une manière générale, de telles données servent à créer de l'information transparente et utile pour les citoyens et décideurs politiques; constituant ainsi un bien public qui profite à la société dans son ensemble. En particulier, les décideurs ont besoin de ces informations afin d'évaluer les programmes existants et de concevoir ou de mettre en œuvre de nouvelles dispositions. Toutefois, ces informations sont hautement confidentielles et sont couvertes par un ensemble de règlements et de lois très strictes. En effet, ces données peuvent concerner tout « agent » du domaine économique et social – que ce soient des individus, des ménages ou des entreprises - et correspondent à des sujets aussi divers que le revenu, le transport, le logement, le patrimoine, le dossier médical, la santé, les aspects sociaux-démographiques, l'emplacement géographique, l'éducation ou encore la carrière professionnelle. À cet égard, la divulgation de renseignements statistiques individuels, appelés micro-données dans le reste de ce texte, en dépit de leur intérêt crucial pour la communauté de la recherche, soulève un certain nombre de questions en raison de leur nature sensible et confidentielle.

Comme beaucoup de pays en Europe ou dans d'autres parties du monde, la France a mis en place des conditions qui permettent un accès particulier pour les chercheurs via le réseau d'archives de données (le « Réseau Quetelet ») à un nombre croissant de micro-données anonymisées. Une partie importante

de ces données sont individuelles et proviennent d'institutions parrainées par le gouvernement telles que l'INSEE et les services statistiques des ministères. L'usage de ces micro-données anonymisées s'est accentué ces dernières années en économie, dans les transports, dans la santé et les sciences sociales. En France, les chercheurs peuvent travailler avec des données statistiques à travers plusieurs canaux :

1. ils peuvent utiliser des données agrégées, totalement anonymisées, qui sont disponibles sur les sites internet officiels (ministères et l'INSEE par exemple) dont l'utilité reste très limitée en raison de leur faible niveau de détail. De toute évidence, ces données qualifiées d'*open data*, fréquemment mises en avant, ne peuvent pas constituer une source acceptable pour couvrir tous les besoins de la recherche. Aussi, de nouveaux modes d'accès contrôlés sont apparus au niveau national pour permettre aux chercheurs d'accéder et de travailler sur des données plus détaillées, tout en assurant le plus strict niveau de sécurité dans le respect des cadres juridiques protégeant la confidentialité.

2. Des fichiers spécifiques sont conçus à des fins de recherche, les "Fichiers de Production et de Recherche" FPR (ou *scientific use files*) et des tableaux sur mesure distribués aux chercheurs par le Réseau Quetelet, l'archive de données française, en collaboration avec l'Insee, plusieurs ministères et autres producteurs de données. Cependant, très souvent, ces micro-données anonymisées ne suffisent pas à la recherche scientifique dont l'exigence en termes de détail est de plus en plus importante. La recherche de pointe au niveau national et international est à ce prix.

3. Les chercheurs peuvent depuis peu, et ce dans des conditions très encadrées, accéder aux données les plus détaillées et donc confidentielles dans un cadre juridique et technique très strict. Comme nous l'avons vu, l'anonymisation des données confidentielles s'accompagne d'une perte importante d'informations qui peuvent être essentielles dans l'analyse statistique. Ne pas anonymiser les données complexifie la question de la protection de la confidentialité qui constitue une préoccupation croissante dans toutes les sociétés européennes, du citoyen aux autorités chargées de la protection de la vie privée (voir le débat en cours sur le futur règlement de l'Union Européenne pour la protection de la vie privée). La prise de conscience croissante des besoins de protection de la vie privée est devenue encore plus générale : les informations contenues dans les "Fichiers Production et Recherche» (directement accessibles par les chercheurs) sont devenus moins détaillées qu'il y a encore une dizaine d'années. Pendant longtemps, seul un nombre très limité de chercheurs ont réussi à obtenir l'accès à des micro-données dans les locaux et sous le contrôle des producteurs de données, le plus souvent les instituts nationaux de statistique et des services statistiques des ministères.

3. L'accès aux micro-données très détaillées et confidentielles

Les *Big data* constituent une nouvelle source de micro-données très détaillée qui sera d'une importance croissante pour la recherche scientifique et l'évaluation des politiques publiques dans un proche avenir. Cependant, l'utilisation de ces données s'accompagne de problèmes spécifiques en raison de leur abondance, leur précision et de leur fréquence de mises à jour via des flux continus d'alimentation. Ces données très détaillées peuvent rapidement devenir volumineuses quand elles comprennent notamment des informations très précises sur la localisation géographique et que, parallèlement, les nouveaux outils d'appariement de données et d'enrichissement d'information augmentent mécaniquement le potentiel de ré-identification, c'est-à-dire la possibilité de retrouver l'identité d'un individu à partir des données qui ne comportent pas d'identifiant direct comme le nom, le prénom ou l'adresse.

