LE CENTRE D'ACCÈS SÉCURISÉ AUX DONNÉES (CASD)

UN SERVICE POUR LA *DATA SCIENCE* ET LA RECHERCHE SCIENTIFIQUE

Kamel Gadouche[*]

L'accès des chercheurs aux données individuelles collectées par le Service statistique public constitue un enjeu scientifique majeur. Ces informations très détaillées exigent un très haut niveau de sécurité pour éviter toute dissémination préjudiciable au citoyen, ou toute utilisation par un tiers non autorisé. Pour répondre à ce besoin de sécurité, l'Insee a créé en 2010 le CASD, Centre d'accès sécurisé aux données : les équipes du centre ont conçu un équipement sécurisé, permettant un accès à distance, tout en garantissant une authentification forte des utilisateurs et un confinement des fichiers. Le CASD, désormais autonome, s'est développé au fil du temps, élargissant son champ à d'autres producteurs et à d'autres types de données très détaillées et sensibles telles les données de santé ou les données administratives. Ce service propose de nouvelles solutions à la problématique des appariements et de la reproductibilité des études sur données confidentielles. Le CASD est de plus en plus utilisé par la communauté de la recherche en France, et son expérience originale, bien que relativement récente comparé à celle de ses partenaires étrangers, lui permet de s'ouvrir à l'international.

Giving researchers access to individual data collected by the Official Statistical System constitutes a major scientific challenge. This very detailed information requires a very high level of security in order to avoid any disclosure which would be prejudicial to the citizen, or any use by an unauthorised third party. To meet this security requirement, INSEE created in 2010 the Secure Data Access Centre (Centre d'accès sécurisé aux données, or CASD) whose teams have designed a secure device, allowing remote access whilst ensuring strong user authentication and confinement of the files. The CASD, now autonomous, has developed over time, extending its perimeter to other data producers and other types of highly detailed, sensitive data such as health data or administrative data. This service provides new solutions to the issue of record linkage and reproducibility of research work based on confidential data. The research community in France uses increasingly CASD's services. The experience is original, although relatively recent compared with that of foreign partners : it enables now CASD to expand on an international level.

[*] Directeur du CASD,
kamel.gadouche@casd.eu

Le premier ouvrage entièrement consacré au secret statistique est paru cette année (Le Gléau, 2019). Il présente un panorama complet pour la France, avec quelques aperçus pour des pays étrangers, des mesures mises en place pour encadrer l'utilisation des données collectées pour établir des statistiques. L'accent est particulièrement mis sur la dualité entre la nécessité de collecter des données individuelles et le besoin de sécurité pour en assurer la confidentialité, que ce soit du point de vue juridique ou technique. Pour les statisticiens du Service statistique public, les garanties sont apportées par leur statut et par leur intégration au sein de l'Insee ou des ministères. Pour les chercheurs, dont les besoins en données très détaillées de la statistique publique sont toujours croissants, se pose la question de l'encadrement juridique de l'accès mais aussi la question des garanties techniques de sécurité pour préserver la confidentialité des données. Nous verrons comment le Centre d'accès sécurisé aux données (CASD) apporte une réponse à ces questions et favorise ainsi le développement de l'accès aux données pour la recherche scientifique. Nous verrons également comment les relations avec les producteurs comme avec les chercheurs se sont développées au fil du temps et ont permis la mise en place de nouveaux usages liés aux nouvelles technologies de *data science*, d'appariement ou encore de certification des résultats.

DONNÉES INDIVIDUELLES ET CONFIDENTIALITÉ

Un grand nombre de données individuelles sur les personnes ou les entreprises sont aujourd'hui collectées par l'Insee et les services statistiques ministériels à des fins de statistique publique, par les administrations dans l'exercice de leurs missions, par les entreprises pour leurs besoins de gestion, par des universités à des fins de recherche dans différents domaines, comme celui de la santé. S'y ajoutent de façon croissante, les données individuelles liées à l'utilisation des moyens électroniques (paiement par carte de crédit…) et donc collectées automatiquement. Toutes ces informations couvrent un grand nombre de domaines particulièrement intéressants pour la recherche : revenus, patrimoine, santé, données comptables des entreprises, informations de localisation géographique, parcours scolaires, trajectoires professionnelles, etc.

Même si elles ne sont pas toutes directement identifiantes (nom, ou identifiants tels que le numéro de sécurité sociale, adresse), un très grand nombre de ces données le sont indirectement du fait de leur précision. Certaines, sensibles au sens de la loi, font peser de par leur nombre un risque d'autant plus grand en cas d'identification pour les personnes ou les entreprises concernées. Pour les entreprises, peu d'informations suffisent le plus souvent à les identifier.

LES BESOINS SPÉCIFIQUES DES CHERCHEURS

Différents secrets inscrits dans des règlements ou des lois couvrent ainsi les données selon les domaines : le secret fiscal, le secret médical, le secret pénal, le secret des affaires, le secret professionnel plus généralement. Dans le cas de la statistique publique, le législateur a inscrit dans la loi l'obligation de confidentialité pour les statisticiens. On appelle cette obligation de confidentialité « secret statistique », une version spécialisée du secret professionnel.
Ces différentes dispositions n'ont généralement pas pris en compte initialement la finalité de recherche. Elles ont été progressivement modifiées pour l'intégrer afin de permettre aux chercheurs d'utiliser ces données très riches pour leurs analyses quantitatives.

Les premières avancées en matière d'accès des chercheurs à des données anonymisées de la statistique publique avec le Réseau Quetelet[1] ont en effet montré très vite que ces fichiers moins détaillés, directement transmis aux chercheurs, s'ils marquaient un progrès important, ne permettaient cependant pas de répondre aux besoins de nombre de projets de recherche. La préoccupation grandissante de la Cnil[2] en matière de données personnelles a conduit à davantage restreindre le niveau de détail de ces données, au point parfois de rendre impossible certains travaux en démographie ou en sociologie urbaine par exemple. Parallèlement, de nouvelles méthodes statistiques, associées à des moyens de calcul plus performants, requéraient notamment chez les économistes des données très détaillées à un moment où il devenait possible de mobiliser pour l'analyse de plus en plus de données administratives utiles à l'évaluation des politiques publiques.

Pour répondre à ces besoins, la loi sur le secret statistique, dite loi de 1951[3], déjà plusieurs fois modifiée, notamment pour permettre l'utilisation par les chercheurs des données d'entreprises, a pris en compte en 2008 la finalité de recherche pour les données des personnes et des ménages. Une modification en ce sens pour l'ensemble des données personnelles avait été réalisée dès 2004 dans la loi Informatique et Libertés[4]. Des modifications d'autres dispositions ont également suivi, notamment dans le domaine des données fiscales.

