L'accès aux données sécurisées : première conférence du CASD
avril 2016

« Vos données sécurisées au cœur de la datascience : usages et perspectives »

Cette conférence est dédiée à Antoine Frachot, Directeur général du Genes entre 2009 et 2016, appelé à d'autres fonctions.

Contact : www.casd.eu

Le 6 avril dernier, le centre d'accès sécurisé aux données (CASD) a organisé une conférence dédiée aux enjeux de l'accès sécurisé aux données pour la recherche scientifique et la data science.
Cette conférence soulève de nombreuses interrogations sur l'utilisation de données personnelles, confidentielles ou sensibles notamment avec l'émergence des technologies "big data".
Données détaillées de l'Insee, données fiscales, données du ministère de la justice, de l'éducation... et données de santé : comment la sécurisation accrue de l'accès à ces données, toujours plus riches, plus précises, plus qualitatives, élargit le domaine des possibles de la datascience, de la recherche, de l'innovation et de l'évaluation...

Très impliquée dans ces problématiques, Axelle Lemaire, Secrétaire d'État chargée du numérique a ouvert cette journée de conférence.

1000 chercheurs et datascientists utilisent actuellement la plate-forme du CASD pour travailler sur des données sécurisées.

Pour la première fois des utilisateurs ont partagé leur expérience d'utilisation du dispositif CASD mais aussi des résultats qu'ils ont obtenus grâce à l'accès aux données. C'est aussi la première fois que des propriétaires de données ont évoqué leurs démarches de mise à disposition de leurs données pour le monde de la recherche et de la data science.

Sécuriser les données, une condition nécessaire pour leurs utilisations.

Au cœur de l'actualité, la sécurisation des données soulève de nombreuses questions quant à la difficulté technique d'anonymiser les données et aux enjeux juridiques liés à l'accès aux données à caractère personnel.
Les présentations ont mis en lumière ces questions et posé les nouvelles perspectives à explorer pour les années à venir.

L'organisation de la conférence a bénéficié d'un soutien humain et financier de la TGIR Progedo.

CASD

SOMMAIRE

LE CASD EN BREF

Le CASD est un équipement permettant aux chercheurs de travailler à distance, de manière hautement 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, secret médical etc. Les données présentes sur le CASD sont donc toutes d'une grande précision, identifiantes ou indirectement identifiantes, et contiennent une grande richesse d'information. La mise à disposition de ces données ne peut se faire que dans des conditions de sécurité très élevée garantissant leur confidentialité ainsi que leur traçabilité.

CASD

LE CASD EN QUELQUES CHIFFRES

Aujourd'hui, plus de 120 sources de données sont disponibles sur le CASD pour près de 350 projets,
soit près de 1000 utilisateurs en France et en Europe.
Bien qu'il soit très difficile d'être exhaustif, il a été possible de recenser plus d'une centaine de publications dans des revues scientifiques s'appuyant sur des travaux réalisés sur le CASD.
Le CASD est une entité du Genes qui compte vingt trois collaborateurs et qui est composée d'un service de gestion de projets, d'un service statistique, d'un service infrastructure et développement informatique et d'une cellule datascience.

EXTRACT


CASD
CENTRE D'ACCÈS SÉCURISÉ AUX DO
CASD
Vos données
au cœur de la datascience :
usages et perspectives

Ouverture de la journée

Axelle Lemaire

Axelle Lemaire, Secrétaire d'Etat chargée du numérique

Sans doute du fait de mes origines québécoises, je n'aime pas utiliser des termes anglo-saxons mais il est difficile de trouver une traduction pas trop longue en français de celui de « data scientist »: scientifique de la donnée, data chercheur… ? Une chose est sûre, la donnée c'est mon dada ! Un autre chercheur qui partage cet intérêt, de manière quasiment activiste, c'est Thomas Piketty. Il a ainsi publié une contribution très intéressante sur le site de la consultation publique mis en place dans le cadre du projet de loi sur la République numérique. A partir de travaux américains démontrant l'étroite corrélation entre la probabilité d'aller à l'université et les revenus des parents, il pose la question de savoir ce qu'il en serait en France et constate qu'il est à l'heure actuelle impossible de le savoir. Seule la possibilité d'apparier les données concernées le permettrait, et nous donneraient ainsi une vision objectivée des inégalités sociales à l'oeuvre. C'est dire l'importance de l'accès aux données pour la recherche, mais aussi la vie de la cité.

Notre siècle marque ainsi une véritable rupture par rapport au siècle précédent : au XXème siècle, la valeur était créée par les ressources naturelles et reposaient sur la rareté. Désormais, nos ressources sont les données, qui sont au contraire abondantes et réutilisables, d'où l'enjeu de leur accessibilité, leur diffusion et leur circulation. On voit bien que les géants de l'internet ont tous fait de la collecte et du croisement des données le coeur de leur business model. Cependant, aucun secteur économique n'y échappe, de l'agriculture à l'automobile en passant par la banque… C'est également un enjeu pour les administrations pour mieux atteindre des objectifs de politique publique. A partir de là une autre nécessité se fait jour : un décloisonnement entre les mondes politique, économique et celui de la recherche, dont les besoins sont interdépendants. En économie, on parle d' « open innovation ».

Les dispositions du projet de loi sur la République numérique permettront l'appariement de ces données administratives sensibles qui constituent un véritable trésor national. Le mouvement avait déjà été amorcé avec l'ouverture des données fiscales en 2013 puis l'accès facilité aux données de santé dans le cadre de la loi de santé. Il subsiste néanmoins des obstacles assez forts : ainsi le traitement des données contenant le NIR exige un passage devant le Conseil d'Etat, ce qui est loin d'être une promenade de santé ! Il faut convaincre sur le fond comme sur la forme de l'opportunité d'autoriser cet accès.

Pour donner un exemple, on parle beaucoup de l'échec de la formation professionnelle, tout en brandissant des chiffres variables selon les analyses. Seul un appariement des données du marché du travail avec la base des élèves de l'Education Nationale permettrait une véritable évaluation de l'impact de la réforme du bac pro, d'autant que l'on dispose de fichiers administratifs de grande qualité technique. C'est pourquoi nous allons rendre possible un appariement basé sur le NIR, sous condition d'une autorisation de la Cnil et d'un hachage du NIR. Cette décision est le fruit de longues discussions avec le Conseil d'Etat et l'Agence nationale de la sécurité des Systèmes d'information. Il s'agit bien d'un enjeu politique qui doit être porté au débat public. Un certain nombre de députés ont exprimé des craintes relatives à la manipulation de certaines données sensibles à des fins douteuses. Pour les apaiser, il est essentiel d'apporter le plus de garanties possibles, sur le plan de la protection des données personnelles et de la sécurisation par la technologie. Mais la mise en place de solutions technologiques doit aussi se conjuguer avec une mise à disposition des données de la manière la plus simple et la plus rapide possible. Il s'agit ainsi de construire un mécanisme d'autorisation unique pour raccourcir les délais, et ce de sorte que les premiers projets de recherche soient approuvés dès début 2017.

Or ce défi est en train d'être gagné grâce à la mobilisation de l'Insee et du GENES. De nouveaux horizons de recherche vont ainsi s'ouvrir.

Ces dispositions s'inscrivent dans un cadre plus large d'ouverture des données, décliné à trois niveaux. Tout d'abord, il y a la mini-révolution de l'open data par défaut : dorénavant les administrations nationales, comme les collectivités locales ou encore les SPIC collectant les données devront mettre en oeuvre une publication systématique des données produites.

S'y ajoutent des mesures d'open data sectorielles. Tout d'abord, l'accès au registre SIRENE de l'Insee sera gratuit à compter du 1er janvier 2017. On voit bien que si nous avions eu une base SIRENE des entreprises enregistrées au Panama en particulier, peut-être que les scandales qui sont révélés aujourd'hui auraient été connus et peut-être que certains comportements auraient été prévenus.

Egalement, au niveau sectoriel, la base de demandes de valeur foncière de la DGFIP - correspondant aux prix de vente des logements - sera ouverte, ainsi que la publication des données agrégées de consommation individuelle des compteurs intelligents Linky déployés par ERDF. Ce projet de loi introduit aussi une nouvelle catégorie de données : les données d'intérêt général, que la France est le premier pays à inscrire dans le droit. Elles auront un caractère mixte, des données émanant d'entreprises privées pouvant être mobilisées au service de l'intérêt général.

Il existe bien sûr des obstacles potentiels évidents à lever par rapport à la propriété de ces données. Mais je suis persuadée que le développement du partage de celles-ci entre les entreprises, en particulier de secteurs distincts, leur ouvrira des perspectives économiques et commerciales, comme elle en ouvrira pour le grand public. Un exemple, les collectivités locales qui signent des concessions de service public auront tout intérêt à pouvoir accéder aux données récoltées pendant l'exécution du contrat, notamment dans le cadre d'un renouvellement.

En ce qui concerne l'accès sécurisé aux données confidentielles indirectement nominatives très sensibles, sur la santé, les revenus, la famille… ce sont des biens à strictement protéger évidemment. En ce sens, le CASD a permis des progrès certains, avec une technologie à mon avis unique. Mais les chercheurs demandent en outre un accès facilité à d'autres catégories de données, celles de Pôle Emploi, de la CNAF… avec des enjeux scientifiques et sociaux considérables à la clé. Dans cette optique, nous travaillons sur une disposition juridique avec l'espoir de pouvoir intégrer un amendement au projet de loi d'ici son passage devant le Sénat dans trois semaines.

Dernier point sur un principe de l'open gouvemement, il se trouve que la France va prendre à l'automne la présidence d'un organisme international de « soft law », l'OGP (Partenariat pour un gouvernement ouvert). L'objectif de la France est d'instaurer une dynamique constructive dans notre pays auprès des administrations pour les encourager à adopter des méthodes d'ouverture et de co-construction, d'intégrer des start up d'Etat et d'aller toujours plus loin en matière de transparence des politiques publiques.

La loi pour la transparence de la vie publique a été adoptée en ce sens tandis que Michel Sapin va faire passer une loi pour protéger les lanceurs d'alertes, contenant également de nombreuses dispositions relatives à la transparence.

En conclusion, sans verser dans la naïveté ou l'angélisme, il me paraît essentiel de tirer le meilleur parti possible des technologies numériques pour orienter les travaux des chercheurs et par conséquent les politiques publiques. Face aux reproches d'irrationalité et d'imprévisibilité qui sont si souvent faits aux décisions politiques, objectiver les débats est un enjeu majeur, y compris pour obliger les hommes et les femmes politiques à assumer pleinement leurs responsabilités en la matière.

Antoine Frachot
Directeur Général du GENES

Alain Trognon
Conseiller scientifique du CASD

Antoine Frachot : historiquement, Alain Trognon peut être considéré comme une sorte de « grand-père » du CASD. Il a connu la période des premières expériences américaines jusqu'au développement par trois ingénieurs de l'administration publique de la technologie du CASD.

Alain Trognon : au cours de la décennie 1980, il y a d'abord eu une convention passée par l'Insee avec le CNRS pour permettre à certains chercheurs et universitaires de venir se pencher sur les bases de l'Insee. En raison du règlement sur le secret statistique, c'était à peu près la seule manière qui leur était offerte d'avoir accès à des données détaillées pour développer des travaux. Mais ils ont aussi pleinement contribué à la construction d'études statistiques.