De nombreux pays en Europe ont partiellement répondu à cet enjeu par la mise en place d'environnements et d'infrastructures spécifiques qui cherchent à concilier les besoins d'accès aux données confidentielles et les exigences de protection de la confidentialité. Au cours des dix dernières

années, de nouveaux modes d'accès sécurisé ont été mis en place dans de nombreux pays. L'exécution à distance (ou *remote execution*) permet aux chercheurs de soumettre des travaux sans accéder directement aux données. Plus récemment encore, des systèmes interactifs d'accès à distance (ou *remote access*), approuvés par les autorités chargées de la protection de la vie privée, ont été mis au point. Les communautés de recherche ont reçu de tels systèmes de *remote access* favorablement car ils permettent l'accès et le traitement des données de manière interactive ce qui réduit les délais de mise au point et d'analyse. L'impossibilité pour les utilisateurs de télécharger les jeux de données et les contrôles de confidentialité effectués sur les résultats intermédiaires et finaux par les Centres d'accès sécurisé répondent aux exigences de sécurité et de confidentialité applicables à ces données.

De récentes publications de chercheurs étrangers dans des revues scientifiques prestigieuses s'appuient sur de telles données. Par exemple, l'article «Économie à l'âge de *Big Data »*, publié dans Science, a souligné qu'un nombre croissant d'articles scientifiques s'appuyaient sur des données privées ou administratives (8% en 2006 à 46% en 2014).

Pendant longtemps, en France, l'accès à ces micro-données statistiques très détaillées, en particulier les micro-données officielles produites par l'INSEE et les ministères, avait été très limitée en raison des restrictions combinées de la Loi Informatique et Libertés (1978) pour la protection des libertés individuelles et de la Loi sur l'obligation, la coordination et le secret en matière de statistiques (1951). Les changements dans la loi de 1978 en 2004 et la modification récente, en 2008, de la loi sur la statistique de 1951 ont rendu légalement possible l'accès aux micro-données indirectement nominatives pour la recherche scientifique. En effet, la modification en 2008 de la loi de 1951 prévoit une exception à la règle du secret statistique pour des accès à des fins de recherche scientifique. Cette modification permet donc aux chercheurs d'accéder aux données sensibles - telles que les données sur les individus, les ménages (comme le recensement) ou les entreprises (par exemple les "DADS") provenant d'enquêtes statistiques ou de documents administratifs à des fins de recherche exclusivement et uniquement à condition de mettre en place un équipement qui permette de garantir la confidentialité des données

La protection physique des données reste une condition nécessaire pour l'accès effectif des chercheurs aux micro-données très détaillées. Plus précisément, comme les micro-données très détaillées sont des données confidentielles car elles permettent techniquement l'identification directe ou indirecte des individus, elles sont bien évidemment soumises à un très haut niveau de confidentialité. Toute forme de fuite de données doit être évitée à tout prix dans la mesure où cela peut avoir un impact important sur la réputation du détenteur des données. La divulgation d'une base de données permettant une identification même indirecte sur un site internet par exemple pourrait rendre les citoyens plus méfiants lorsqu'ils répondent à un questionnaire statistique, voire les faire renoncer à répondre aux enquêtes statistiques. Par conséquent, bien que très utiles à des fins de recherche, ces micro-données ne peuvent pas être librement communiquées aux chercheurs. Des solutions d'accès aux données hautement sécurisés et fiables sont nécessaires comme cela a déjà été évoqué pour permettre aux chercheurs d'accéder à ces micro-données. Contrairement à d'autres pays d'Europe ou d'Amérique du Nord qui ont construit des solutions et des installations spécifiques pour répondre à cet objectif, aucune solution n'était disponible en France lorsque le nouveau cadre juridique a été publié en 2008.

4. Les moyens d'accès

Trouver et maintenir l'équilibre entre un niveau élevé de sécurité et les besoins des chercheurs est un enjeu crucial pour la mise en œuvre pratique d'un système qui permette aux chercheurs de travailler avec des micro-données confidentielles. La demande croissante de l'interdisciplinarité, se traduisant par davantage d'interfaces entre les secteurs contenant des données hautement sensibles - tels que ceux du secteur privé, du secteur de la santé ou de l'administration fiscale et des finances -, rend cet enjeu encore plus crucial.

Les centres sécurisés sous forme locaux «physiques» ont été introduits aux États-Unis durant les années 80, à la suite des travaux pionniers du Cornell Restricted Access Data Center (CRADC) de l'Université Cornell. Ensuite, les centres d'accès sécurisé physique ont été mis en œuvre au Canada par Statistique Canada dans les années 11990 et dans certains pays européens comme les Pays-Bas ou l'Allemagne dans les années 2000. Pour accéder aux données, les chercheurs doivent se rendre dans ces «centres physiques» sécurisés (également appelés *safe center*). Une fois leur travail terminé et vérifié par les opérateurs, ils ne sont autorisés à emporter que des tables de résultats suffisamment agrégés pour se conformer à la confidentialité statistiques (secret statistique). Aussi sûre qu'elle puisse être, cette solution présente des inconvénients évidents (présence physique obligatoire ce qui signifie bien souvent voyager pour un grand nombre de chercheurs, files d'attente) et elle est de plus très coûteuse pour les chercheurs - qui doivent couvrir leurs frais de voyage et de séjour sur place - et elle entraîne des coûts de gestion élevés.

L'exécution à distance (*remote execution*) était une première tentative d'offrir une alternative - encore perfectible - aux systèmes d'accès physique. Même si les chercheurs ne doivent pas être physiquement sur place, ils ne disposent pas d'un accès direct aux micro-données : l'accès est accordé uniquement à un petit sous-échantillon de données fictives avec une structure identique aux données réelles qui permet aux utilisateurs d'écrire et de tester leurs scripts avant de les envoyer aux opérateurs de centres de données. Ces derniers les vérifient et les exécutent pour le compte du chercheur. Il en résulte un processus lourd et long au cours duquel les chercheurs sont bien souvent obliger de présenter plusieurs fois leurs programmes avant d'obtenir leurs résultats intermédiaires qui, de surcroît, peuvent nécessiter davantage de raffinements et d'autres itérations : un processus maintenant considéré comme trop complexe et lourd et donc insuffisant pour répondre aux besoins des chercheurs.