Encadré 1. Quelques repères

Le CASD accueille aujourd'hui environ 500 projets de recherche menés en France (Amiens, Lyon, Marseille, Dijon, Paris…) et à l'étranger (Royaume-Uni, Allemagne, Pays-Bas, Pologne, Espagne, Italie…), ce qui représente environ 350 sites déployés. Ce sont en tout près de 1 500 utilisateurs qui s'appuient sur le CASD pour l'accès aux données confidentielles.

Juin 1999 – Rapport de Roxane Silberman sur l'accès aux données pour la recherche.

Octobre 2007 – L'Insee initie un projet de pilote de CASD à la suite de la revue par les pairs de la mise en œuvre du code de bonnes pratiques de la statistique européenne.

Juillet 2008 – Modification de la loi 51-711 pour permettre l'accès des chercheurs aux données sur les individus et les ménages.

Octobre 2009 – Premier Comité du secret statistique concernant les données sur les individus et les ménages – Annonce de la mise en place de la facturation à l'usage pour couvrir les coûts.

Février 2010 – Mise en production du CASD pour 30 projets.

Janvier 2011 – Le CASD est lauréat de l'appel à projet Équipex (Équipement d'Excellence du programme Investissements d'avenir) et obtient un financement de 4 M€ pour se développer.

Mars 2012 – Création de l'entité CASD au sein du Genes.

Septembre 2014 – Mise à disposition des données fiscales.

Octobre 2016 – Loi pour une République numérique dont l'article 36 étend les compétences du Comité du secret statistique aux données administratives.

Décembre 2018 – Création du Groupe d'Intérêt Public CASD.

1. Réseau français des centres de données pour les sciences sociales, « Quetelet Progedo Diffusion » lui a succédé pour diffuser les données françaises en sciences humaines et sociales auprès de la communauté de recherche.
2. Commission nationale de l'informatique et des libertés.
3. Loi n° 51-711 du 7 juin 1951 modifiée sur l'Obligation, la coordination et le secret en matière de statistiques.
4. Loi n° 78-17 du 6 janvier 1978 relative à l'Informatique, aux fichiers et aux libertés.

Les aménagements juridiques ne suffisent pas à eux seuls à garantir la confidentialité des données. L'accès effectif doit s'accompagner de mesures de sécurité appropriées permettant d'apporter des garanties techniques additionnelles. Ces mesures requièrent un niveau d'exigence élevé, d'emblée plus facile à appliquer au sein de la statistique publique de par sa fonction qu'à l'extérieur de celle-ci. D'où des difficultés qu'il fallait prendre en compte pour garantir l'application de telles mesures dans des établissements, comme les universités et les centres de recherche, avant de leur transmettre les données. C'est dans le but d'étendre le dispositif de sécurité à l'extérieur du Service statistique public, et d'apporter ainsi une réponse au besoin d'accès aux données des chercheurs, que l'Insee et le Genes[5] ont créé en 2010 le Centre d'accès sécurisé aux données (CASD), initialement centré sur l'accès aux données issues de la statistique publique (voir (Le Gléau et Royer, 2011) et ***encadré 1***).

CONFIDENTIALITÉ *VS* OUVERTURE : PREMIÈRES PISTES

Afin de résoudre le conflit entre confidentialité des données et volonté d'utilisation plus large de ces données par les chercheurs, dès les années quatre-vingt-dix, certains pays comme les États-Unis, le Canada, la Grande-Bretagne ou l'Allemagne ont mis en place des centres d'accès sécurisé sous forme de locaux isolés : les utilisateurs doivent s'y rendre *physiquement* pour y travailler avec des contrôles très stricts à l'entrée comme à la sortie des locaux. Toute sortie de résultats ne peut notamment être récupérée qu'après une vérification effectuée par des opérateurs pour garantir le respect du secret statistique. Même si l'accès aux données était possible, il était cependant très contraignant pour les chercheurs de devoir se déplacer, parfois loin, pour accéder aux données.

Pour pallier les inconvénients de tels dispositifs, des expérimentations ont commencé à la fin des années quatre-vingt-dix pour développer des systèmes qui permettraient un accès à la fois distant et sécurisé. Dès 1999, un rapport rendu au ministre en charge de l'Enseignement supérieur et de la recherche (Silberman, 1999) soulignait la nécessité pour les chercheurs d'utiliser des données très détaillées. Il indiquait l'existence de telles expérimentations à l'étranger, notamment à Statistique Québec. À partir des années deux-mille, on trouvait déjà ce type de système aux États-Unis (le *NORC*[6] *Data Enclave* à Chicago) et dans plusieurs pays européens. Le cas du Danemark était souvent cité par des chercheurs du Crest[7] au moment où la question de création d'un tel dispositif commençait à être discutée en France.

Ces systèmes différaient dans leurs implémentations techniques et en raison des différentes législations nationales de protection des données, mais leurs caractéristiques restaient assez similaires. En particulier, ils s'appuyaient sur des logiciels dédiés d'accès distant (*remote access* en anglais). De tels systèmes nécessitent l'installation de logiciels sur des postes de travail non maîtrisés. Ainsi, ils n'apportent pas les garanties de sécurité suffisantes et sont complexes à mettre en œuvre : ils génèrent souvent des problèmes de compatibilité, de conflits d'installation ou de maintenance.

5. Le Groupe des Écoles Nationales d'Économie et Statistique (Genes) est un établissement public d'enseignement supérieur et de recherche rattaché au ministère de l'économie et des finances, dont l'Insee assure la tutelle technique.
6. NORC : *National Opinion Research Center* (Lane et Shipp, 2007).
7. Le Centre de recherche en économie et statistique (Crest) est un centre de recherche dépendant du Genes.