Ensuite, dans les années 90, il se crée en Amérique du Nord des CAS « physiques », c'est-à-dire des îlots dans les universités comme Cornell, qui permettent aux chercheurs d'avoir accès aux données dans des conditions sécurisées. Parallèlement, le développement d'Internet et de l'échange d'information va aboutir à la mise en place par les statisticiens de systèmes d'interrogation à distance en Europe du Nord et en Amérique du Nord.

La statistique publique française, de son côté, se tient à l'écart de ces deux options. Ce n'est que dans les années 2000 que le GENES, fort de sa position à la charnière de l'enseignement supérieur, de la recherche et de la statistique publique prend l'initiative, avec le concours du CREST qui est pleinement ouvert sur l'international, d'accueillir des chercheurs en son sein tout en garantissant le respect du secret statistique par la mise en place de serveurs sécurisés. L'expertise par nos informaticiens des systèmes d'interrogation à distance existants ayant révélé des lacunes en matière de sécurité, il a fallu se tourner vers une autre solution. Une fois celle-ci trouvée, c'est vers l'enseignement supérieur que nous nous sommes tournés pour la mise en place d'une expérimentation dans le domaine des sciences humaines et sociales, particulièrement avides de pouvoir disposer de données. L'initiative pilote conduite avec Paris Sciences Eco, Toulouse Sciences Eco et l'INED a ainsi permis de regarder si l'outil développé était acceptable et gérable pour le chercheur dans la mesure où cela change ses habitudes de travail.

Ensuite, il s'agissait de donner confiance aux producteurs de données et aux entreprises en associant à cette opération un audit de sécurité. A partir de là, je me retire...

Antoine Frachot : le CASD a donc été développé et a remporté la compétition « Equipement d'excellence », ce qui a permis de le doter de moyens assortis de l'objectif ambitieux d'accompagner toujours plus de chercheurs et de projets de recherche. En 2011, un brevet est déposé, tandis que l'année 2012 est celle du déploiement plus général, à l'échelle de toute la communauté scientifique française, mais aussi étrangère. De son côté, le gouvernement a œuvré à l'ouverture des données fiscales en 2013, puis des données de santé récemment, et s'attelle en ce moment au projet de loi sur la République numérique. En 2014-15, le CASD a mis au point un pilote de centre d'accès sécurisé européen entre l'Italie, l'Allemagne et la France. L'objectif est de construire à terme un réseau européen à l'échelle d'une vingtaine de pays. En 2014 aussi, les premières entreprises privées testent la technologie pour donner accès à leurs propres données afin de répondre au besoin de faire travailler ensemble chercheurs, experts ou consultants… sur celles-ci sans risque qu'elle s'échappent.

Pour ma part, je considère le CASD comme une start up de l'administration française et je suis fier que celle-ci puisse en créer. Il faut souligner pour finir que tout ça n'aurait pas existé sans les trois ingénieurs qui l'ont construit, Kamel Gadouche, Philippe Donnay, Eric Debonnel … alors qu'à l'étranger ce sont des bataillons de centaines de personnes qui étaient mobilisées sur la question des accès sécurisés. Comme on n'avait pas les moyens en France, ils ont relevé le défi et il faut les en remercier vivement.

Les enjeux de l'accès aux données sécurisées : introduction

Yannick Moreau

Présidente du Conseil national de l'information statistique (Cnis),
Présidente de section honoraire au Conseil d'Etat.

Yannick Moreau est ancienne élève de l'école HECJF, de l'ENA et docteur en droit. Le Conseil d'État est sa première affectation et elle y a passé la moitié de sa vie professionnelle. Elle en a présidé la section sociale de 2006 à 2011. Elle a exercé d'autres fonctions dans des cabinets ministériels, des administrations et des entreprises publiques. Elle a présidé le Conseil d'orientation des retraites de 2000 à 2006.

Pourquoi le Cnis s'intéresse-t-il aux questions de l'accès aux données ? D'abord, deux mots sur cette institution. La période suivant la Seconde Guerre mondiale est une période de transformation de l'appareil public avec une démarche associant souvent les partenaires sociaux. Le constat que l'appareil statistique français est très insuffisant conduit à la volonté d'améliorer le recensement de la population et de développer les enquêtes auprès des entreprises et des ménages. Ceci ne peut être mis en œuvre sans associer les représentants de l'économie et de la société dont la représentation est ainsi prévue par la loi de 1951. C'est le rôle du Cnis aujourd'hui d'organiser la concertation entre les représentants de l'économie, de la société et les producteurs de statistiques. La concertation est organisée autour de commissions mais aussi de manière plus informelle ; par exemple, à une assemblée plénière, on peut voir « débarquer » un chercheur, Roxane Silberman par exemple, pour demander que le Cnis s'occupe de faciliter l'accès aux chercheurs des données qui servent de base aux statistiques. Ceci aboutira à ce qu'un rapport soit réalisé sur ce sujet ; le Cnis a ainsi servi de catalyseur.

Le Cnis s'est ainsi intéressé activement à cette question de l'accès aux données servant de base aux statistiques publiques parce que c'est dans son ADN en tant que lieu de rencontre sur les statistiques publiques.

Pendant un certain temps, l'accès s'est fait de manière informelle pour des chercheurs nouant une relation particulière avec l'Insee ou les services statistiques ministériels. Cette solution n'était pas satisfaisante. Pour ouvrir plus largement, il était cependant nécessaire de garantir techniquement et juridiquement la protection des secrets prévus par la loi, notamment le secret statistique, et l'égalité d'accès.

L'intervention de la loi de 2008 qui a donné au Comité du secret statistique le rôle d'examiner les demandes des chercheurs et les garanties associées ainsi que la création du CASD, ont rendu les choses possibles.

Désormais, l'accès aux fichiers des bases de données statistiques est très largement effectif. Pour les statistiques de la Banque de France, le processus est moins avancé mais la mise en place d'un comité d'accès vient d'être officialisée sur son site. La question de l'ouverture des données bancaires et financières a fait l'objet d'un groupe de travail du Cnis en 2015.

Un deuxième pas important a été franchi avec l'accès aux données administratives fiscales. En ce sens, la loi de 2013 est à saluer comme volontariste dans un domaine qui n'était pas a priori le plus ouvert , mais aussi comme la première des lois qui vont permettre un accès organisé aux données de gestion des administrations ; l'accès est organisé comme pour les statistiques autour du CASD et du Comité du secret statistique.

Le troisième pas important est l'accès aux données de santé pour lequel la loi a prévu des règles et une organisation particulière. Il existe donc actuellement trois voies juridiques distinctes, pour les statistiques et la fiscalité, d'une part, pour la la Banque de France, d'autre part, et pour la santé, enfin.

Reste cependant non couvert le champ important des autres fichiers administratifs, par exemple, des fichiers des caisses nationales famille et vieillesse pour lesquelles il existe des règles de secret professionnel à concilier techniquement et juridiquement avec une ouverture contrôlée aux données. C'est l'objet du projet d'amendement gouvernemental à la loi numérique en cours de discussion dont a parlé Madame Axelle Lemaire.

Il reste donc à élargir le champ des données auxquelles les chercheurs peuvent avoir accès. Mais il y a aussi des impératifs concernant le bon fonctionnement des trois systèmes d'accès existants. Il me semble qu'il faut tout d'abord assurer une certaine transparence de la part des différents comités concernant leur composition, le nombre de demandes qu'elles reçoivent, acceptent ou refusent… bref, nous avons besoins de données sur l'accès aux données ! Ensuite, il est important que les responsables de ces trois circuits se parlent pour tirer parti des meilleures pratiques des uns et des autres. Et, soyons fous, il faudrait même aller vers l'évaluation de ces pratiques.

Le Cnis peut être le lieu de dialogue entre ces instances pour qu'elles fonctionnent le mieux possible et de la façon la plus transparente possible.

La tension entre le besoin de sécurité et l'ouverture des données s'exprime-t-elle partout de la même façon ?

Je n'ai pas une bonne connaissance personnelle de ce qui se passe ailleurs. Mais j'ai tendance à penser que se passer d'un minimum d'encadrement juridique, serait prendre le risque de laisser les craintes prendre des proportions contre-productives et que cette nécessité doit se manifester aussi dans les autres pays. Le débat parlementaire autour de ces questions est nécessaire et normal dans une culture démocratique.

L'accès aux données sécurisées, un enjeu essentiel pour les sciences sociales

Thomas Piketty

Directeur d'études à l'EHESS et Professeur à l'École d'économie de Paris/Paris School of Economics

Thomas Piketty est directeur d'études à l'EHESS et professeur à l'Ecole d'économie de Paris. Il a publié de nombreux articles de recherche dans des revues internationales telles que le Quarterly Journal of Economics, Journal of Political Economy, American Economic Review, Review of Economic Studies, ainsi qu'une dizaine de livres. Il est l'auteur de travaux historiques et théoriques consacrés à la relation entre développement économique et répartition des richesses. Il est notamment l'initiateur de la littérature récente sur l'évolution sur longue période de la part des hauts revenus dans le revenu national (maintenant disponible dans la World Wealth and Income Database). Ces travaux ont conduit à remettre en cause radicalement l'hypothèse optimiste de Kuznets sur le lien entre développement et inégalités, et à mettre en évidence l'importance des institutions politiques, sociales et fiscales dans la dynamique historique de la répartition des richesses. Il est également l'auteur du best-seller international "Le Capital au 21e siècle".

La question : un exemple concret des enjeux de l'accès aux données. Je vais vous présenter un exemple qui en témoigne à travers les possibilités offertes par l'appariement en matière d'étude des inégalités. Des travaux américains ont ainsi mis en évidence une ligne droite entre les revenus des parents et l'accès à l'enseignement supérieur : on passe de quasiment 20% de chances d'aller à l'université pour les enfants des parents les plus pauvres à quasiment 90% pour les enfants des parents les plus aisés. Evidemment, ce résultat spectaculaire révèle aussi un écart considérable entre les discours publics autour de la méritocratie, de la mobilité sociale et la réalité.

En France, la courbe serait sans doute un peu moins accentuée notamment en raison des droits d'inscription très élevés dont il faut s'acquitter dans les facultés américaines, mais justement, le problème c'est qu'on ne sait pas. Dans la mesure où il n'y a pas d'appariement possible entre les déclarations de revenus et les numéros d'étudiants, il n'y a pas non plus de comparaison possible et donc de possibilité de soumettre à la critique un certain nombre de positions prises dans l'espace public.

Comment la France se situe-t-elle en matière d'accès aux données par rapport aux autres pays ?

Grâce au CASD, nous sommes plutôt en avance. L'accès aux données tel qu'il est permis par le CASD est bénéfique aux chercheurs, en particulier les jeunes thésards qui étaient il y a encore quelques années confrontés à de fortes restrictions. Aujourd'hui, cet accès est possible dans des délais relativement brefs et de très bonnes conditions techniques. A titre d'exemple, l'accès aux données fiscales est beaucoup plus difficile aux Etats-Unis.

Quel équilibre viser entre la sécurité et les usages ?