Par conséquent, des expérimentations ont commencé dans les années 2000 pour développer des systèmes qui permettraient un accès distant sécurisé interactif avec un accès direct aux données pour les chercheurs, en particulier aux États-Unis (la « NORC données Enclave » à Chicago) et dans plusieurs pays européens (Royaume-Uni, Danemark, Pays-Bas, Suède). Les systèmes diffèrent dans leurs implémentations techniques et en raison des différentes situations par rapport à la législation nationale de protection des données, mais leur fonctionnement est assez similaire : un logiciel d'accès à distance installé sur le poste de travail du chercheur, une connexion au serveur distant via internet utilisant un canal sécurisé, un ensemble de logiciels scientifiques pour traiter les données et un confinement des données (les données ne peuvent pas être téléchargées).

L'INSEE et son département de recherche (le GENES) ont décidé en 2008 d'étudier la possibilité d'implémenter un centre d'accès sécurisé aux données. C'est ainsi qu'en 2009, la France a développé son propre système d'accès à distance, appelé CASD («Centre d'Accès Sécurisé aux Données ") pour permettre aux chercheurs d'accéder et d'exploiter les données confidentielles principalement issues de la statistique publique.

5. Le Centre d'accès sécurisé aux données (CASD)

Le CASD est un équipement permettant aux chercheurs de travailler à distance, de manière sécurisée, sur des données individuelles très détaillées. Ces données sont confidentielles car elles sont le plus souvent couvertes par un secret : secret professionnel, secret des affaires, secret statistique, secret fiscal ou secret médical. Les données du CASD sont donc toutes d'une grande précision, identifiantes ou indirectement identifiantes, et sont riches en information. La mise à disposition de ces données ne peut se faire que dans des conditions de sécurité très élevée, garantissant la confidentialité ainsi que la traçabilité de ces données. Il s'agit donc de permettre aux utilisateurs de travailler à distance sur les données tout en garantissant qu'aucune donnée ne puisse être récupérée sans avoir été vérifiée par un opérateur. Grâce à une solution technologique, qui a donné lieu au dépôt d'un brevet, le GENES et l' INSEE, ont pu, en 2010 créer un centre d'accès sécurisé aux données offrant à la communauté

scientifique un accès à des données détaillées et confidentielles de manière hautement sécurisée. Le CASD a été conçu pour répondre à deux objectifs principaux : (a) répondre aux fortes exigences de sécurité et d'ergonomie pour la diffusion des données confidentielles dans un contexte budgétaire très contraint ; (b) assurer l'authentification, le confinement et la traçabilité des données pour offrir les garanties de sécurité nécessaires aux producteurs de données confidentielles. En particulier, le principe qu'à tout moment et quoi qu'il arrive, chacun doit prendre ses responsabilités en cas de faute, est la meilleure garantie de sécurité des données, au-delà de son aspect purement dissuasif. Le confinement est un élément essentiel pour garantir la traçabilité des données : une fois les données « à l'air libre », il devient quasiment impossible de les tracer. Le confinement est notamment assuré en empêchant techniquement l'utilisateur de récupérer des fichiers de données détaillées via par exemple un copier/coller, un partage réseau, un accès internet, une clé USB ou une imprimante.

Pour répondre à toutes ces contraintes, L'INSEE et le GENES ont conçu un boitier totalement sécurisé et autonome, la SD-Box, ayant pour unique fonction de donner un accès distant, après authentification par carte à puce et biométrie, à des moyens de traitement de données confidentielles confinées au sein des locaux techniques du CASD. Cet endroit de stockage et de traitement des données est appelé « bulle » ou parfois « enceinte ». Son principe de fonctionnement est qu'aucune donnée ne peut sortir de cette bulle sans contrôle et ceci afin de prévenir tout risque d'évasion de fichiers de données. Le contrôle d'accès de l'utilisateur est réalisé à l'aide d'une authentification forte s'appuyant sur une carte à puce contenant un certificat de sécurité et un lecteur biométrique d'empreintes digitales. Conformément à la loi, ce traitement a fait l'objet d'une autorisation de la commission informatique et liberté (CNIL - délibération n°2014-369). Le système de « bulle » sécurisée hermétique crée une isolation totale du boiter, fonctionnant en circuit fermé, sans contact avec l'extérieur.

Les principes qui ont conduit à la conception de l'équipement CASD peuvent être illustrés par une analogie avec un équipement couramment utilisé par les chimistes : afin de travailler avec un produit dangereux nécessitant une atmosphère particulière, les chimistes utilisent ce qu'on appelle une boîte à gants (*glovebox*). Il s'agit d'une enceinte étanche dans laquelle sont placés les produits et les outils nécessaires à la manipulation de ceux-ci. De longs gants étanches sont intégrés à une paroi transparente de l'enceinte afin que le chercheur puisse glisser ses mains à l'intérieur de ces gants et puisse ainsi interagir avec les éléments qui sont présents dans l'enceinte et ainsi en préserver le confinement. On peut souligner que pour cet usage, l'utilisateur est bien identifié, il peut voir les produits, il dispose des outils pour manipuler les produits et il peut travailler dans un environnement qui lui est familier. Il ne peut cependant pas introduire, ni extraire de produit sans une procédure particulière et finalement, il est protégé du risque de dissémination de ces produits chimiques.