UNE TECHNOLOGIE POUR OUVRIR LE CHAMP DES POSSIBLES

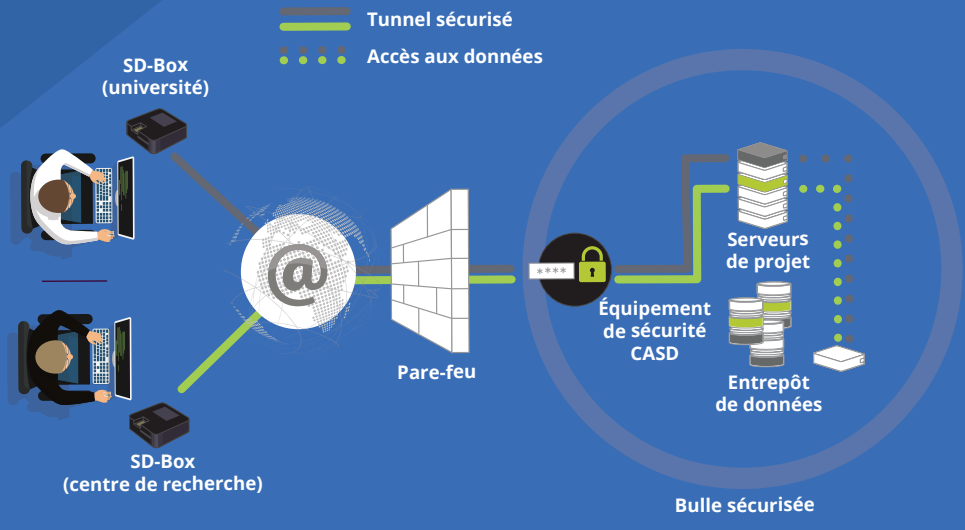
Ces contraintes techniques ont conduit l'Insee et le Genes, poussés également en ce sens par les moyens restreints dont ils disposaient initialement, à concevoir un équipement spécifique, permettant de répondre au besoin d'accès tout en évitant les inconvénients précédemment cités. C'est ainsi que la France a développé son propre système d'accès à distance pour permettre aux chercheurs d'accéder et d'exploiter les données confidentielles principalement issues de la statistique publique.

Au lieu d'utiliser des solutions logicielles tierces, l'équipe projet du CASD a conçu un boîtier informatique spécifiquement pour cet usage particulier d'accès distant sécurisé à des données confidentielles : **la SD-Box.** Une fois en possession d'une SD-Box, l'utilisateur n'a plus qu'à se connecter pour avoir accès à distance à des moyens de traitement sur les données confidentielles confinées au sein de locaux techniques sécurisés. Cet endroit de stockage et de traitement des données est appelé **bulle sécurisée *(figure 1)***. Le principe de cette bulle est qu'aucune donnée ne peut en sortir sans une procédure de contrôle adaptée. L'authentification de l'utilisateur est réalisée à l'aide d'un dispositif s'appuyant sur une carte à puce, contenant un certificat de sécurité, et sur un lecteur biométrique d'empreintes digitales. Conformément à la loi, ce traitement a fait l'objet d'une autorisation

Figure 1. La SD-Box et la bulle sécurisée : un dispositif breveté, conçu par l'Insee et le Genes

Le dispositif technologique du CASD a fait l'objet d'un brevet d'invention déposé par l'Insee le 12 octobre 2009 auprès de l'Institut national de la propriété industrielle (Inpi) sous le numéro FR0957127 (Système informatique d'accès à des données confidentielles par au moins un boîtier distant) et d'une demande d'extensions internationales déposée le 11 octobre 2010, sous le n° WO2011045516A1.

Il a été transféré par l'Insee au Genes à sa création en 2011 et fait désormais l'objet d'une licence exclusive accordée au CASD.



de la Cnil. Le système de bulle sécurisée crée une isolation totale du boîtier, le tout fonctionnant en circuit fermé, sans contact avec l'extérieur : cela permet de garantir une sécurité élevée de bout en bout.

UNE SÉCURISATION EXTRÊMEMENT POUSSÉE ET CERTIFIÉE…

La technologie développée présente l'avantage de pouvoir réaliser une certification de sécurité ISO 27 001, référence internationale en la matière[8], extrêmement poussée parce que chaque composant de la chaîne d'accès est entièrement maîtrisé. Par exemple pour établir une connexion, il faut être localisé dans un établissement lié contractuellement avec le CASD, disposer d'une SD-Box à jour et authentifiée, mais également disposer d'une carte à puce biométrique et d'un compte utilisateur valide. Le pilotage des risques peut être réalisé de manière efficace grâce à un dispositif technique et organisationnel entièrement contrôlé. Ce modèle d'architecture unifiée a été validé par plusieurs audits de sécurité réalisés par des sociétés spécialisées qui ont tous souligné le très haut niveau de sécurité du dispositif.

In fine, le fait d'avoir une technologie qui assure l'authentification, le confinement et la traçabilité des données apporte des garanties essentielles pour la diffusion sécurisée de données confidentielles. Le confinement est un prérequis technique primordial pour garantir la traçabilité des données : une fois les données à l'air libre, elles peuvent être copiées sans limite, à un coût marginal quasi nul. Il devient dès lors impossible de les tracer.

… QUI N'ENTRAVE PAS LES USAGES DES CHERCHEURS

Le confinement des données ne doit pas créer des conditions d'utilisation si restrictives qu'elles compliqueraient considérablement, voire empêcheraient, la réalisation de certains travaux. Les chercheurs doivent pouvoir disposer de tous les outils nécessaires ainsi que d'une puissance de calcul adaptée. Ce dernier point a été une véritable préoccupation du CASD dès la conception de l'architecture et continue de l'être dans sa gestion courante. Contrairement à un service classique de mise à disposition d'environnement de calcul, comme on peut en voir par exemple sur le *cloud*, l'exigence de confinement du CASD n'offre pas la possibilité à l'utilisateur d'installer par lui-même des logiciels. C'est une contrainte forte pour les chercheurs et c'est pour cela qu'elle doit être compensée par une offre large de logiciels scientifiques mis à leur disposition et la possibilité d'en ajouter si nécessaire et dans des délais assez courts. Il en est de même pour la puissance de calcul qui doit être paramétrable en fonction des besoins, du type de traitement et du volume des données.

À la sécurité technique, s'ajoute une sécurité juridique. Avant de pouvoir accéder à l'infrastructure technique, les chercheurs doivent procéder à un ensemble de démarches qui engagent par ailleurs leur responsabilité personnelle.

ACCÈS AUX DONNÉES : PROCÉDURE ADMINISTRATIVE…

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[9] pour obtenir la levée du secret statistique ou fiscal pour les membres de ce projet. Pour les autres

8. La norme ISO 27 001 permet aux entreprises et aux administrations d'obtenir une certification qui atteste de la mise en place effective d'un système de management de la sécurité de l'information.
9. *https://www.comite-du-secret.fr*

données administratives, le comité peut également être saisi par le producteur de ces données.