Le CASD apporte des garanties pour sécuriser les partenaires et les producteurs de données, notamment via la prise des empreintes digitales, l'étanchéité entre les machines, le fait de passer par le serveur du CASD pour traiter les données… Mais il faut aussi grâce à cette technologie qu'on arrête de considérer les chercheurs comme des délinquants en puissance. Ce système les place en situation de démontrer l'esprit de responsabilité dans lequel ils travaillent. Il est essentiel qu'il soit impossible de trahir la confidentialité des données, et que des sanctions soient appliquées le cas échéant.

Quels sont les progrès encore à réaliser ?

Des progrès importants ont déjà été faits. La question de l'appariement est très importante, de nombreux fichiers ne prenant du sens que croisés. La vigilance est de mise, car plus tôt les choses seront effectives, moins on s'exposera à un retour en arrière. Il existe aussi des domaines où l'accès aux données est encore insuffisant, par exemple en matière fiscale ce qui touche le patrimoine, des déclarations de succession… Il s'agit d'informations qui ne sont pas correctement conservées et numérisées, alors que derrière, il y a des enjeux fondamentaux sur l'accès au logement, à la propriété. En ce qui concerne l'accès à l'enseignement supérieur, on a un logiciel, l'APB (Admission Post Bac) sur le fonctionnement duquel il n'y a aucun débat public faute d'accès aux données qu'il contient.

Espérons que le CASD va contribuer à rassurer les administrations qui utilisent des excuses plus ou moins convaincantes pour ne pas fournir ce type de données. Enfin, dans le domaine des archives, actuellement, les systèmes existants tant au niveau national qu'au niveau des ministères n'ont pas pleinement accompli leur révolution numérique et on continue de perdre des fichiers, ce qui est terrible pour la mémoire nationale. Là encore, il y a une carte à jouer pour le CASD même si ce n'était pas prévu dans ses missions initiales.

Les hommes et femmes politiques écoutent-ils suffisamment les chercheurs ?

Je ne sais pas, Il n'y a pas de règle, Axelle Lemaire qui est présente, je suis sûr les écoute beaucoup… d'autres les écoutent moins. On entend même certains qui invoquent le manque de temps pour lire, alors… Il ne faut pas se plaindre comme cela, il faut d'abord que les chercheurs fonctionnent dans une logique d'offre en améliorant la qualité des recherches. Ce qui compte en priorité, c'est, plutôt que de chercher à influencer les hommes politiques, qu'ils se préoccupent d'informer l'opinion publique sur les enjeux essentiels de politique publique.

Axelle Lemaire : je partage ces réflexions : oui, on peut aller plus loin, mais il est appréciable que Thomas Piketty rappelle que tout n'est pas noir en France. Et il faut trouver des solutions juridiques et technologiques pour continuer à avancer.

L'utilisation des données sécurisées dans le domaine de l'éducation

Camille Terrier

Doctorante à l'École d'Économie de Paris

Camille Terrier est doctorante à l'École d'Économie de Paris, actuellement accueillie à la London School of Economics. Ses travaux portent sur différents domaines en économie du travail et économie de l'éducation. Parmi ses travaux en cours, Camille (avec Olivier Tercieux et Julien Combe) analyse le mécanisme utilisé pour affecter les enseignants dans les écoles. Une autre de ses recherches porte sur les biais de genre existants dans les notes des enseignants, et leur effet sur le progrès des élèves, et leurs choix de disciplines.

Mes recherches sur données françaises portent d'une part sur l'impact des pratiques évaluatives des enseignants sur les progrès des élèves et leurs décisions d'orientation, d'autre part sur le processus d'affectation des enseignants du secondaire. Dans ce deuxième cas, il s'agissait de mettre en évidence, avec mes deux co-auteurs Julien Combe et Olivier Tercieux, les propriétés et les limites de l'algorithme utilisé actuellement par l'Education nationale et à partir de là de suggérer un algorithme alternatif puis d'évaluer l'amélioration possible. Cette étude ne pouvait se faire qu'à partir de données sur la mobilité des enseignants. Uniques et jusqu'alors inexploitées car récoltées et conservées par la Direction Générale des Ressources Humaines (DGRH) du ministère de l'Education Nationale, elles ont pu être mises à disposition de la DEPP, qui a accepté de les déposer au CASD afin que je puisse y accéder depuis l'Angleterre où je me suis installée en troisième année de thèse. Les données en question contiennent quatre informations-clés : les vœux de mobilité des enseignants, leur barème, les places vacantes dans les académies/établissement et les affectations initiales des enseignants. Sur le plan technique, nous avons bénéficié de deux SD-Box à Paris et à Londres avec des logiciels installés sur les serveurs et un accès simultané. L'ensemble nous a été fourni dans un délai de moins de six mois. Tout au long de leur utilisation, nous avons importé régulièrement des fichiers, et fait tourner de gros calculs, ce qui a nécessité d'augmenter la mémoire à notre disposition.

Nos recherches ont ainsi montré que nous aurions pu, en 2013, accroître de 40% le mouvement inter-académique du second degré au niveau national. Cependant, une telle hausse sans précautions additionnelles risque d'accroître la surreprésentation des jeunes enseignants dans les académies les moins attractives. En tenant compte de ce risque, nous trouvons qu'une procédure alternative d'affectation permettrait toutefois d'augmenter de plus de 30 % le nombre d'enseignants obtenant une nouvelle affectation. Ce travail de trois ans a abouti à la conception d'un outil de pilotage et de simulation sur les possibilités de mouvement dans les différentes académies, mais aussi à la rédaction (en cours) de deux articles de recherche et à la publication d'une Note d'Information de l'Institut des Politiques Publiques.

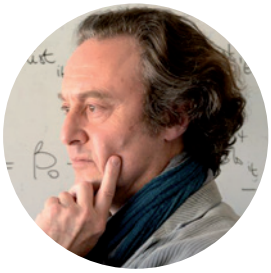
Dans le cadre de mes travaux de recherche, la contribution du CASD a été fondamentale.

L'Education nationale pourrait-elle être amenée selon vous à envisager l'usage de cet algorithme alternatif ?

Si cet outil pouvait être utile pour réformer les politiques publiques, tant mieux, et c'est l'une des raisons pour lesquelles nous avons produit une note pour l'IPP qui vulgarise le résultat de notre étude. Nous espérons que cette recherche générera un intérêt de la part du ministère mais aussi des syndicats d'enseignants.

Les Britanniques s'intéressent-ils à l'accès aux données pour gérer leur système d'affectation des enseignants ?

En France, pour avoir accès aux données éducatives, il faut avoir un pied dans l'institution, d'où le fait de prendre contact avec la DEPP. En Angleterre, en revanche, le Ministère a mis en ligne des formulaires pour l'accès à des données très similaires, mais il faut être néanmoins physiquement présent dans les universités qui sont elles-mêmes responsables de la sécurisation de ces données via des serveurs locaux.



Un bref historique de l'accès aux données sécurisées pour les chercheurs et les nouveaux potentiels ouverts pour la France au niveau international

Francis Kramarz

Directeur du CREST, professeur d'économie à l'ENSAE et Polytechnique

Directeur du Centre de Recherche en Économie et Statistique (CREST), Francis Kramarz est professeur d'économie à l'ENSAE et professeur associé à l'Ecole Polytechnique. Il dirige également le Labex ECODEC (ENSAE, HEC, Ecole Polytechnique), « réguler l'économie au service de la société ».
Nommé par le Premier Ministre comme expert au Conseil d'Orientation pour l'Emploi, à la Commission d'Experts sur le SMIC (jusque 2013) et au Conseil d'Analyse Economique (jusque 2013), il est élu Fellow de l'Econometric Society en 2013.
Francis Kramarz a publié de nombreux articles dans des revues internationales sur les thèmes de l'emploi, des revenus, des politiques de compensation des entreprises, de la concurrence, du commerce international, etc.
Co-auteur du rapport « Vers une Sécurité Sociale Professionnelle » (avec P. Cahuc, 2005) sur la flex-sécurité, il a publié en 2008 l'ouvrage Working Hours and Job Sharing in the EU and USA (avec T. Boeri et M. Burda) et « Ni en Emploi, Ni en Formation » (avec M. Viarengo, 2015) aux Presses de Sciences-Po.

En ce qui concerne ce qui existait avant le CASD, je vais prendre l'exemple des DADS et des données d'entreprise.
Donc un jour, un chercheur, aujourd'hui chef scientist du Bureau du Census aux Etats-Unis, voulait prendre une année sabbatique en France. Il avait trouvé une note de Stéphane Lolivier, de l'Insee, qui disait qu'en France il existe des données qui permettent, simultanément, de savoir où travaille chaque individu et de connaître l'ensemble des individus qui travaillent dans une même entreprise, soit ce qu'on appelle aujourd'hui les DADS. Et on s'est rendu compte qu'en économie du travail, nombre des questions qui se posent classiquement - par exemple le salaire, est-il lié à l'individu ou à l'entreprise dans lequel il travaille ? - ne pouvaient être résolues qu'en ayant simultanément des données sur l'individu et des données sur l'entreprise. Or, un certain nombre de personnes se servaient de ce type de données pour faire des études sur les salaires, mais sans jamais utiliser la composante entreprises. On m'a alors proposé de travailler avec ce chercheur de l'université Cornell et j'ai tout de suite senti que ça pouvait être intéressant. De fait, l'impact a été considérable sur ma carrière puisque nous avons écrit un des articles les plus cités en économie du travail. Nous avons aussi à cette occasion développé des techniques statistiques pour utiliser ce genre de données.
Concernant les données d'import/export par entreprise, même si j'entendais les gens de Harvard parler du «biais technologique », je trouvais que la question du commerce international se posait.

Mais il fallait avoir des données sur les imports et les exports, qu'ont pu me fournir deux personnes qui construisaient des statistiques pour la comptabilité nationale sur les comptes extérieurs. Cette étude sur le comportement des entreprises à l'international a aussi débouché sur un des articles les plus enseignés en commerce international. Très bien, sauf que ça implique des conflits d'intérêts massifs. Certes, je n'étais pas le seul à pouvoir y accéder, mais… c'est ainsi que très naturellement s'est imposée la nécessité d'élargir l'accès aux données. L'environnement était très favorable à cette idée que c'était un bien commun dont devait pouvoir faire bénéficier le plus grand nombre.
Sur le versant des forces du système actuel, la France a les atouts pour devenir un potentiel leader mondial pour l'étude des entreprises et de leurs salariés, ce qui n'a pas échappé aux chercheurs en Angleterre ou en Italie. Ainsi, avec mes co-auteurs italiens Chiara, Giacinta, et Gio nous avons examiné comment les groupes d'entreprises peuvent assurer l'emploi de leurs salariés face à des chocs négatifs ou positifs. Nous avons utilisé les DADS, les bilans d'entreprise et la connaissance de la constitution des groupes. La question était : est-ce-que les groupes d'entreprises fournissent des mécanismes assurantiels, c'est-à-dire est-ce que, quand une entreprise ferme, les salariés bougent vers une autre entreprise du groupe ou de manière aléatoire sur le marché du travail ? Notre étude met en évidence les frictions sur le marché externe du travail, frictions dont pâtissent les plus petites entreprises, les entreprises jeunes qui ont du mal à recruter en phase d'expansion.

Un autre exemple en creux : avec mes co-auteurs autrichiens, Andreas et Josef, nous nous sommes penchés sur un système de sécurisation des parcours en Autriche grâce à l'existence d'un compte indemnité transférable d'une entreprise à l'autre, même en cas de démission. Notre travail montre que les salariés d'une entreprise qui va mal n'attendent pas qu'elles meurent pour partir. Or si les données appariées individus-entreprises (style DADS) sont faciles d'accès, l'absence d'un CASD en Autriche rend l'accès aux données d'entreprises impossible, ce qui nous empêche de caractériser les sociétés qui bénéficient de ce système.