Cette analogie est très illustrative de ce qu'est un centre d'accès sécurisé aux données. L'objectif est de maintenir la confidentialité des données en les confinant et en garantissant l'identité de l'utilisateur même lorsqu'il s'agit d'une utilisation à distance. Pour son travail, l'utilisateur peut voir les données, dispose des outils logiciels pour les manipuler dans de bonnes conditions, dans un environnement qui lui est familier. Pour garantir le confinement, le dispositif empêche techniquement l'utilisateur d'introduire ou de récupérer de lui-même des fichiers de données (par téléchargement, copier/coller, impression ou clés USB). Des procédures spécifiques sont prévues pour insérer des scripts, des données ou des nomenclatures, et pour sortir des fichiers de résultats qui ne contiennent plus de données confidentielles. Le confinement, ainsi qu'une authentification forte, sont nécessaires pour garantir un haut niveau de sécurité et préserver la confidentialité des données.

Le CASD met aujourd'hui à disposition des données des ministères de la justice, de l'éducation, de l'agriculture et des finances pour les données fiscales. Pour ces dernières, il a été nécessaire de modifier la loi et de publier un décret en 2014 pour qu'elles puissent être mises à disposition des chercheurs. Le décret d'application précise explicitement que l'accès ne peut s'effectuer qu'au moyen du centre d'accès sécurisé aux données (CASD). De nouvelles sources sont en permanence ajoutées pour les besoins de la recherche.

À ce jour, les utilisateurs du CASD ont des profils variés : chercheurs, consultants, *datascientist*, médecins travaillant sur des données de santé ou géostatisticiens. La diversité des secteurs et profils intéressés par la méthode d'accès sécurisé est révélatrice du caractère transversal des préoccupations en matière de sécurité des données.

Si initialement la technologie ne permettait pas d'effectuer des traitements *Big Data*, elle a intégré depuis 2013 des outils comme Hadoop ou Spark. Le CASD peut alors s'apparenter à un *datalab* qui offre un accès sécurisé à des données tout en fournissant un grand nombre d'outils de traitement issus des technologies du big data. Le passage au *Big Data* complique souvent les aspects techniques d'accès aux données, d'où la pertinence d'une plateforme qui applique des conditions d'accessibilité précises à chacun des partenaires. Un volume de données très important augmente potentiellement les risques de ré-identification et rend l'anonymisation souvent beaucoup plus complexe à mettre à œuvre. Les traitements et les règles d'accès doivent être adaptés en conséquence.

5.1. Concernant l'éducation

Les recherches sur l'éducation peuvent s'appuyer sur deux types de panels au sens général du terme :

- le panel au sens habituel, qui correspond à un suivi de chaque individu dans le temps, est adéquat pour analyser la progression des élèves dans le système scolaire ou universitaire, ainsi que pour analyser l'effet de l'éducation sur les salaires, l'activité ou l'insertion sur le marché du travail, tout en corrigeant pour l'endogénéité du niveau éducatif choisi par l'individu

- le panel au sens élargi du terme, qui correspond aux différents enfants d'une même famille. Est adéquat pour analyser les choix d'éducation au sein des familles ; l'effet « individuel » correspond alors à un effet familial.

La DEP (Direction de l'Évaluation et de la Prospective), la DEPP (Direction de l'Évaluation, de la Prospective et de la Performance) et le CEREQ (Centre d'Études et de Recherche sur les Qualifications) produisent régulièrement des panels permettant de suivre les élèves dans leur cursus scolaire, puis dans certains cas dans leur cursus universitaire, puis dans certains cas, dans leur insertion sur le marché du travail.

Dans le cas des différents enfants d'une même famille, les différences d'âge entre enfants permettent de tirer parti de l'évolution de l'offre d'éducation au cours du temps. Si des chocs exogènes considérables ont affecté l'offre éducative, cela permet d'utiliser des techniques d'expérience quasi-naturelle, comme dans Picard et Wolff (2010, 2014). L'analyse des choix d'éducation au sein des familles nécessite de disposer de données décrivant l'éducation de tous les enfants d'une fratrie, ce qui implique en particulier de s'assurer que les parents ont achevé de constituer leur descendance. Mais, lorsque l'on peut raisonnablement supposer que les parents sont trop âgés pour avoir de nouveaux enfants dans le futur, les aînés sont souvent dans une tranche d'âge où ils ont déjà quitté le domicile parental, se retrouvant ainsi exclus de la plupart des enquêtes.

La difficulté pour le chercheur est alors de trouver des données renseignant sur le niveau éducatif de tous les enfants d'une fratrie, y compris ceux qui ont déjà quitté le domicile parental. Dans les pays anglophones, de telles données peuvent être reconstituées grâce aux grands panels que sont le PSID ou le BHPS, mais il est plus difficile de trouver ou de reconstituer de telles données en France. Picard et Wolff (2010) ont trouvé de telles données dans le cas de l'Albanie, ce qui leur a permis d'analyser les conséquences de l'histoire mouvementée de ce pays sur les inégalités intra-familiales d'éducation. Dans le cas de la France, Picard et Wolff (2014) ont eu recours à des techniques de pseudo-panel pour combiner trois enquêtes Patrimoine disponibles au Centre Maurice Halbwachs, afin de travailler sur un historique suffisamment long pour observer des évolutions significatives de l'offre d'éducation. Une bonne profondeur historique est en effet nécessaire pour distinguer l'effet du rang dans la fratrie de l'effet de la date de naissance (qui détermine l'offre éducative). Ils ont alors pu constater que, malgré la

tendance à l'augmentation du niveau éducatif au cours du temps, les aînés sont plus éduqués que leurs cadets, toutes choses égales par ailleurs.