> *« Vérifier que le projet soumis peut être qualifié de projet de recherche scientifique et que les porteurs du projet sont des chercheurs. »*

À cet égard, il faut rappeler que le statut de chercheur couvre certes des institutions variées (universités, instituts, etc.), mais qu'à la différence de celui de statisticien public, il n'est pas défini par les textes (loi ou décrets). C'est pour cette raison que l'Insee avait décidé d'instaurer le Comité du secret statistique et de lui confier la tâche de vérifier que le projet soumis peut être qualifié de projet de recherche scientifique et que les porteurs du projet sont des chercheurs.

Ce comité, où siègent notamment les producteurs des données et des représentants des chercheurs, examine un ensemble de critères défini dans les textes, dont la finalité de la recherche proposée, la pertinence des données pour lesquelles l'accès a été demandé et la qualification de chercheur du demandeur.

Après cette instruction, le comité émet un avis, lequel est suivi d'une décision de l'administration des Archives nationales, ou du ministre du Budget s'il s'agit de données fiscales. La Cnil intervient également s'il s'agit de données personnelles. *Quelle que soit la procédure, l'accord du service producteur des données est requis.*

… PUIS SÉANCE D'ENRÔLEMENT

Avant tout accès au CASD, les chercheurs, une fois habilités, doivent suivre dans les locaux du CASD à Palaiseau, une session de formation et de sensibilisation dite *séance d'enrôlement* durant laquelle ils sont sensibilisés aux lois de protection de la confidentialité et au respect des règles du secret statistique, à l'image de séances de formation telles qu'il en existe dans plusieurs centres à l'étranger.

Il s'ensuit une présentation des règles de sécurité informatique figurant dans les conditions d'utilisation du CASD signées par chaque utilisateur : accès strictement personnel, obligation de retirer sa carte d'authentification lorsque l'on s'absente (même un court instant) du poste SD-Box…

On y présente aussi les conditions d'hébergement du boîtier, bien que ces conditions fassent l'objet d'un contrat entre le CASD et l'établissement où sera installée la SD-Box. Par exemple, on demande que la SD-Box soit installée dans un local fermant à clé, que l'écran ne soit visible que par son utilisateur, etc. À la fin de la séance, le chercheur obtient sa carte d'accès et il lui appose, suivant une procédure encadrée par les ingénieurs du CASD, ses empreintes digitales ***(figure 2)***.

UNE AUTONOMIE QUI N'EMPÊCHE PAS LE CONTRÔLE

Après l'enrôlement, les chercheurs reçoivent dans leur établissement un boîtier SD-Box. Il leur suffit alors de lui brancher un écran et un clavier et de le connecter au réseau. Ils peuvent immédiatement commencer à travailler et faire leurs analyses avec une réelle autonomie. À cela près que les utilisateurs ne peuvent techniquement récupérer aucun fichier à partir

de leur boîtier ***(figure 3)***. Celui-ci est isolé de tout autre dispositif. L'impression de fichiers, le transfert de fichiers, les opérations de copier-coller sont impossibles. Lorsque leur travail est suffisamment avancé et qu'ils désirent récupérer des fichiers de résultats, les chercheurs utilisent un programme du CASD. Celui-ci, entre autres, poste les résultats dans un espace du serveur réservé à cet usage :

- Dans le cas d'une procédure dite de *contrôle a priori,* les gestionnaires du CASD vérifient que l'utilisateur a bien réalisé le nécessaire pour garantir que les fichiers de résultats satisfont aux règles de confidentialité définies par le producteur des données, et si c'est bien le cas, les lui transmettent.

- Il existe une procédure automatique, sans vérifications manuelles, pour certaines catégories de données, comme les données de santé. Les utilisateurs doivent renseigner un formulaire électronique où ils s'engagent avoir respecté les règles de confidentialité. Le fichier leur est alors transmis automatiquement par notification de messagerie, accompagnée d'un lien de téléchargement sécurisé. Une copie de ces fichiers est conservée par le CASD pendant une durée de cinq ans pour permettre de réaliser des *contrôles a posteriori*.

Dans les deux cas, les contrôles portent exclusivement sur la confidentialité des données, et en aucun cas sur la qualité ou la pertinence scientifique des travaux.

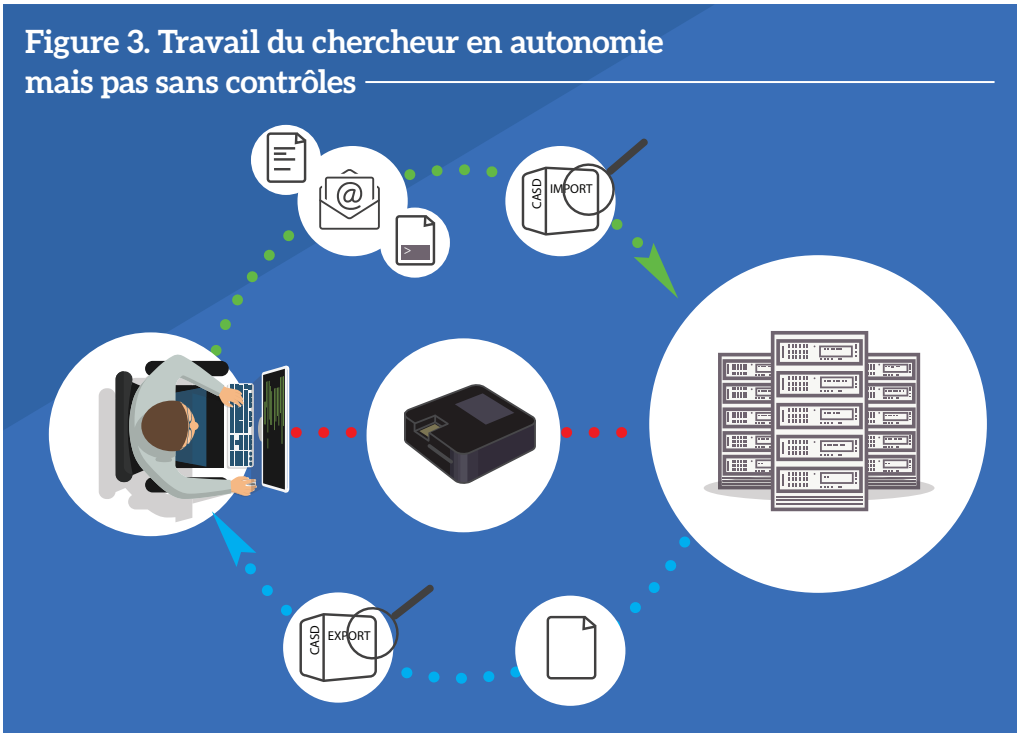
Figure 2. Les séances d'enrôlement engagent la responsabilité du chercheur