Un point que je souhaitais aussi soulever : on n'évalue que des politiques qui existent. Quand on discute de la loi El Khomri, par exemple de l'effet des restrictions d'embauches et de licenciements sur la création d'emploi, on peut ainsi toujours dire qu'on n'a pas de preuve française, mais on sait qu'il existe un lien positif entre flexibilité et création d'emploi d'après des expériences nord-américaines.

En termes de manques sur les données, je ne vais pas attendre que les données soient accessibles sur le CASD pour travailler sur des questions qui me semblent intéressantes. Donc je vais le faire sur des données suédoises. Ainsi, nous avons mené avec mes co-auteurs suédois un travail sur la force des liens forts en matière d'emploi des moins diplômés. Grâce aux informations dont nous disposions sur les personnes,mais aussi avec l'équivalent des DADS, nous avons montré que 15% des moins diplômés trouvent leur premier vrai emploi dans le même établissement que leur père ou mère et qu'ils restent plus longtemps dans l'entreprise, du privé la plupart du temps, que leurs camarades de classe entrés dans la même entreprise. De la même manière, nous pouvons examiner ce qui se passe lorsque lorsqu'une mort surprise affecte un salarié dans une entreprise : ses collègues prennent moins de congés maladie dans les mois qui suivent. Vous voyez donc qu'on peut poser nombre de questions avec ce genre de données, mais je préférerais le faire sur les données françaises. Et avec à ce qui se met en place grâce au travail de multiples personnes, on va pouvoir le faire… et avoir des réponses qui seront certainement différentes.

Ce qui ressort de votre intervention, c'est que plus on partage les données, plus on fait des découvertes inattendues qui remettent en cause des idées reçues, des conservatismes…

En fait il existe déjà un nombre important de gens qui savent, d'où le titre d'une de mes présentations à Toulouse : « Le seul pays où les Noirs n'ont pas de couleur ». Le problème c'est qu'on ne pourra lutter contre les discriminations que lorsqu'on enlèvera du pouvoir à la majorité blanche et qu'on en donnera un peu à la minorité plus colorée…

Dès lors est-ce-que la mise en place de statistiques ethniques vous semble pertinente ?

La question, c'est peut-on faire des fichiers administratifs ethniques ? Je n'y serais pas le plus opposé malgré les préventions de ceux qui invoquent « les heures les plus sombres de notre histoire », je n'ai pas connu mes grand-parents paternels de ce fait… A mes yeux il s'agit d'une excuse pour protéger ceux qui ont du pouvoir actuellement.

Est-ce-que cela veut dire que ces statistiques qu'on s'interdit de faire pour des raisons idéologiques, de méthodologie, de paresse intellectuelle…

Les technologies disponibles au CASD offriraient-elles des garanties suffisantes pour les mettre en place ?

Je ne sais pas. Une bonne démocratie avec un bon système éducatif, c'est aussi une garantie. Mais il faut les deux.

L'expérience du dataLAB pour les données massives (big data) sécurisées de RTE, gestionnaire du réseau éléctrique haute tension français

Samir Issad

Responsable d'études R&D au sein de RTE

Samir ISSAD est responsable d'études R&D au sein de RTE, gestionnaire du réseau électrique haute et très haute tension français. Ingénieur des Mines, spécialiste en mathématique appliquée puis en informatique, il a travaillé sur les thématiques de la sureté du système électrique en exploitation d'une part (tensions hautes, congestions, maitrise de l'équilibre offre-demande) puis en développement moyen/long terme (évolutions des infrastructures électriques) au travers des méthodes d'analyses probabilistes. Il s'occupe à présent de problématique Gestion d'Actif ou comment optimiser les investissements en renouvellement/maintenance, maitriser les risques et accompagner l'insertion des énergies renouvelable et la transition énergétique. Il met en particulier la datascience au service de ces missions en installant ses outils comme des leviers de performance opérationnelle pour RTE.

Energie renouvelable, biodiversité, mix énergétique, maîtrise de la demande, marchés de l'énergie, développement des territoires… sont quelques-uns des mots clés illustrant les défis auxquels est confronté notre système électrique. Quelles performances du système de l'infrastructure électrique pour répondre à ces enjeux ? C'est à ce titre que la donnée joue un rôle décisif, nous conduisant à élaborer une stratégie numérique autour de trois initiatives : la première est appelée hub numérique, car une donnée utile est une donnée qui va circuler, être accessible au bon endroit, au bon moment. En second lieu, le mal numérique correspond à la contribution de RTE au service public de la donnée. On dispose par exemple d'une plateforme open data accessible sur Internet. Mais une donnée utile, c'est aussi une donnée placée au coeur du processus d'innovation : c'est ce qu'on appelle le Lab numérique. Le but, c'est de se donner les moyens pour travailler cette donnée, prototyper, développer de nouvelles plus values. L'expérimentation qu'on a menée avec le CASD s'inscrit dans cette démarche de Lab. Il s'agit de jouer avec les données afin d'évaluer les bénéfices possibles pour les différents acteurs du système électrique.

Pour ma part, à la R&D de RTE, je travaille notamment sur la problématique de gestion des actifs : matériels, ouvrages… qui sont nombreux et divers sur tout le territoire. Ces infrastructures entretiennent elles-mêmes toutes sortes d'interactions avec l'environnement, le milieu industriel, agricole, etc. Elles nécessitent des investissements et des moyens pour un renouvellement et une maintenance efficaces dans des conditions parfois difficiles. Dans ce contexte, la donnée au travers de son analyse, nous permet de produire des modèles descriptifs, pour observer ce qui se passe pour nos réseaux, chez les différents acteurs du système, etc. dans une diversité de situations, à court ou à long terme, des modèles prédictifs, qui nous aident à anticiper et à optimiser nos décisions, et enfin des modèles prescriptifs qui nous permettent de mieux spécifier à nos fournisseurs quels types de matériels nous souhaitons, aux exploitants comment les utiliser, au maximum de leurs performances mais en sécurité… L'idée derrière l'expérimentation entamée avec le CASD en 20014-15 et qui se poursuit aujourd'hui, c'est aussi qu'on ne pense pas pouvoir tout faire tout seuls, car nous avons besoin de travailler avec des compétences spécifiques, universitaires par exemple. C'est ainsi que nous coopérons avec la filiale de valorisation du GENES, Datastorm, des écoles de statistiques, l'ENSAE et l'ENSAI, mais aussi des PME. Quant au CASD, il fournit l'infrastructure sécurisée autour de laquelle nous nous réunissons pour travailler sur des données confidentielles, celles que nos consommateurs et producteurs, qui vivent dans un monde concurrentiel, ne veulent pas voir divulguées. Au sein du CASD, nous avons donc développé un prototype qui nous permet de mener des approches exploratoires sur les données du système électrique, croisées avec d'autres sources de données. Grâce aux nouvelles technologies, on peut le faire mieux et plus rapidement ce qui ouvre de nouvelles perspectives quant aux solutions que nous pouvons développer pour répondre à nos missions.

Et ça, c'est donc possible grâce à la puissance de calcul, au cloud, à des nouveaux logiciels… Mais qu'est-ce-qui a permis de passer de quelques heures ou quelques jours contre des semaines ou des mois avant ?

C'est la possibilité de réunir sous une même infrastructure ces exigences que sont les moyens de calcul et de stockage, des outils innovants… le tout dans une enceinte sécurisée, sans risque de dissémination des données. Et le CASD nous apporte aussi des compétences en informatique et en data science pour faire émerger des prototypes.

Pour conclure, la transition électrique, la mutation du système électrique et le virage numérique constituent trois vagues qui se conjuguent pour répondre aux enjeux actuels. Pour RTE, développer, exploiter et maintenir les infrastructures avec les moyens que la collectivité que lui concède - RTE investit chaque année 1,5 milliard d'euros dans son réseau pour la maintenance et le renouvellement - cela passe par savoir comment les dépenser en ciblant au mieux les besoins de demain. Quant aux données, elles sont en elles-mêmes un patrimoine, ce dont les gens me semblent être aujourd'hui convaincus. De plus en plus, nos métiers nous sollicitent autour de leur usage et nous allons continuer à tout faire pour que cette culture de la donnée irrigue l'entreprise au delà de la R&D.

Pourquoi le CASD au lieu d'une infrastructure interne ?

Nous y avons réfléchi mais nous avons besoin de travailler avec des acteurs externes, notamment pour s'adjoindre des compétences, se retrouver dans un lieu sécurisé. Ouvrir des voies d'accès aux données de RTE à des tiers, ce n'est pas simple pour des raisons de sécurité. On s'est aussi posé la question d'aller louer des serveurs standards sur le marché existants, par exemple un serveur américain, mais cela soulevait de multiples questions, autour de la législation, de l'envoi de la donnée aux Etats-Unis, etc. alors que le CASD offre des garanties.

Le pylône est-il devenu intelligent ?

En fait, le pylône est intrinsèquement porteur de données : on sait quand ils ont été installés, quels sont leurs caractéristiques. Mais on ne se servait pas suffisamment de ces informations jusqu'à présent. Or elles peuvent nous permettre par exemple de savoir placer le « bon » pylône au « bon » endroit vis-à-vis du risque climatique. Nous avons cessé de simplement stocker la donnée, désormais nous nous en servons.

Pour bénéficier de tous les avantages de la donnée, faut-il un investissement du « big boss » ?

Oui, la stratégie numérique est prise à bras-le-corps au plus haut niveau. L'idée que l'exploitation, la maintenance, les RH, les achats… produisent des données implique un décloisonnement des métiers qui est impulsé fortement par la direction.

Faites-vous des simulations : de gestion de crise, de reprises d'activité ?

Oui.

Est-ce que les données provoquent du changement opérationnel ? Vous avez un exemple précis ?

En ce moment, par exemple, nous travaillons sur les conducteurs électriques : il s'agit de savoir s'il faut les remplacer à un certain âge. Or avec la donnée, on connaît leur vie passée, on a la possibilité d'exploiter au mieux le matériel sans incident.

Cela veut dire qu'il y a moins d'incidents… sinon, quel intérêt ?

Nous n'avons pas tant d'incidents que ça, ce qui complique d'ailleurs la tâche pour faire de la datascience ! Donc l'idée c'est déjà de maintenir notre niveau de performance actuel.

Quelle quantité mensuelle de données de RTE est-elle stockée au CASD ?

La question n'est pas si simple car la donnée circule, peut être stockée de différentes manières, redondée… Disons environ deux teraoctets mensuel.

Les données à caractère personnel : comment concilier richesse de l'information et protection des données sensibles ?

Sophie Vulliet-Tavernier

Directeur des relations avec les publics et la recherche, Commission Nationale de l'Informatique et des Libertés (Cnil)

Directeur des relations avec les publics et de la recherche à la Cnil. Cette direction est chargée de la gestion et de la valorisation des connaissances et de l'information de tous les publics qui sollicitent la Cnil , notamment par des publications, l'animation d'un réseau d'experts pluridisciplinaires et le développement de partenariats avec les universités et la recherche. Elle assure aussi le pilotage et l'animation des activités visant à promouvoir l'éducation au numérique en lien avec d'autres partenaires.