6. Techniques d'appariement : de nouvelles données

Les administrations publiques françaises ont à disposition de nombreux fichiers individuels, établis et mis à jour pour les besoins de leur propre gestion. Ces fichiers sont en général de bonne qualité : fichiers établis pour des usages fiscaux (IRPP, ISF, taxe d'habitation, déclarations de succession et de donation), sociaux (déclarations annuelles de données sociales, fichier des demandeurs d'emploi ou des allocataires de l'assurance chômage, fichier des caisses d'allocations familiales), scolaires (base des élèves ou des étudiants, résultats des examens), de santé (fichiers de l'assurance maladie, des causes de décès).

Ces fichiers sont en eux-mêmes une source très riche d'information pour les études et la recherche. Ils permettent de répondre à de nombreuses questions que se posent les pouvoirs publics ou les parlementaires, au moment de prendre une décision, d'adopter une nouvelle loi ou d'évaluer les effets de mesures déjà prises. Cependant, leur puissance d'explication et d'information est démultipliée dès qu'il était possible de les apparier, c'est-à-dire d'enrichir les informations recueillies pour un individu dans un fichier par des informations disponibles pour ce même individu dans un autre fichier. Certaines études ou évaluations ne sont même envisageables qu'à condition d'effectuer un appariement. C'est le cas pour l'étude des liens entre les revenus salariaux et les revenus de remplacement (chômage, indemnités journalières d'assurance maladie, retraites), des liens entre la trajectoire scolaire et la trajectoire professionnelle ultérieure d'un individu ou encore des liens entre domiciliation et transport. On peut aussi citer des projets de recherche qui concernent l'évaluation a posteriori de réformes ou encore des projets qui s'intéressent à l'estimation complète des coûts engendrés par la prise en charge d'une mesure dans le domaine du transport par exemple.

Les exemples ci-dessus illustrent en partie l'utilité des appariements de fichiers individuels pour concevoir, mettre sur pied et évaluer des politiques publiques dans de nombreux domaines. Les appariements sont bien plus puissants que des enquêtes spécifiquement mises sur pied pour répondre à des questions prédéfinies. De telles enquêtes sont en effet très coûteuses, sujettes à des taux de réponses et à une qualité aléatoires et ne peuvent évidemment porter que sur un échantillon beaucoup plus restreint que les fichiers administratifs, dont une caractéristique notable est leur exhaustivité concernant la population ciblée.

Pourtant, aujourd'hui, il n'existe que peu d'études ou de recherches en France fondées sur les appariements de tels fichiers.

L'appariement peut en réalité se faire de deux manières :

- soit à partir d'informations donnant une simple probabilité d'identification des personnes,
- soit à partir d'un numéro permettant l'identification certaine des individus.

L'identification de la personne dans les deux fichiers peut se faire par exemple à partir du nom, si celui-ci est disponible. Malheureusement cette identification ne pourra être que probabiliste. Il existe en effet de nombreux homonymes pour la plupart des noms, même enrichis du prénom. On peut bien sûr ajouter l'adresse ou la date de naissance pour départager les homonymes. Cela n'est pas toujours suffisant. Et il y a bien peu de fichiers comportant toutes ces informations. La plupart de ceux qui ont été mentionnés ci-dessus ne les comportent pas (par exemple, le fichier des déclarations annuelles de données sociales). Même pour les rares fichiers qui comprennent ces données, il arrive assez souvent qu'il y ait des différences d'orthographe (nom accentué ou non, prénom d'usage ou de l'état civil), si bien que l'appariement à partir de ces informations « naturelles » ne se fait que de façon très exceptionnelle.

En fait, l'appariement qui permet réellement de faire correspondre des individus figurant dans deux fichiers, se fait en général sur un numéro d'identification. Ce peut être le NIR (Numéro d'Inscription au Répertoire national d'identification des personnes physiques) ou un numéro dérivé de celui-ci. Ce numéro (qu'il s'agisse du NIR ou de ses dérivés) ne présente pas les inconvénients mentionnés ci-dessus. Il est unique pour un individu, et il n'y a pas deux individus qui portent le même NIR. La France a, de ce point de vue, un avantage important, puisque l'attribution du NIR, à la naissance, est de très bonne qualité. Celle-ci a été vérifiée au moment de la mise en place de la carte de sécurité sociale électronique et est maintenue à un haut niveau depuis. Aussi, le NIR figure dans un grand nombre de fichiers, ou le lien avec celui-ci peut être plus ou moins facilement effectué.

Cependant, la loi de 1978 relative à l'informatique, aux fichiers et aux libertés, prévoit que l'utilisation du NIR (par exemple pour effectuer un appariement) ne peut être mise en œuvre que si le traitement a été autorisé par un décret en Conseil d'État, après avis motivé et publié de la Commission nationale de l'informatique et des libertés (CNIL), dès lors que ce traitement de données à caractère personnel est mis en œuvre pour le compte de l'État, d'une personne morale de droit public ou d'une personne morale de droit privé gérant un service public.