LES AVANTAGES POUR LES PRODUCTEURS DE DONNÉES…

Le fait que la diffusion sécurisée des données soit assurée par un tiers comme le CASD permet au producteur, dont ce n'est en général qu'une mission auxiliaire, de ne pas avoir à investir trop lourdement dans une infrastructure pour offrir ce service. Cela permet aussi de mutualiser ce service entre plusieurs producteurs de données afin d'en minimiser les coûts d'investissement et de fonctionnement ; il est à noter qu'il n'y a pas de coût d'entrée, ni même de coût d'exploitation pour un producteur de données souhaitant mettre à disposition sur le CASD ses données à des fins de recherche scientifique. Le CASD se charge également de contractualiser avec les chercheurs ***(figure 4)***. Dans bien des cas il n'est plus nécessaire pour les producteurs de conclure une convention avec les chercheurs pour permettre l'accès à leurs données. Et pour les producteurs qui souhaitent tout de même avoir une convention avec les chercheurs, celle-ci s'en trouve très nettement allégée car ne contenant plus de clauses spécifiques détaillées sur la technique et la sécurité.

Depuis la mise en application du RGPD[10], les garanties de sécurité sont devenues des exigences juridiques fortes. Un modèle standardisé permet de réduire considérablement ces formalités et facilite ainsi la mise en conformité aux exigences des registres de traitements ou des études d'impact.

Figure 3. Travail du chercheur en autonomie mais pas sans contrôles



10. Règlement général de protection des données.

… ET POUR LES CHERCHEURS

Cette configuration offre également le très grand avantage de rendre possible l'utilisation de données provenant de plusieurs producteurs par utilisation conjointe ou par appariement au sein d'un même environnement de travail.

Les chercheurs n'ont d'ailleurs pas attendu longtemps pour en profiter : en 2013, déjà 16 % des projets utilisaient les sources de deux ou trois producteurs, à un moment où celles de l'Insee constituaient encore l'essentiel des données déposées au CASD. Depuis, sur un nombre de projets presque quatre fois plus important, la proportion de ceux utilisant les sources de plusieurs producteurs est montée à 52 % avec désormais des projets incluant les sources de quatre voire cinq producteurs de données. Actuellement, 171 projets utilisent conjointement les données de l'Insee et de la DGFiP.

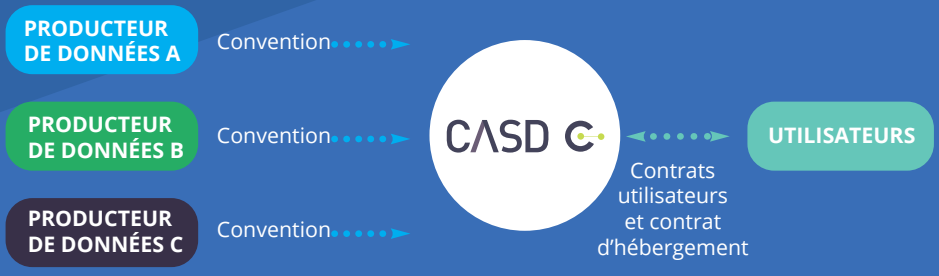
Cette possibilité constitue un avantage pour les chercheurs, comparé à une situation comme celle du Royaume-Uni où les dispositifs d'accès aux données fiscales et aux données de la statistique publique sont différents. En effet, à l'étranger, dans la plupart des pays comparables, le développement de plusieurs centres d'accès sécurisés en silos a conduit à rendre complexe pour les chercheurs la réalisation d'appariements ou l'utilisation conjointe de plusieurs sources de données provenant de plusieurs producteurs. Cette difficulté s'explique notamment parce qu'à l'étranger, historiquement, les premiers centres d'accès sécurisés ont été créés sous forme de centres physiques avant de devenir des centres d'accès à distance.

Ce modèle de spécialisation et de mutualisation présente un avantage indéniable pour l'utilisation des données françaises. À notre connaissance, cette offre est unique au monde à cette échelle. Cela explique que de plus en plus de chercheurs européens demandent désormais l'accès aux données françaises.

DES APPARIEMENTS POUR DÉMULTIPLIER LES POSSIBILITÉS…

Les données disponibles sur le CASD sont en soi des sources très riches d'information pour les études et la recherche. Cependant, leur puissance d'information et d'explication se trouve démultipliée en les appariant, c'est-à-dire en enrichissant les informations recueillies pour un individu dans un fichier par celles disponibles pour ce même individu dans un

Figure 4. L'intermédiation du CASD allège la charge du producteur et facilite la mutualisation des sources



autre fichier. Certaines études ou évaluations ne sont possibles qu'à condition d'effectuer préalablement un tel appariement. C'est le cas par exemple lorsqu'il s'agit d'étudier les liens entre les revenus salariaux et les revenus de remplacement (chômage, indemnités journalières, d'assurance maladie, retraites), entre trajectoire scolaire et trajectoire professionnelle d'un individu ou entre la santé et le travail.

« C'est le cas par exemple lorsqu'il s'agit d'étudier les liens entre les revenus salariaux et les revenus de remplacement (chômage, indemnités journalières, d'assurance maladie, retraites…). »

De tels appariements de fichiers individuels sont particulièrement utiles et même nécessaires pour concevoir, mettre en place et évaluer des politiques publiques dans de nombreux domaines. Ils offrent des avantages par rapport à des enquêtes qui seraient spécifiquement conçues pour répondre à ces questions prédéfinies. De telles enquêtes seraient en effet très coûteuses au regard des moyens dont disposent les chercheurs et ne pourraient évidemment porter que sur un échantillon beaucoup plus restreint que les fichiers administratifs, dont une des caractéristiques est d'être souvent exhaustifs sur la population concernée.

Jusqu'alors, il n'existait cependant que très peu d'études ou de recherches en France fondées sur les appariements de tels fichiers. En réalité, l'appariement qui permet réellement de faire correspondre des individus figurant dans deux fichiers, se fait en général sur un numéro d'identification comme le NIR (Numéro d'Inscription au Répertoire national d'identification des personnes physiques). Celui-ci figure dans un grand nombre de fichiers. Cependant son utilisation était jusqu'en 2017 très restreinte, car elle nécessitait préalablement la publication d'un décret en Conseil d'État autorisant le traitement.