Précédemment directeur des affaires juridiques, internationales et de l'expertise de la Cnil puis responsable de la direction des études, de l'innovation et de la prospective.

Membre de la chaire de recherche de l'institut mines-telecoms « valeurs et politique des informations personnelles » et membre du CERNA de l'Alliance Allistene.

La problématique de la conciliation entre l'accès aux données à des fins de recherche et la protection des données suscite encore beaucoup de malentendus et d'idées fausses. C'est pour cela que je voudrais clarifier cette problématique, en évoquant précisément trois points. Le premier, c'est la frontière entre données personnelles, anonymat et pseudonymat, le deuxième porte sur la question de la démarche d'analyse « informatique et libertés » que devrait avoir tout chercheur, et le troisième est relatif à la prise en compte des besoins et spécificités de la recherche, à la fois dans la régulation actuelle et la régulation à venir. En effet, nous avons, en France, la Loi Informatiques et Libertés (LIL), au niveau européen une directive mais qui est en fin de vie puisqu'un nouveau règlement européen a été adopté et devrait entrer en application dans les deux ans à venir. Avant d'y venir, un bref rappel sur la Commission nationale informatique et libertés. Il s'agit d'une autorité administrative indépendante avec un triple rôle de contrôle, de sanction et de conseil. Elle comprend, outre le défenseur des droits, 17 membres de tous horizons, d'ailleurs plus nombreux à être issus de la recherche qu'avant. Ses missions ne cessent d'évoluer pour mieux répondre aux besoins des usagers, contribuer à l'éducation au numérique, développer la concertation, la réflexion prospective et éthique… mais aussi mieux accompagner la recherche et l'innovation. A ce titre, elle a noué différents partenariats, est membre du Comité du secret statistique, et mène des actions de sensibilisation auprès des milieux universitaires et de la recherche. Mais elle participe également à des travaux de recherche et en accompagne d'autres. Il faut développer cette démarche d'accompagnement Le nouveau règlement européen nous y incite. Certes, aujourd'hui, il est reproché à la Cnil un certain délai dans le traitement des demandes des chercheurs. Aussi, pour avancer ensemble, nous encourageons vivement ces derniers à venir nous voir pour nous expliquer leurs projets et lever ainsi les incompréhensions.

En ce qui concerne l'usage des données, je voudrais revenir à quelques définitions. Tout d'abord, les données à caractère personnel recouvrent toute information relative à une personne physique identifiée ou identifiable, directement ou indirectement. En ce qui concerne l'identification indirecte elle passe aussi bien par l'apparence physique, l'identité culturelle, sociale, génétique… ou encore technique, via les adresse IP par exemple. La directive européenne actuelle considère que pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en oeuvre pour identifier ladite personne, tandis que la LIL prend plus largement en compte l'ensemble des moyens en vue de permettre l'identification auxquels peut avoir accès ou dont dispose le responsable du traitement ou toute autre personne. Quant au concept de pseudonymisation, qui renvoie aux données indirectes et aux identifiants à caractère personnel, il a suscité un vif débat autour du projet de règlement européen, certains lobbies, notamment, souhaitant que les données pseudonymes ne soient pas considérées comme des données à caractère personnel. Finalement, le règlement définit le pseudonymat comme le traitement de données à caractère personnel de sorte qu'elles ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires. Quant à l'anonymat, il est lié au retrait de tous les éléments permettant d'identifier une personne et à l'impossibilité de les ré-identifier. La doctrine de la Cnil en la matière s'est élaborée de manière progressive, en lien avec les évolutions législatives et technologiques.

L'une de ses premières approches a vu le jour dans le domaine de la santé et de la recherche médicale, avec la problématique du VIH dans les années 85-90. La prise de conscience qu'il fallait réussir à conjuguer l'essor des recherches et la protection de l'anonymat a conduit à développer des techniques de pseudonymisation. Par la suite, une polémique a éclaté autour du PMSI, lorsqu'on s'est rendu compte que l'anonymisation de cette base de données hospitalières n'était pas si effective. Après encore d'autres évolutions au fil de l'eau, l'avis du G29 du 10 avril 2014 a relancé le débat en définissant trois critères permettant d'aboutir à l'anonymat total: l'individualisation, la corrélation et l'inférence. Actuellement, par rapport au développement de l'open data et du big data, nous travaillons à l'élaboration de conseils pratiques pour vulgariser ces techniques, exemples à la clé, en concertation notamment avec des équipes de recherche qui travaillent sur ces sujets.

En ce qui concerne la démarche d'analyse « Informatique et Libertés » que devrait avoir tout chercheur amené à utiliser des données personnelles, elle consiste tout simplement à prendre appui sur les principes de protection des données (finalité, pertinence, sécurité, conservation, droits des personnes) pour se poser les « bonnes questions », par exemple, à réfléchir, au regard des objectifs de recherche que l'on poursuit et de la méthodologie que l'on entend suivre , à ce dont on a réellement besoin: de données réelles ou fictives ? De données directement identifiantes ou pseudonymes ? Peut-on arriver à travailler sur des données anonymes ? Mais il faut battre en brèche l'idée que la Cnil ne prônerait que des recherches anonymes : non, tout dépend du protocole... Au-delà, il y a un aspect important, celui de la pertinence des données par rapport à la finalité, ce que le projet de règlement traduit par le terme de minimisation. Or, ce domaine de la recherche et des statistiques présente à cet égard des spécificités car on est par définition dans le cadre d'hypothèses à tester, d'où la nécessité parfois d'avoir à recueillir a priori un grand nombre de données sans que l'on sache véritablement si celles-ci seront pertinentes. Donc j'insiste encore une fois pour que les chercheurs n'hésitent pas à venir argumenter leurs projets auprès de la Cnil, surtout s'il s'agit de données sensibles (santé, ethniques, sexuelles…)… A ce propos, je voudrais revenir sur cette question de savoir si l'on peut mener des études sur la mesure de la diversité et pour progresser vers l'égalité des chances. La Cnil a émis des recommandations sur ce sujet en 2006 et a coproduit avec le Défenseur des Droits un guide pratique pour les chercheurs et les entreprises expliquant comment mener ce type d'enquête. Donc, contrairement aux idées reçues, il y a des possibilités de le faire en France. Pour finir, on entend dire, trop souvent, que la recherche est le parent pauvre de la régulation dans le domaine de la protection des données. Or aussi bien dans la loi que dans le projet de réglement, il y a une exception pour la recherche qui est reconnue. Elle touche à la réutilisation, à des fins de recherche, des données déjà collectées à d'autres fins, à la possibilité de traiter des données sensibles, de bénéficier de dérogations en matière d'information, d'exercice des droits des personnes, de durée de conservation… et ce moyennant des garanties techniques, légales, organisationnelles… le règlement insiste sur l'aspect mini, pseudo, chiffrement pour le respect des droits des personnes: prise en compte et nécessité de développer des travaux de recherche sur ces sujets de protection et d'anonymisation. Un mouvement se dessine d'ailleurs en ce sens autour de l'INRIA, de l'Institut des Télécoms, à l'étranger… Quant au dispositif du CASD, il se situe dans la ligne souhaitée par le règlement et la Cnil. Au delà de l'aspect recherche, développer des processus de chiffrement, d'anonymisation peut être aussi un atout pour la compétitivité de notre économie numérique en Europe.

Du côté de la loi, un assouplissement des formalités pour l'utilisation du NIR se dessine, mais au delà, il faudra d'autres évolutions législatives. Quel contrôle par exemple pour ce qui est hors recherche en santé et recherche statistique ? Je pense aux sciences sociales: aujourd'hui, il y a nécessité, par exemple, de fixer clairement les conditions dans lesquelles les données soi-disant « publiques » issues du web social peuvent être utilisées à des fins de recherche, … Au delà, à l'égal de ce qui existe dans le domaine de la recherche médicale, faut-il ou non instaurer dans le champ des sciences humaines et des recherches sur le numérique, à côté de la Cnil, des comités d'éthique qui puissent conseiller les chercheurs ?

Comment être compétitif par rapport aux délais actuels de réponse de la Cnil ?

Aujourd'hui, nous sommes effectivement en surcharge concernant l'instruction des dossiers, et c'est pourquoi avec le projet de loi pour une République numérique on s'oriente vers la mise au point de procédures déclaratives simplifiées. Avec le règlement européen, les formalités déclaratives vont très largement disparaître au profit de référentiels, de codes de conduite, d'études d'impact vie privée, en bref d'une plus grande responsabilisation des entreprises et administrations… … Enfin, dans ce débat un peu technique, n'oublions pas les personnes dont les données sont utilisées , aujourd'hui encore trop peu associées et notamment trop peu informées des résultats des recherches entreprises avec l'aide de leurs données . Là aussi la transparence s'impose …

« Computational privacy » ou comment le comportement humain limite les possibilités d'anonymisation

Yves-Alexandre de Montjoye

Research Scientist MIT Media Lab et Harvard University

Yves-Alexandre de Montjoye est chercheur en mathématiques appliquées au MIT Media Lab et à Harvard. Il développe des méthodes stochastiques pour l'analyse de métadonnées comportementales : données de mobilités, transactions financières, communications dans les réseaux sociaux. Ses recherches ont reçu une couverture médiatique dans BBC News, CNN, The New York Times, Wall Street Journal, Foreign Policy, Le Monde, Die Spiegel, ainsi que dans les rapports du World Economic Forum et des Nations Unies. Avant de rejoindre le MIT, Yves-Alexandre était chercheur au Santa Fe Institute (NM). Sur une période de 6 ans, il a obtenu un Master en Mathématiques appliquées de l'université de Louvain, un diplome d'ingénieur de l'Ecole Centrale Paris (Centralien), son Master en ingénieurie mathématique de la KULeuven (Belgique), ainsi que son Bachelier en ingénieur de l'université de Louvain.

Mes préoccupations portent sur la manière dont le comportement humain peut limiter la portée de ce que l'on met en place pour protéger la vie privée, particulièrement dans le domaine des big data.

L'utilisation des données sensibles à des fins de recherche soulève en effet plusieurs défis, dont le premier porte sur l'unicité et les limites de l'anonymisation. Une des grandes données de la big data, c'est la géolocalisation, très souvent utilisée de manière « anonyme ». Historiquement, en effet, on a réussi à trouver un équilibre entre la protection de la vie privée et l'utilisation des données en les anonymisant.

Pour anonymiser des données, on retire tout d'abord les identifiants directs comme le nom, numéro de téléphone, de carte de crédit, ou l'adresse, etc des bases de données. On va ensuite empêcher les ré-identifications indirectes, par exemple si nous n'avons dans l'échantillon qu'une seule femme de 92 ans, en ajoutant du bruit, en généralisant, par exemple en utilisant une tranche d'âge (plus de 80 ans) plutôt qu'une date de naissance (le 12 juin 1924). Ces démarches permettent d'obtenir des données « anonymes », une notion qui distingue légalement les données personnelles et les données « non personnelles ».