La plupart des organismes de recherche, notamment universitaires, relèvent donc de cet article, puisqu'ils dépendent d'une personne morale de droit public. Ils doivent alors faire autoriser l'appariement par un décret en Conseil d'État. Dans la pratique, cette exigence s'est révélée insurmontable et aucun organisme universitaire ou de recherche n'a pu obtenir qu'un ministre prenne l'initiative de faire accepter un décret en Conseil d'État pour permettre un appariement dans le cadre d'une étude scientifique. Seules certaines administrations telles que l'INSEE ont pu, grâce à l'appui du ministre dont elles dépendent, franchir l'obstacle représenté par cette exigence.

Cette disposition, prévue pour protéger le citoyen contre un usage intrusif de l'administration dans sa vie privée, a donc comme conséquence de priver l'ensemble de la recherche française de la richesse des fichiers dont dispose son administration, et malgré l'existence d'un identifiant national (le NIR) que l'on peut considérer comme d'excellente qualité.

7. Nouveaux cadres techniques, organisationnels et juridiques pour les appariements

La loi pour une république numérique, promulguée en octobre 2016, vise à simplifier la procédure d'appariement dans le domaine de la recherche scientifique tout en garantissant un haut niveau de protection des données personnelles grâce à la mise en place d'un cadre de sécurité organisationnel et informatique très strict. Pour la recherche scientifique publique, il s'agit de rendre possible les appariements de données en respectant les exigences de sécurité nécessaires pour leur réalisation.

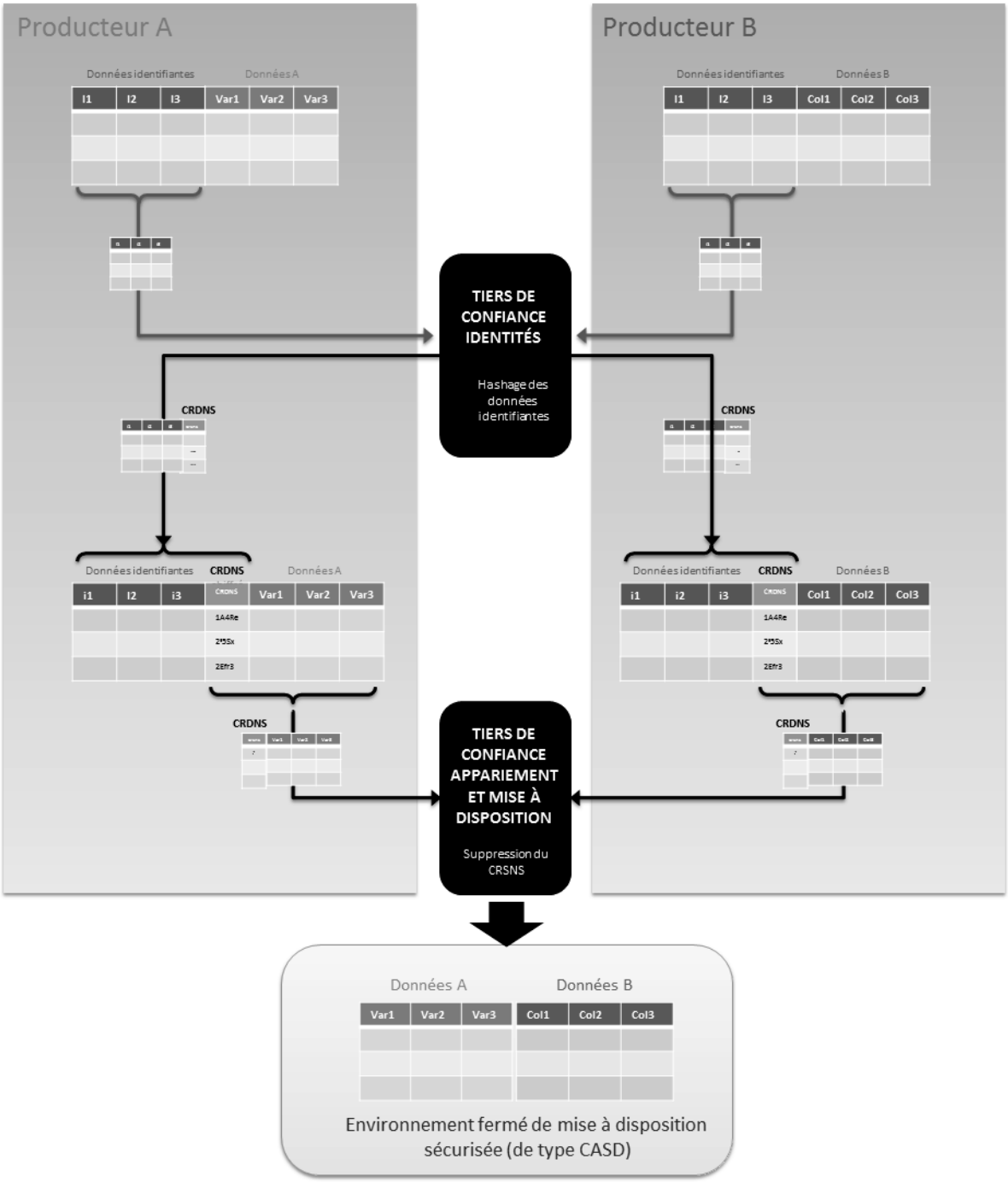
Ce nouveau cadre juridique va assouplir les restrictions pour des traitements portant sur des données utilisant un dérivé du NIR. En effet, le NIR est un indicateur partiellement signifiant (sexe, âge, lieu de naissance) et, lors de sa création, il n'était prévu que pour des usages dans le domaine social. La loi sur la santé a élargi ce domaine à celui d'un identifiant de santé pour la prise en charge des individus à des fins sanitaires et médico-sociales. La CNIL ne souhaite pas généraliser de manière systématique son usage. Cependant, il existe des techniques permettant de faire correspondre à un NIR un autre indicateur appelé « NIR haché » selon un processus qui permet à chaque NIR d'avoir un correspondant unique, tout en évitant au maximum les collisions (le fait que deux NIR différents donnent des NIR « hachés » identiques), mais qui ne permet pas de recalculer le NIR d'origine à partir du NIR « haché ». Le hachage du NIR permettrait de rendre les mêmes services que l'usage du NIR lui-même, mais avec un risque considérablement atténué d'identifier une personne à partir de ce NIR « haché ». Cependant, aujourd'hui, l'utilisation du NIR « haché » est soumise aux mêmes règles que celles du NIR, puisqu'il est issu d'un traitement effectué sur celui-ci. La nouvelle loi va donc assouplir l'usage du NIR « haché »,