… EN UTILISANT LE «NIR HACHÉ»

Depuis la loi pour une République numérique adoptée en 2016 et son décret d'application paru un an plus tard[11], il est devenu juridiquement possible de réaliser 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[12] a élargi ce domaine à celui d'un identifiant de santé des personnes pour leur prise en charge à des fins sanitaires et médico-sociales. La Cnil reste néanmoins toujours vigilante et préfère limiter son usage. Mais il existe des techniques permettant de faire correspondre à un NIR un autre indicateur, appelé « NIR haché » selon un processus asymétrique qui permet à chaque NIR d'avoir un correspondant unique, mais qui ne permet pas de recalculer le NIR d'origine à partir du « NIR haché ». Le hachage du NIR permet de rendre les mêmes services que l'usage du NIR lui-même, mais avec un risque considérablement atténué d'identifier une personne.

11. Loi n° 2016-1321 du 7 octobre 2016 (article 34) modifiant les articles 22 et 25 de la loi Informatique et libertés en vigueur, et décret n° 2016-1930 du 28 décembre 2016, portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche.
12. Loi n° 2016-41 du 26 janvier 2016 de Modernisation de notre système de santé (article 147).

Pour utiliser le « NIR haché », il faut recourir à un tiers de confiance pour la gestion des clés secrètes nécessaire aux opérations cryptographiques de *hachage*. Une clé secrète différente est créée pour chaque projet de recherche, aboutissant à des NIR chiffrés différents pour chaque appariement. Le résultat de l'appariement de deux fichiers est davantage ré-identifiant que les fichiers initiaux pris séparément. Cela oblige à prendre des précautions particulières quant à sa diffusion. C'est pour cette raison que la loi prévoit qu'un second tiers, comme le CASD, soit sollicité pour la réalisation effective de l'appariement à partir du NIR haché ainsi que pour la mise à disposition sécurisée des données une fois celles-ci appariées ***(figure 5)***.

INNOVER POUR CERTIFIER LES RECHERCHES FONDÉES SUR DES DONNÉES CONFIDENTIELLES…

Depuis l'ouverture sécurisée des données confidentielles, la question de la reproductibilité des calculs, garante de la scientificité de la démarche, s'est posée. Les journaux scientifiques demandent en effet que les chercheurs déposent données et code pour que les résultats publiés puissent être soumis à vérification par des tiers, ce qui pose bien évidemment des difficultés s'agissant de données confidentielles (Pérignon *et alii,* 2019).

Jusqu'ici la vérification des résultats devait se faire après soumission aux relecteurs (*reviewers*) désignés par l'éditeur. Or pour que le *reviewer* puisse faire ces vérifications, il doit suivre la procédure d'accréditation par le Comité du secret statistique et se déplacer au CASD pour se faire enrôler. Ce processus peut prendre des mois. En pratique, aucun *reviewer* n'a eu le temps ni les ressources nécessaires à consacrer à ce type de vérification.

Le CASD a conclu un partenariat avec une agence de certification, Cascad[13], pour mettre en place une solution de certification de la reproductibilité d'une recherche qui s'appuie sur des données confidentielles.

L'agence spécialisée et accréditée vérifie la conformité des résultats, en amont de leur soumission à la revue scientifique. La certification est attribuée à l'issue d'un processus d'évaluation mené par un spécialiste du langage de programmation utilisé par le chercheur, à partir des données sources présentes sur le CASD et de l'ensemble des codes informatiques mis à disposition par le chercheur. Elle doit permettre d'augmenter les chances de publication d'un article dans les revues académiques.

Grâce au soutien du Comité du secret statistique et de l'ensemble des producteurs, un pilote a pu débuter en avril 2019 pour une durée d'un an. Le principe de ce pilote est qu'après habilitation du certificateur, le CASD lui ouvre un accès à un environnement sécurisé pour chaque demande sur la durée nécessaire à la certification. Les environnements ainsi créés sont fermés à la fin de chaque certification. Les programmes et les données associées sont estampillés et archivés de manière chiffrée sur une période de cinq ans.

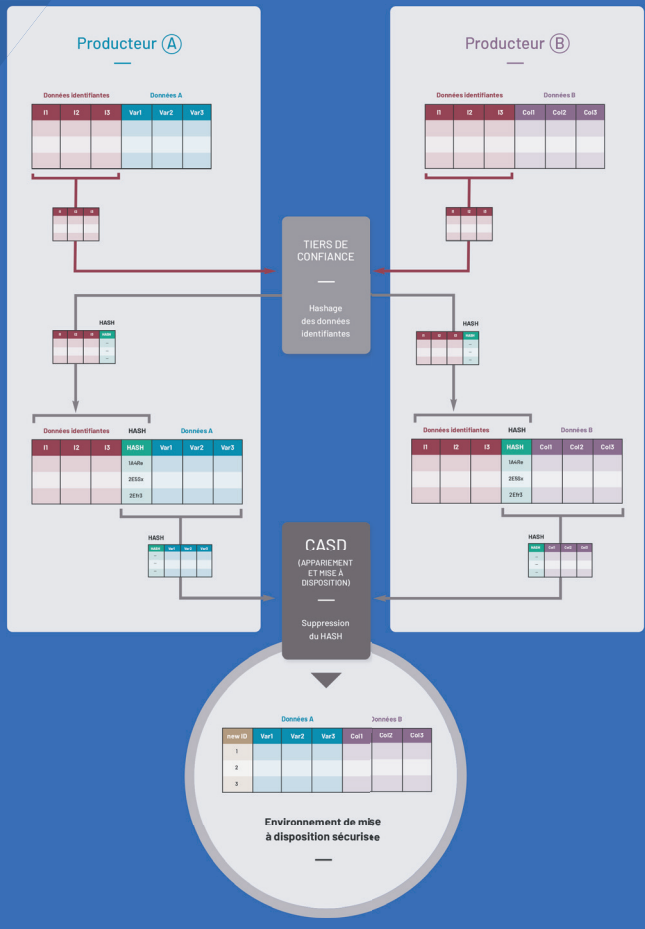
Au moment où la question de la reproductibilité des résultats de la recherche est souvent mise en avant dans nombres de domaines, la systématisation de la certification offerte par ce service permettrait une avancée considérable[14].