Cependant, différents travaux ont montré que des données qu'on pensait anonymes étaient en fait réidentifiables. On peut citer par exemple les données de recherche du moteur AOL, notamment en raison de la propension des gens à tendance à se chercher eux-mêmes, et à faire des recherches relativement locales. Même constat avec la base de données partagée par Netflix, dans la mesure où quand on regarde un film, on a tendance à le noter, à laisser des critiques sur plusieurs sites spécialisés après…

Dès lors, la question est de savoir si ses ré-identifications sont due à une anonymisation qui n'a pas été correctement effectuée

(comme l'affirment les sociétés spécialisées dans l'anonymization), ou si des éléments plus fondamentaux liés aux big data empêchent de trouver cet équilibre entre la protection de la vie privée et l'utilisation des données en les anonymisant.

On peut, par la métrique, quantifier le risque moyen de réidentification dès lors qu'on est en possession d'un certain nombre d'informations, équivalant à des « points », sur les personnes. Dans ce cadre, l'unicité mesure la probabilité, à partir de la connaissance d'un nombre de « points », d'identifier une personne unique. On peut ainsi considérer qu'une information est la présence dans un lieu et à un moment précis. A partir de là, en travaillant sur une base de données téléphonique d'1,5 million de personne sur 15 mois, combien de points faut-il pour identifier une seule personne à un endroit et un moment donnés ? La réponse est quatre, dans 95% des cas. Le même type d'expérience avec des résultats similaires a été mené avec des données de carte bancaire. Donc, avec le big data, retirer tous les identifiants n'est pas suffisant pour protéger la vie privée. La deuxième étape sera donc de rajouter du bruit, en réduisant la résolution spatiale et temporelle. Le résultat probable de réidentification se comporte alors comme une loi de puissance : si on ajoute un peu de bruit, on baisse significativement la possibilité de réidentification, mais le rendement est ensuite décroissant. Dès lors, il suffit pour l'attaquant de collecter quelques points supplémentaires pour ré-identifier une personne avec une forte probabilité malgré l'ajout de bruit. Là encore, cette précaution n'est pas suffisante.

Le deuxième risque est celui de l'inférence, à partir de données que l'on considère a priori comme pas si sensibles que ça, celles du téléphone par exemple. Dans le passé, c'était en effet le cas. Cette fois, nos recherches ont consistées à utiliser des métadonnées d'appels, pour voir ce qu'il est possible de prédire sur la personne selon la façon dont elle se sert de son téléphone. Après avoir soumis des étudiants à un test de personnalité classique, on a utilisé leurs données téléphoniques pour construire des indicateurs comportementaux en fonction du temps passé au téléphone, du nombre de personnes appelées, de la répartition entre appels et envois de SMS, du temps mis à répondre à ces derniers… tout en faisant appel à la géolocalisation pour définir une taille de région telle qu'on sait que la personne s'y trouve 95% du temps. Grâce à l'algorithme construit sur cette base, on a pu inférer dans une proportion significative à quel degré se situe une personne sur l'échelle de tel ou tel trait de caractère, par exemple prédire le degré de neurotisme d'une personne jusqu'à 1.7 mieux qu'au hasard ou son degré d'extraversion avec 61% de chances. Ce résultat, que nous avons répliqué dans des études à grande échelle, met en évidence le fait que la sensibilité des données big data est bien plus importante que ce que l'on peut penser a priori.

Dès lors, comment redéfinir le « trade off » entre vie privée et utilisation des données ? Le paradigme actuel, qui passe par l'anonymisation et l'utilisation de données jugées peu sensibles, ayant montré ses limites, il faut en changer y compris légalement au niveau Français et Européen. Sans abandonner la promesse d'anonymisation, il s'agit de bâtir une infrastructure sur le principe de l'« oignon » qui permet de garantir une utilisation des données dans le respect de leur anonymat. C'est ce que fait le CASD.

Est-ce du ressort du politique et de la loi de fixer le « trade off » ?

Absolument, le « trade off » relève d'un choix sociétal, mais il est essentiel que celui-ci se fasse sur des bases techniques solides. L'anonymisation de données n'offre pas à mon sens un équilibre satisfaisant. Il est nécessaire de changer de paradigme afin que l'on puisse légalement garantir que les données soient utilisées de manière anonyme non pas parce que les données elle-mêmes sont anonymes mais parce que l'on a construit une infrastructure et des mécanismes de sécurité et d'auditing qui garantissent une utilisation « anonymisante » de ses données.

Présentation en image de la technologie CASD

Philippe Donnay

Responsable R&D, CASD

Philippe Donnay a commencé sa carrière à la direction régionale Insee d'Ile de France. Il a ensuite rejoint la direction informatique du Genes pour concevoir et mettre en place une architecture informatique répondant aux exigences des entités du groupe en particulier dans le secteur « concurrentiel » de la recherche scientifique. Il s'est spécialisé dans l'expertise système, le réseau et l'optimisation des performances des logiciels scientifiques de traitement de données. En 2008, il a rejoint l'équipe de projet CASD pour participer à sa création dont il est l'un des principaux contributeurs. Il se consacre désormais entièrement aux travaux de recherche et développement de la technologie CASD et de sa valorisation en France et à l'étranger.

La technologie du CASD est la réponse à un défi : comment donner accès aux données sans donner les données ? Les protagonistes : les producteurs de données d'un côté, les chercheurs de l'autre, et au milieu le futur CASD. Pour les premiers, mettre à disposition les données impliquait des risques comme la perte, le vol, la tentation offerte de les stocker, de les dupliquer ou encore de les garder plus longtemps que prévu. Quant aux chercheurs, ils ont besoin de ces données et ils doivent pouvoir travailler avec de façon confortable.

Ces positions connues, il fallait aller voir ce qui se faisait à l'étranger. Certains font de l'accès physique, l'utilisateur se déplace physiquement chez le producteurs de données, ce qui n'est pas une solution confortable, d'autres pratiquent l'exécution à distance sans à aucun moment voir les données (remote execution), ce qui peut fonctionner sur certains usages mais reste trop contraignant pour les chercheurs.

La solution, était donc de créer un vrai accès complet à distance depuis un poste de travail dans l'institution de l'utilisateur, permettant de visualiser les données et de pouvoir travailler avec elles comme si on les avait en sa possession. Les autres institutions le faisaient en remettant au chercheur un logiciel et un lecteur de carte à installer. Mais les responsables des services informatiques des institutions de recherche étaient réticents à les installer, et à juste titre, au vu de leurs propres impératifs de sécurité. Il fallait trouver mieux...

Nous avons donc travaillé sur la technologie du CASD et mis au point la SD Box, un équipement complet remis au chercheur, installable très facilement sans contact avec l'infrastructure locale. Il peut être isolé du réseau local et fonctionne de manière autonome, avec un objectif principal: établir une communication sécurisée avec le CASD. Ensuite, il suffit au chercheur de rentrer sa carte et d'apposer son empreinte pour pouvoir travailler. En cas de panne, on lui envoie un nouveau boîtier (il ne contient pas de données). Par ailleurs, nous avons voulu proposer un environnement convivial, persistant, similaire en tous points à celui que le chercheur utilise quotidiennement dans son institution.

Pour travailler sur les données, la durée de l'habilitation du chercheur peut varier de un à trois ans, une durée pendant laquelle il peut accéder aux données quand il le souhaite, depuis tout boitier autorisé, grâce à sa carte personnelle biométrique. L'infrastructure est prévue pour supporter un crash éventuel causé par un projet, qui n'impactera pas les machines affectés aux autres projets.

Nous avons ainsi fait en sorte qu'on puisse accéder aux données en faisant disparaître au maximum, dans l'usage, les contraintes associés à la sécurité et à la sensibilité des données.

Quels sont les délais pour avoir accès aux données disponibles au CASD ? Selon le type de donnée, le feu vert peut prendre plusieurs mois ou être donné très rapidement en quelques semaines.

N'y-at-il pas un risque de stigmatisation sur des critères idéologiques associées au projet de recherche ?

Les seuls critères sont la faisabilité et la sensibilité des données.

Comment gérez-vous la grande variété des logiciels avec des contraintes différentes à chaque fois ?

Pour bénéficier d'un maximum de souplesse, chaque serveur de projet est autonome. Si quelqu'un a besoin d'un logiciel qui n'est pas dans le socle commun, il sera fourni, et si le problème est relatif à un contrat de licence, nous négocions avec l'éditeur, parfois le chercheur dispose de sa propre licence valide dans notre environnement.

Un accès au CASD est-il possible pour les journalistes ?

Cela dépend de la demande mais il n'y a pas d'obstacle a priori.

Comment faire lorsqu'on publie dans des revues où l'on est obligé de fournir les données utilisées au referee pour vérification ?

Si l'on donne les résultats intermédiaires ou finaux et qu'ils respectent les contraintes de confidentialité, il n'y a pas de problème. Selon la source, il faut demander une autorisation de sortie. Sinon, il faut anticiper pour que le referee dispose lui-même d'une carte d'accès au CASD en passant par le circuit d'autorisation concerné.

Quel est le coût d'accès aux services du CASD ?

Il s'agit d'une tarification plus ou moins à la carte, en fonction de la configuration souhaitée.

Kamel Gadouche : cela revient en moyenne à 70 euros par utilisateur et par mois tant que nous bénéficions de la subvention Equipex. A partir de 2019, ce tarif sera de l'ordre de 100 euros. Ceci comprend le serveur, le boîtier, les logiciels, le coût des sorties et l'assistance.

Samir Issad : l'habilitation est-elle obligatoirement ad hominem ou peut-on envisager une procédure pour les institutions ou les entreprises ? Par exemple si je veux travailler avec une université, un centre de recherche…

Roxane Silberman : en fait, c'est la finalité d'un projet qui détermine l'habilitation, celle-ci pouvant dès lors être accordée à plusieurs chercheurs travaillant sur un même projet. Habiliter une institution nous ferait sortir de cette logique.

Comment garantir l'inviolabilité d'un système unique centralisé ?

Philippe Donnay : on peut toujours trouver des contre-exemples quand on parle d'inviolabilité. C'est une question de proportionnalité. Nous offrons déjà des garanties très importantes avec plusieurs types de protection qui sont à la mesure du type de données actuellement hébergées. Il y a aussi une question de coût, et de confort de travail : une interface qui devient trop contraignante risque de pousser le chercheur à essayer d'utiliser des moyens détournés pour obtenir les données.

Sophie Vulliet-Tavernier : l'ANSI a-t-elle validé vos procédures ?

C'est prévu. On est actuellement dans un processus de certification ISO27001. Enfin on fait auditer notre système et on va le refaire cette année.

Une version plus petite du boîtier sera-t-elle disponible ? On en est déjà à la quatrième génération, nous avons supprimé le ventilateur par exemple…

Ce dispositif peut-il être utilisé dans d'autres contextes que celui de la recherche ? Par exemple pour accéder aux dossiers médicaux ?

Oui, nous avons construit un socle, une infrastructure sécurisée. Il est tout à fait possible de concevoir d'autres types de services pour d'autres activités qui se posent sur ce socle.

TABLE RONDE


LA SECURISATION
SESSION 2
Table ronde
des producteurs de données
JEAN GAEREMYNCK, Conseiller d'État
MICHEL ISNARD, Insee
BRICE LEPETIT, DGFiP
BÉATRICE SÉDILLOT, Ministère de l'Agriculture
CORINNE PROST, DARES
CHRISTINE CHAMBAZ, Ministère de la Justice
CASD
CENTRE D'ACCÈS SÉCURISÉ AUX DONNÉES

Michel Isnard

Chef de l'Unité des Affaires Juridiques et Contentieuses de l'Insee

Michel Isnard est actuellement chef de l'Unité des Affaires Juridiques et Contentieuses de l'Insee. A ce poste, il assure notamment le secrétariat du comité du secret statistique en charge de donner un avis sur les différentes demandes de communications de données confidentielles faites par les chercheurs. Il a auparavant travaillé sur le projet de rénovation du recensement de la population et à la production de l'indice de production.