tout en maintenant les mêmes règles pour l'usage du NIR lui-même. Dans un tel cas, il faut veiller aux conditions dans lesquelles le NIR a été « haché », afin de s'assurer que le « hachage » est bien irréversible et d'éviter que l'organisme qui a effectué ce « hachage » puisse avoir accès à des informations confidentielles qu'il pourrait relier au NIR d'origine. La nouvelle loi prévoit le recours à un *Key management authority* pour la gestion des clés secrètes de chiffrement et pour la réalisation des opérations cryptographiques. La clé associée à l'opération cryptographique sera spécifique à chaque projet de recherche, c'est-à-dire qu'une nouvelle clé devrait être produite pour chaque projet de recherche. Une clé secrète différente sera créée pour chaque projet de recherche publique, aboutissant à des NIR chiffrés différents pour chaque appariement. Par conséquence, il sera créé, pour chaque appariement, un code recherche dédié non signifiant (CRDNS) spécifique. La procédure d'appariement de données sera la suivante :

1. pour chaque étude, l'Autorité de gestion des clés génère une clé de hachage dédiée,
2. chaque producteur de données attribue un numéro aléatoire unique à chaque enregistrement afin d'obtenir un identifiant unique qui ne se rapporte à aucune autre information spécifique, appelé identificateur "neutre",
3. chaque producteur de données envoie une table qui ne contient que l'identifiant «neutre» et les numéros d'identification à l'Autorité de gestion des clés. En parallèle, ils envoient seulement l'identifiant «neutre» et les micro-données (sans la numérotation identifiant) au tiers de confiance,
4. l'Autorité de gestion des clés hache les numéros d'identification de chaque producteur de données avec la clé de hachage avant d'envoyer l'identifiant « neutre » ainsi que le code de hachage (à savoir sans la numérotation identifiant) au tiers de confiance,
5. à partir de là, le tiers de confiance a reçu toutes les tables nécessaires pour effectuer la mise en correspondance des données sans identification possible.

Figure 5.1. – Organisation de l'anonymisation des données pour un appariement à visée scientifique -

Dans ce schéma, le premier tiers de confiance n'a connaissance que des variables identifiantes et le deuxième tiers de confiance n'a connaissance que de données sans aucune information sur les identités.



Le résultat de l'appariement de deux fichiers, qui sont déjà initialement riches en informations sur les individus, est davantage ré-identifiant que les fichiers initiaux pris séparément. Cela oblige à prendre des précautions particulières quant à sa diffusion. C'est ainsi que la nouvelle loi prévoit qu'un second tiers de confiance soit sollicité pour la réalisation de l'appariement sur le NIR « haché » ainsi que pour la mise à disposition sécurisée des données une fois celles-ci appariées.

La nouvelle loi numérique vise ainsi, pour des travaux de recherche scientifique, à remplacer la procédure de décret en Conseil d'Etat (pris après avis de la CNIL) par une simple procédure d'autorisation auprès de la CNIL. Il faut noter que lorsqu'il s'agit de données issues de la statistique publique ou de données fiscales, un projet de recherche doit d'abord être soumis au Comité du secret statistique pour obtenir la levée du secret statistique ou fiscal. Celui-ci examine notamment la finalité

de la recherche proposée, la qualité des chercheurs, la sécurité mise en place, la pertinence des données pour lesquelles l'accès a été demandé.

Le procédé présenté est déjà mis en œuvre au cas par cas, mais sans harmonisation des procédures. Il est permis par décret en Conseil d'Etat dans le domaine de la santé. La mise en place d'une organisation mutualisée et sécurisée pour la recherche scientifique facilitera la réalisation des appariements et réduira les coûts associés.

L'impact d'une telle mesure sera considérable pour la communauté des chercheurs et, par extension, pour l'ensemble des pouvoirs publics qui bénéficieront des analyses scientifiques et objectives alors réalisables.