13. Cascad *(Certification Agency for Scientific Code and Data)* est une structure d'appui à la recherche, à but non lucratif, financée par différentes institutions françaises dont le CNRS, HEC Paris et l'Université d'Orléans.
14. Voir également la consultation publique lancée en juillet 2019 par la Cnil auprès des chercheurs sur les traitements de données à des fins de recherche scientifique (Cnil, 2019).

Figure 5. Un premier tiers de confiance produit une clé secrète et le CASD procède à l'appariement

Lorsque le CASD est acteur de la mise à disposition sécurisée des données appariées, la procédure est la suivante :

- pour chaque étude, le premier tiers de confiance génère une clé de hachage dédiée à l'appariement ;
- 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 » ;
- chaque producteur de données envoie une table qui ne contient que l'identifiant neutre et le NIR au tiers de confiance. En parallèle, ils envoient l'identifiant neutre (sans le NIR) et les micro-données au CASD ;
- le tiers de confiance hache les NIR des fichiers de chaque producteur de données avec la clé de hachage et envoie le NIR haché accompagné de l'identifiant neutre au CASD ;
- le CASD a alors reçu toutes les tables nécessaires pour effectuer la mise en correspondance des données sans à aucun moment avoir connaissance du NIR et peut les mettre à disposition des chercheurs du projet.



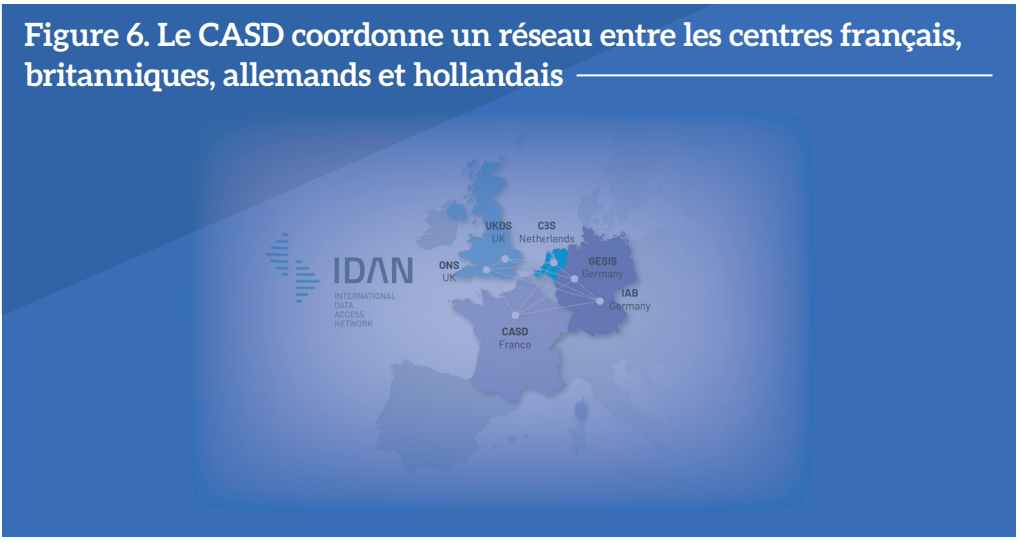
OUVERTURES INTERNATIONALES

Sur le plan international, les producteurs de données ont donné leur accord pour un accès à distance transnational aux chercheurs de l'Union européenne et des pays associés de l'AELE[15]. Les chercheurs ont récemment été autorisés à travailler depuis les États-Unis et le Canada, avec quelques conditions supplémentaires[16], sur les données de l'Insee et du ministère de l'Agriculture.

En bonne position de ce point de vue sur le plan international, après avoir participé au projet européen *Data without Boundaries (DwB)*[17], le CASD coordonne la mise en place d'une collaboration entre centres d'accès sécurisé français, britanniques, allemands et hollandais ***(figure 6)*** : l'objectif d'IDAN (*International Data Access Network*) est de faciliter l'accès aux données sécurisées de ces pays pour les chercheurs, leur évitant des déplacements et leur permettant de mobiliser plus facilement les données de plusieurs pays à partir du site de chacun des partenaires du réseau. Ainsi, il devrait être possible d'ici la fin de l'année 2019, à partir de chacun des centres, d'accéder aux données de tous les autres centres.

Créé tardivement par rapport à d'autres grands pays en Europe et en Amérique du Nord, le CASD bénéficie aujourd'hui d'une position de pointe à la fois par sa technologie et par le nombre de données qui y sont déposées. Il a ainsi fait partie des quatre centres sécurisés auditionnés par la Commission américaine en charge du rapport sur l'ouverture des données administratives pour l'évaluation des politiques publiques et la recherche (American Congress, 2017). Le nombre croissant de données qui sont disponibles au CASD témoigne de la confiance créée du côté des producteurs, gage de futurs développements notamment dans le domaine de

Figure 6. Le CASD coordonne un réseau entre les centres français, britanniques, allemands et hollandais



15. AELE : Association européenne de libre-échange.
16. Répondre à au moins une des deux conditions suivantes : soit être citoyen européen, soit avoir un établissement européen impliqué dans le projet.
17. Il s'agit d'un projet dans le cadre du septième programme-cadre pour 2007-2013 de l'Union européenne (FP7), pour la recherche et le développement technologique (Alvheim *et alii*, 2012).

la santé à un moment où l'articulation entre ce domaine et celui des sciences économiques et sociales est de plus en plus important. Du côté des chercheurs, les craintes initialement exprimées devant les contraintes à passer par un système sécurisé ont été très largement compensées par le nombre et la richesse des nouvelles données ouvertes permettant la réalisation de projets importants utilisant les données françaises y compris à l'international.

Encadré 2. Statut et financement du CASD

Le CASD est un équipement labellisé en 2011 Équipex (Équipement d'Excellence du programme Investissements d'avenir) et a bénéficié à ce titre d'un financement jusqu'en 2019 pour se développer.

Le règlement de l'appel à projet Équipex exigeait notamment la mise en place d'un dispositif de facturation pour assurer l'autofinancement du service au-delà de 2019. C'est ce qu'a fait le CASD à partir de 2012 en facturant ses services. La facturation annuelle moyenne aujourd'hui est d'un peu plus d'un millier d'euros par utilisateur. Cette facturation sert à couvrir partiellement les frais d'exploitation, l'autre partie étant couverte par les contributions des partenaires du projet à savoir l'Insee, le Genes, le CNRS, l'École polytechnique et HEC Paris.