Béatrice Sédillot

Chef du service de la statistique et de la prospective, Ministère chargé de l'agriculture

Inspectrice générale de l'Insee, Béatrice Sédillot a occupé précédemment les fonctions de sous directrice en charge des politique d'emploi puis chef de service/adjointe au directeur à la Dares au Ministère du travail (2003-2013), chef de la division Redistribution et politiques sociales à l'Insee (1998-2003), et chargée d'études à la Direction de la prévision au Ministère des finances (1995-1998).

Christine Chambaz

Chargée de la sous-direction de la statistique et des études, Secrétariat général du Ministère de la Justice

Christine Chambaz a fait l'essentiel de sa carrière dans le champ des statistiques sociales : à l'Insee comme responsable d'enquête, chargée d'études, puis comme chef de la division applications et projets pour les statistiques sociales (2003-2006) puis chef de la division Etudes sociales (2006-2008) ; à la Drees comme chef du bureau études structurelles et évaluations (1999-2003), puis à la Dares comme chef du département « salaires et conventions salariales » (2008-2011). De 2011 à 2014, elle est Directrice des statistiques, des études et de la recherche (DSER) à la Cnaf, avant de rejoindre la Drees comme chargée de mission.Depuis janvier 2016, Christine Chambaz est chargée de la sous-directrion de la statistique et des études au sein du secrétariat général du Ministère de la justice.

Brice Lepetit

Chef du bureau des études statistiques en matière fiscale, Direction Générale des Finances Publiques

Administrateur de l'Insee, Brice Lepetit est, depuis décembre 2015, chef du bureau des études statistiques en matière fiscale au sein de la direction générale des finances publiques (DGFiP). Le bureau, service statistique ministériel, est notamment en charge de la prévision et du suivi des opérations financières de la DGFiP, de la diffusion des données fiscales et de la réalisation de simulations relatives à la plupart des impôts et taxes afin notamment de quantifier les impacts budgétaires des dispositions fiscales. Auparavant, il a travaillé 6 ans à la direction de la sécurité sociale en tant que chef du bureau des régimes professionnels de retraite et des institutions de la protection sociale complémentaire et en tant que chef du bureau des recettes fiscales. Brice Lepetit est ancien élève de l'Ecole Nationale de la Statistique et de l'Analyse de l'Information (ENSAI).

Corinne Prost

Chef de service, adjointe à la directrice de la Dares, Ministère du Travail, de l'Emploi, de la Formation professionnelle et du Dialogue social

Corinne Prost est chef de service, adjointe à la directrice de la DARES depuis janvier 2016.
Ancienne élève de l'École polytechnique et ENSAE 1998, Corinne Prost est titulaire d'un DEA de

macroéconomie. En 2003, elle rejoint le bureau du marché du travail et des politiques de l'emploi à la direction de la prévision et de l'analyse économique. Aux États-Unis depuis septembre 2004, Corinne Prost a effectué des recherches en économie à l'université de Cornell en tant que « visiting scholar », ainsi que « research associate » au Cornell Institute for Social and Economic Research (Ciser). Entre 2007 et 2011, elle est responsable de la division Emploi à l'Insee. Puis de 2011 et 2015, elle occupe le poste de chef du département des études économiques à l'Insee. Elle est parallèlement chercheure associée au CREST et poursuit des travaux de recherche sur le marché du travail et l'économie de l'éducation.

Jean Gaeremynck

Conseillé d'État, président du Comité du secret statistique

Table ronde des producteurs de données présidée par Jean Gaeremynck

La tonalité générale de cette matinée a été que l'accès aux données est un bien en soi et va dans le sens de l'histoire. Je m'empresse de dire que je suis d'accord avec cela, mais je voudrais souligner ici que cela ne va pas de soi. Car l'accès aux données, quelles que soient les données, c'est une dérogation au secret, et que le secret, souvent, c'est la confiance ! Ainsi historiquement les enquêtes statistiques publiques ont été lourdement protégées par le secret (loi de 1951). Puis est intervenu (1984) un dispositif dérogatoire d'accès aux données individuelles d'enquête avec l'institution du comité du secret statistique. Dans ce dispositif il faut l'accord des services producteurs, l'avis du comité du secret statistique (le plus souvent favorable) et la décision appartient à l'administration des archives.

Ce dispositif est dérogatoire à plusieurs titres. D'abord, comme exception au secret imposé par la loi et, je viens de le dire, nécessaire à la confiance. Ensuite, comme entorse au lien privilégié entre le recueil de données dans le cadre d'une enquête statistique menée sous le régime de 1951, et une finalité précise consistant en certains travaux d'exploitation statistique pour les besoins desquels l'enquête est faite. Dans cette approche on pourrait dire que la collecte de données n'a pas de valeur en soi et qu'elle s'efface devant les travaux statistiques qui sont sa raison d'être.

Or ce dispositif dérogatoire a connu une mutation quantitative lié à des causes scientifiques et techniques. La conséquence est une nouvelle conception de la finalité du recueil de données, qui n'est plus de mener certains travaux statistique pré désignés, mais une finalité plus générale consistant en des travaux historiques ou scientifiques. De ce fait, les données individuelles d'enquête, et par extension toutes les données détenues par les acteurs des politiques publiques acquièrent une énorme valeur potentielle en soi : c'est en quelque sorte une mine qui ne demande qu'à être exploitée. Il se passe aussi que si la technologie est à l'origine de la mutation quantitative en démultipliant la puissance du traitement des données, elle vient conforter aussi les préoccupations historiques fondamentales relatives à la confiance, au secret et à la sécurité. On peut ainsi interpréter le CASD comme une prouesse technologique admirable en ce qu'elle réconcilie ces fondamentaux avec les besoins de la recherche. A noter que la mutation quantitative a été aussi facilitée par le fait que les chercheurs ont toujours respecté la confidentialité et qu'il n'y a pratiquement jamais eu d'incident. En ce sens le fonctionnement du dispositif dérogatoire n'a pas affecté la confiance.

Ce dispositif donc, fonctionne, mais l'affirmation dans la période récente des besoins des chercheurs dans certains domaines, a montré qu'ils ne pouvaient plus se satisfaire uniquement des données recueillies par les enquêtes statistiques. Cela a conduit le législateur à créer des dispositifs spécifiques dans les domaines du fiscal et de la santé.

L'impact de cette mutation sur les acteurs du système revêt plusieurs aspects. Le rôle des producteurs de données évolue, notamment dans le sens du conseil aux porteurs de projets, candidats à la procédure dérogatoire d'accès aux données. Bien que de création récente, le CASD, tout en révélant sa capacité à faire face à la demande croissante d'hébergement des sources statistiques, a lui-même évolué pour répondre à une demande croissante d'accueil et de conseil des chercheurs. Parallèlement il s'est engagé dans une régulation tarifaire. Quant au Comité du secret statistique, il accompagne lui-même résolument la mutation, conforté par l'absence d'incidents en matière de secret.

Une démarche qui passe par l'affirmation des services producteurs comme point de passage obligé et l'accent mis sur leur rôle de conseil, un intérêt pour le CASD assorti d'une nécessaire vigilance sur les conditions d'accès à ce service, une intégration réussie de la procédure spécifique d'accès aux données fiscales et enfin une adaptation des méthodes internes pour réduire les délais.
Reste toutefois une lacune de taille à combler dans le dispositif général d'accès aux données individuelles, lorsque les demandes de chercheurs portent sur l'accès à des données détenues par des organismes publics ou chargés d'une mission de service public, et issues de la gestion administratives: les textes relatifs au comité du secret ne lui donnent pas compétence juridique en la matière. Les dispositifs législatifs spécifiques mentionnés ci-dessus, relatifs aux données fiscales et de santé sont intéressant, mais sectoriels. Il y a un besoin plus général, et il manque sans doute quelque chose à notre droit. C'est un point qu'on ne plus ignorer aujourd'hui.

Comment les services producteurs de données voient-ils l'évolution de leur rôle ? Quelles sont leurs propres attentes et quelles sont celles qui se portent vers eux ?

Béatrice Sédillot (chef du service de la statistique et de la prospective du Ministère de l'Agriculture) : le service statistique du Ministère de l'Agriculture est un des utilisateurs historiques du CASD. Celui-ci a commencé se déployer en 2012, or c'est la date à laquelle s'est posée la question de la diffusion du recensement agricole 2010 - une opération qui a lieu tous les 10 ans, avec une somme considérable de données collectées et de demandes les concernant. Au-delà de notre propre utilisation, nous avons à cœur de donner accès aux données que nous produisons dans les meilleures conditions, tout en garantissant le respect du secret statistique. La création du CASD est une opportunité qui est arrivée au bon moment pour pouvoir mettre en œuvre cette diffusion. Avant, nous entretenions déjà des relations importantes avec le monde de la recherche, l'INRA par exemple, mais organisées sous la forme de conventions bilatérales et impliquant des procédures spécifiques de mises à disposition (sur postes sécurisés par exemple). Nous avons donc fait ce choix pour le recensement et mais nous l'avons aussi adopté pour toutes sortes d'autres enquêtes que nous réalisons et qui peuvent être mobilisées par la recherche.

Michel Isnard (chef de l'unité des affaires juridiques et contentieuses de l'Insee) : jusqu'en 2008, la loi, en vertu des textes de 1951, interdisait l'accès aux données des ménages même si des choses se faisaient via le réseau Quételet. L'arrivée du CASD a permis assez facilement cet accès, ouvrant la possibilité de faire des recherches sur les mêmes données que celles de l'Insee. Il s'agit d'une évolution qualitative et quantitative, dont l'impact commence à se faire sentir sur les travaux des chercheurs. Quant aux données confidentielles des entreprises, la France, à rebours de l'Europe, les diffuse depuis 1984.

Dans quelle mesure l'appétit des chercheurs pour les données modifie-t-il vos pratiques ?

Corinne Prost (chef de service, adjointe à la directrice de la DARES, Ministère du Travail et de l'Emploi) : notre rôle change un peu, la mise à disposition des données de la DARES depuis le 1er janvier 2016 bousculant nos habitudes. Nous avions des liens privilégiés avec les chercheurs qui venaient travailler chez nous sur nos données. Le CASD démythifie cet accès aux données en y donnant accès à plus de chercheurs, ce qui nous éloigne de l'utilisation qui en est faite tout en renforçant le formalisme : nous devons documenter davantage les données transmises, la façon dont nous les produisons, et guider les chercheurs, en vue d'une utilisation correcte. D'autres moyens vont devoir être trouvés pour maintenir le contact avec les chercheurs et avoir un retour sur l'utilisation des données.

Ce contact avec les chercheurs autour de la mise à disposition des données, est-elle à vos yeux une composante un peu marginale ou incontournable de votre rôle ?

CP : la diffusion des donnés ayant pris plus d'ampleur, cette démarche qui était déjà partiellement mise en place avant devient une mission plus centrale aujourd'hui, avec, surtout, un accompagnement est à organiser.