Conclusions

Le mouvement de fond actuel d'ouverture de l'accès aux données permettra de nouvelles possibilités en matière de recherche scientifique notamment en ce qui concerne l'évaluation des politiques publiques. L'impact des récentes lois (la loi pour une République numérique, loi de santé en 2016) sera considérable pour l'élargissement de l'éventail des sources accessibles aux chercheurs ainsi que pour les appariements. Il faudra cependant que les modalités d'application pratiques en France soient correctement mises en œuvre avec une forte implication du monde de la recherche. Il est important que la diffusion des données soit réalisée par un organisme tiers, distinct des producteurs de données car ceux-ci, dont ce n'est en général pas le cœur de métier, ne peuvent pas investir lourdement dans une infrastructure dédiée pour offrir ce service dans de bonnes conditions pour les chercheurs (délais, disponibilités, ergonomie et logiciels). Il est aussi important qu'un service d'accès aux données confidentielles soit le plus largement mutualisé entre plusieurs producteurs de données afin d'en minimiser les coûts d'investissement et de fonctionnement et, surtout, de rendre possible au sein d'un même environnement de travail l'utilisation de données provenant de plusieurs producteurs (par utilisation conjointe ou par appariement). On observe dans de nombreux pays étrangers que le développement de plusieurs centres d'accès sécurisé en silos rend presque impossible la réalisation d'appariements de plusieurs sources de données provenant de différents producteurs de données. Historiquement, de nombreux centres d'accès aux données ont été créés sous forme de centres physiques avant de devenir des centres d'accès à distance. Le passage par un tiers spécialisé et indépendant, en lien avec le monde académique, permet de formaliser et d'homogénéiser les modalités d'accès entre les producteurs de données et les utilisateurs de ces données. Cette organisation a fait ses preuves depuis 2010 pour la diffusion de données statistiques et administratives, ouvrant le champ à des études innovantes qui s'appuient sur des appariements de plusieurs sources de données de différents producteurs. Le succès rencontré par le CASD est indéniable puisque de plus en plus de chercheurs européens demandent l'accès aux données françaises via le CASD.

Partie en retard en matière d'accès aux données, la France rattrape son retard grâce aux récentes grandes avancées juridiques et techniques accomplies pour l'accès aux données. La communauté scientifique devra dans les années à venir être très vigilante sur les modalités d'application des textes et les modalités pratiques de mise en œuvre afin de réellement confirmer les perspectives nouvelles et prometteuses qui s'ouvrent actuellement.

Remerciements

Cette recherche a bénéficié du soutien du Labex MME-DII (ANR11-LBX-0023-01).

Références

Alibay N., Picard N. et Trannoy A. (2005). “Evaluation des effets de l'intéressement Aubry sur l'activité des bénéficiaires des minima sociaux à la Réunion”, Revue Economique, mai 2005.

N. Alibay, N. Picard, A. Trannoy (2006). Evaluation des effets de l’alignement du RMI et de l’API sur l’emploi des bénéficiaires des minima sociaux. Document de travail du THEMA, 2006-10.

Blundell R and Macurdy T. (1999). ‘Labor Supply: A Review of Alternative Approaches’, *Handbook of Labor Economics*, 3, 1560-1695.

Chamberlain G. (1984). ‘Panel data’, *Handbook of econometrics*, 2, 1248-1318.

Combes P.-P., Duranton G., Gobillon L., Puga D. and Roux S. (2012). ‘The productivity advantages of large cities: Distinguishing agglomeration from firm selection’, *Econometrica*, 80, 2543-2594.

Combes P.-P., Duranton G., Gobillon (2011). The identification of agglomeration economies, *Journal of Economic Geography*, 11(2), pages 253-266.

de Palma A., N. Picard (2005). “Route choice decision under travel time uncertainty”, *Transportation Research Part A: Policy and Practice*, 39(4), 295-324.

de Palma A., N. Picard, A. Ziegelmeyer (2011). “Individual and couple decision behavior under risk: evidence on the dynamics of power balance”, *Theory and Decision*, 70(1), 45-64.

de Palma, A., M. Abdellaoui, G. Attanasi, M. Ben-Akiva, I. Erev, H. Fehr-Duda, D. Fok, C. Fox, R. Hertwig, N. Picard, P. Wakker, J. Walker, M. Weber (2014). Beware of black swans: Taking stock of the description–experience gap in decision under uncertainty, *Marketing Letters*, 25(3), 269-280.

Maxence Guesdon, Eric Benzenine, Kamel Gadouche and Catherine Quantin, “Securizing data linkage in french public statistics”, BMC Medical Informatics and DecisionMaking (2016)

Picard N., Quantin C., Riandey B., Solaz A. (2010). “Les besoins des démo-économistes en matière d’appariements sécurisés”, *Courrier de Statistiques*, 129.

Picard N. *et al.* (2012). PREDIT 09 MT CV 13, Trajets et mobilité des ménages : choix individuels et collectifs, THEMA (2009-2012). http://www.predit.prd.fr/predit4/publication/43694

Picard, N., A. de Palma, S. Dantan (2013). Intra-household discrete choice models of mode choice and residential location, *International Journal of Transport Economics*, XL(3), 419-445.

Picard, N., F.-C. Wolff (2014). Les inégalités intrafamiliales d'éducation en France, *Revue Economique*, 94-6), 813-840.

Picard N., F.C. Wolff (2010). “Measuring educational inequalities: A method and an application to Albania”, *Journal of Population Economics*, 23(3), 989-1023.