À la fin de l'année 2018, les partenaires du projet ont décidé de créer pour le CASD une structuredédiée, en constituant un Groupement d'intérêt public (GIP), personne morale de droit public àbut non lucratif dotée de l'autonomie administrative et financière. La transformation du CASDen GIP dans la continuité du consortium Équipex lui permet de disposer d'un mode de fonctionnementplus souple et adapté tout en apportant des garanties aux partenaires publics en tant que personne publique.

La création d'une telle structure a permis d'inscrire dans les textes les missions du CASD, à savoir: organiser et mettre en oeuvre des services d'accès sécurisé pour les données confidentielles àdes fins non lucratives de recherche, d'étude, d'évaluation ou d'innovation, activités qualifiées deservices à la recherche, principalement publiques, et valoriser la technologie développée poursécuriser l'accès aux données dans le secteur privé.

La mise en application récente du nouveau règlement européen de protection des données (RGPD) laisse présager un besoin accru de sécurisation pour tous ceux qui souhaitent travailler sur des données très détaillées. Le CASD en se dotant d'une structure autonome sera ainsi plus agile pour fournir le service nécessaire à la recherche.

BIBLIOGRAPHIE

ALVHEIM, Atle, BOND, Steve, GADOUCHE, Kamel, GÜRKE, Christopher et SCHILLER, David, 2012. *Report on the state of the art of current SC in Europe.* Septembre 2012. Project DwB, funded under : FP7-Infrastructures

AMERICAN CONGRESS, 2017. *The Promise of Evidence-Based Policymaking.* [en ligne]. Septembre 2017. Rapport de la Commission sur l'Élaboration de politiques fondées sur des données probantes. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.cep.gov/report/cep-final-report.pdf*

CAPELLE-BLANCARD, Günther et BELLANDO, Raphaëlle, 2015. *L'accès aux données bancaires et financières : une mission de service public.* [en ligne]. Juillet 2015. Rapport du groupe de travail du CNIS. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2015_140_acces-aux-donnees-bancaires.pdf*

COMMISSION NATIONALE INFORMATIQUE ET LIBERTÉS (CNIL), 2016. *Communication cadre relative au Big Data.* [en ligne]. 18 février 2016. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.casd.eu/wp/wp-content/uploads/2016-02_Communication_cadre_relative_au_big_data.pdf*

COMMISSION NATIONALE INFORMATIQUE ET LIBERTÉS (CNIL), 2019. *Régime juridique applicable aux traitements poursuivant une finalité de recherche scientifique (hors santé).* [en ligne]. [Consulté le 26 septembre 2019]. Disponible à l'adresse : *https://www.cnil.fr/sites/default/files/atoms/files/consultation_publique_-_presentation_du_regime_juridique_applicable_aux_traitements_a_des_fins_de_recherche.pdf*

GUESDON, Maxence, BENZENINE, Eric, GADOUCHE, Kamel, et QUANTIN, Catherine, 2016. *Securizing data linkage in french public statistics*. [en ligne]. 6 octobre 2016. BMC Medical Informatics and Decision Making. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://hal.inria.fr/hal-01377334/document*

LANE, Julia et SHIPP, Stephanie, 2007. Using a Remote Access Data Enclave for Data Dissemination. In : *The International Journal of Digital Curation.* [en ligne]. 27 juillet 2007. N°1, Volume 2 | 2007, pp. 128 134. [Consulté le 14 octobre 2019]. Disponible à l'adresse : *http://www.ijdc.net/article/download/31/20/*

LE GLÉAU, Jean-Pierre, 2019. *Le secret statistique.* 2 mai 2019. EDP Sciences, Collection Le monde des données. ISBN 978-2-75982-342-0

LE GLÉAU, Jean-Pierre et ROYER, Jean-François, 2011. Le centre d'accès sécurisé aux données de la statistique publique française : un nouvel outil pour les chercheurs. In : *Courrier des statistiques.* [en ligne]. Mai 2011. N°130, pp. 1-5. [Consulté le 14 octobre 2019]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/8288/1/cs130e.pdf*

LOTH, André *et alii,* 2015. *Données de santé : anonymat et risque de ré-identification.* [en ligne]. 6 juillet 2015. Dossiers solidarité et santé, Drees, N° 64. [Consulté le 14 octobre 2019]. Disponible à l'adresse : *https://drees.solidarites-sante.gouv.fr/IMG/pdf/dss64-2.pdf*

DE MONTJOYE, Yves-Alexandre *et alii,* 2018. *On the privacy conscientious use of mobile phone data.* [en ligne]. 11 décembre 2018. Nature, Scientific Data, Comment. [Consulté le 14 octobre 2019]. Disponible à l'adresse : *https://www.nature.com/articles/sdata2018286.pdf*

MOREL-À-L'HUISSIER, Pierre et PETIT, Valérie, 2018. *Rapport d'information sur l'évaluation sur l'évaluation des dispositifs d'évaluation des politiques publiques.* [en ligne]. 15 mars 2018. Assemblée Nationale. [Consulté le 14 octobre 2019]. Disponible à l'adresse : *http://www.assemblee-nationale.fr/15/pdf/rap-info/i0771.pdf*

PÉRIGNON, Christophe, GADOUCHE, Kamel, HURLIN, Christophe, SILBERMAN, Roxane et DEBONNEL, Éric, 2019. Certify reproducibility with confidential data. In : *Science.* [en ligne]. Juillet 2019. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://science.sciencemag.org/content/365/6449/127/tab-pdf*

SILBERMAN, Roxane, 1999. *Les sciences sociales et leurs données.* [en ligne]. Juin 1999. Rapport, Ministère de l'éducation nationale, de la recherche et de la technologie. [Consulté le 14 octobre 2019]. Disponible à l'adresse : *http://media.education.gouv.fr/file/96/5/5965.pdf*

SILBERMAN, Roxane, 2013. Transnational Access to Official Micro-data : The Data without Boundaries European Network. In : KLEINER, Brian, RENSCHLER, Isabelle, WERNLI, Boris, FARAGO, Peter, JOYE, Dominique, 2013. *Understanding Research Infrastructures in the Social Sciences.* Seismo Press, Social Sciences and Social Issues AG, Zurich, pp. 47-66. ISBN 978-3-03777-133-4

FONDEMENTS JURIDIQUES

Loi n° 51-711 du 7 juin 1951 sur l'Obligation, la coordination et le secret en matière de statistiques

Loi n° 78-17 du 6 janvier 1978 relative à l'Informatique, aux fichiers et aux libertés

Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la Protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la Protection des données)

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique

Loi n° 2016-41 du 26 janvier 2016 de Modernisation de notre système de santé