JG : en ce qui concerne les données fiscales, c'était un trésor bien « coffré » jusqu'à ce qu'intervienne le législateur. Nous savons, au CSS, que vous avez remarquablement joué le jeu.

Mais de l'intérieur comment cela a-t-il été vécu ?

Brice Lepetit (chef du bureau des études statistiques en matière fiscale, DGFIP) : c'est en effet pour nous une façon assez nouvelle de considérer nos données que de les communiquer, notamment les données individuelles car les données agrégées étaient déjà disponibles. Cela implique donc de faire évoluer les mentalités, et de revoir notre processus d'acquisition de données en questionnant leur qualité. Il s'agit aussi d'aller au-delà de la mise à disposition de données brutes en développant l'accompagnement des utilisateurs. Les spécificités en matière de législation et de gestion fiscales doivent être bien prises en compte pour une utilisation efficiente.

Le chercheur n'accède pas forcément à la base telle qu'elle est utilisée pour vos opérations.
Qu'est-ce-qui se passe entre temps ?

BL : outre l'anonymisation, il y a en effet un « nettoyage » nécessaire, les éléments de pure gestion sont retirés. Nous effectuons aussi quelques rapprochements pour être certains que les données fournies sont de qualité et pour s'assurer de leur valeur ajoutée.

Christine Chambaz (chargée de la sous-direction de la statistique et des études au secrétariat général du Ministère de la Justice) : en matière de justice, la diffusion des données demeure encore restreinte. Pour l'essentiel, les chercheurs viennent chez nous pour accéder à des fichiers qui sont d'une part volumineux et d'autre part perçus comme très sensibles au sein du Ministère. Nous rejoignons la DGFIP de ce point de vue et avons beaucoup à apprendre de leur démarche. Il s'agit aussi de données portant sur des champs très différents. Les données relatives à la justice civile sont perçues comme moins sensibles, encore que… quand elles concernent le familial par exemple, ou dans le domaine économique, du côté de l'entreprise. C'est surtout dans le registre pénal que les données sont considérées comme sensibles. Elles peuvent impacter la vie des personnes mais aussi des services. Dans notre activité de mise à disposition des données, il est donc important de concilier une approche « activité des services » et une approche « impact sur le justiciable ». La complexité de ces données doit nous pousser à assortir la diffusion de nos fichiers d'une reformulation de certains concepts ainsi que d'un accompagnement fort des chercheurs.

Par nature, les données de la justice sont en effet très sensibles mais d'un autre côté, elles présentent un intérêt social et scientifique considérable.
Ne va-t-il pas falloir définir une doctrine en la matière lorsque leur mise à disposition montera en puissance ?

CC : il faut déjà convaincre en interne des bonnes conditions de sécurité dans lesquelles nous pouvons le faire. L'existence du CASD représente un argument fort dans la discussion avec les directions détentrices des données, notamment celle du casier judiciaire. La mise à disposition peut d'ailleurs supposer d'accroître l'anonymat de ces données, notamment en restreignant le nombre de modalités des nomenclatures des affaires portées au casier. Par ailleurs, une des grandes difficultés, c'est que les données administratives construites à des fins de gestion sont assorties d'un certain droit à l'oubli, mais les informations anonymisées restent dans les fichiers statistiques, où l'on a du coup une mémoire plus longue du passé judiciaire des individus. On ouvre dès lors la possibilité d'évaluations qui vont au-delà des textes de loi. Par exemple, les personnes peuvent revenir devant la justice au-delà de la récidive légale - qui répond par nature à une définition inscrite dans la loi -, et le chercheur ou le statisticien pourra analyser ce retour alors qu'il a disparu de la mémoire du juge.

Comment allez-vous faire ? Procéderez-vous à une sélection des chercheurs ?

CC : si nos données devaient être déposées au CASD, les chercheurs devront de toute façon passer devant le CSS et il y aura une analyse fine de la finalité des recherches.
Mais a priori le CSS n'est pas compétent…

MI : cela dépend, il pourrait être compétent pour un certain nombre de sources mais pas pour la totalité.

JG : si certaines dispositions sont clairement fixées par la loi concernant le rôle du CSS, il faut bien comprendre que sur tout un pan des données recherchées par les chercheurs, il n'existe pas de dispositif, le CSS n'est pas concerné, du coup il faut aller voir ailleurs, mais qui ?

Faut il considérer le recours au CASD comme une voie unique et privilégiée d'accès aux données sachant qu'elle est très récente, qu'il y avait des voies et moyens avant..?

MI : deux remarques : il existe des voies autres comme les fichiers anonymisés de production et recherche qui passent par le réseau Quételet, et qu'il est nécessaire de continuer à utiliser. Le CASD est limité aux données confidentielles. Ensuite, j'évoquerais une expérience internationale en Allemagne qui a impliqué plusieurs institut de statistiques ayant réparti leurs données dans plusieurs centres d'accès sécurisés. La question que cela pose, c'est-ce-que ce sont des clones, est-ce qu'un seul endroit permet de simplifier les démarches ? Côté chercheurs, le fait de centraliser est important. Derrière, il y a aussi un problème de coût à prendre en compte : il faut que l'activité du CASD soit rémunérée. Quételet par exemple bénéficie d'une subvention du milieu de la recherche, ce qui permet de rendre l'accès gratuit pour les chercheurs, mais il a un budget de fonctionnement.

Comment prévoir des architectures pour le CASD si on n'y met pas le coût ?

BS : avant la création du CASD, nous mettions déjà nos données à disposition des chercheurs. Néanmoins, nous avons décidé depuis 2012 de privilégier le CASD car la procédure est simplifiée et sécurisée, avec un seul déchargement pour les producteurs et des modalités d'accès qui dépendent des projets de recherche. Ce choix a suscité les réticences de certains laboratoires qui travaillaient déjà avec nous car il représentait une contrainte supplémentaire pour eux.

Mais globalement, l'accès par le CASD favorise un traitement plus équilibré des différents laboratoires de recherche et instituts techniques. Si la question du coût d'accès est parfois soulevée par les chercheurs, le CASD est pour nous la voie privilégiée, avec des avantages indéniables.

CP : pour moi, il faut conserver plusieurs voies d'accès. Il est vrai que la tentation est forte d'avoir un seul interlocuteur comme le CASD pour éviter d'avoir à jongler avec plusieurs types de fichiers… Cependant, Quételet reste pertinent pour un certain nombre de fichiers, les contraintes assez fortes du CASD ne se justifiant pas pour des données anonymisées et donc qui ne sont pas indirectement nominatives.

BL : en tant que « jeune » diffuseur de données, le CASD nous permet de nous concentrer sur la production plutôt que sur la diffusion, avec des aspects de simplicité et de respect très fort de la sécurité qui le rendent très attractif.

CC : la question est double. Tout d'abord, le choix du centre d'accès dépend du type de donnée. Le CASD n'est pas un passage obligé, il peut y avoir des solutions plus simples pour certaines sources. Mais pour une source considérée, il est vrai que la déposer une fois à un seul endroit simplifie les choses.

Si le CASD est une solution technique de plus en plus développée qui devient un point de passage obligé pour certaines données, quid de l'égalité d'accès à cet outil ?

JG : le CSS reste vigilant sur les conditions techniques, les délais, la qualité, le coût que ça représente…

Plus généralement, quelles conséquences peut avoir le développement de l'accès aux données individuelles en termes de développement de la recherche universitaire, d'enrichissement démocratique, d'évaluation des politiques publiques ?

MI : pour ma part, je suis frappé par le fait que les travaux des chercheurs qui nous font des demandes soient autant en prise avec certains des sujets les plus sensibles des politiques publiques. Par exemple, parmi les projets examinés lors de la dernière réunion du CSS, figurent le bien être en France, le lien entre les inégalités de revenus et les dons aux partis politiques, le suivi des femmes à l'adolescence, l'impact du licenciement économique sur les carrières…
On observe une multiplication des dossiers avec des demandes très précises en termes d'impact sociétal, ce qui donne une idée de l'importance de l'ouverture des données. Sur l'évaluation des politiques publiques, elle peut se faire essentiellement a posteriori avec les données statistiques, en prenant en compte aussi leur délai de production.

CP : Cette ouverture est en effet très précieuse pour le débat démocratique. Il est intéressant de noter que le travail de recherche mentionnépar Thomas Piketty de mesure, grâce à des données appariées, de l'accès à l'enseignement supérieur en fonction du revenu des parents pourrait aussi bien être des résultats statistiques produits par l'Insee ou un service statistique ministériel. C'est aussi un avantage de l'accès élargi aux données: il peut jouer le rôle d'aiguillon pour le système de la statistique publique en l'incitant àdévelopper des approches plus novatrices.

BL : jusqu'ici, le Ministère avait le monopole de l'évaluation des dispositifs fiscaux, et l'analyse ex post des dispositifs se révélait relativement limitée. En jouant sur le capital humain, l'ouverture des données aux chercheurs va susciter des analyses plus fines et plus riches. Au delà de la question de la donnée, l'exemple de la mise à disposition récente du code source de l'impôt sur le revenu nous montre que ce type de démarche crée du lien social et de l'adhésion en misant sur la transparence et l'enrichissement mutuel.

CC : le fait de livrer des données à la recherche permet de profiter d'une forme de spontanéité des chercheurs, qui imaginent assez facilement de mettre en regard différentes sources. Une grande partie des données produites par l'administration à vocation de gestion ont le défaut d'être un peu sèches : les rapprochements permettront de les enrichir.

Quel lien éventuel peut-on établir entre la problématique d'ouverture des données administratives avec les affaires de type Wikileaks ou Panama Papers ?

JG : elles peuvent constituer un accélérateur pour mettre des données dans le domaine public. Mais attention, dans le cadre des Panama Papers, ce ne sont pas les données directes, incluant des données à caractère personnel, qui sont disponibles mais le retraitement effectué par un consortium de journalistes… Dans ces cas-là on est quand même loin de la statistique publique.

Franck von Lennep

Directeur de la Direction de la recherche, des études, de l'évaluation et des statistiques, Ministère des affaires sociales et de la santé

Directeur de la recherche, des études, de l'évaluation et des statistiques (DREES) au Ministère des Affaires sociales et de la Santé depuis mars 2012. Diplômé de l'ENSAE, Franck von Lennep a exercé plusieurs responsabilités : chargé de mission au Conseil d'orientation des retraites (COR) ; responsable du département veille et stratégie à la Caisse nationale d'assurance maladie (CNAMTS) ; secrétaire général du Haut conseil à la famille et conseiller en charge des comptes sociaux au cabinet du ministre du Budget et des Comptes publics.

Tous les sujets qui ont été présentés ce matin font partie de ceux sur lesquels nous travaillons pour faire avancer le projet porté par le Ministère de la Santé sur l'ouverture des données.
Celui-ci est inscrit dans l'article 193 de la Loi de modernisation de notre santé qui a été promulguée en janvier 2016. Nous travaillons maintenant sur les textes d'application. Dans le cadre de la préparation de cet article, j'avais co-piloté la commission open data et assisté à ce titre à la présentation du CASD. Nous avions alors beaucoup orienté les réflexions sur la sécurité des données, qui doit être garantie dans le cadre de la multiplication des usages, puisqu'ils peuvent être à fins d'épidémiologie, de surveillance…mais aussi individuelles, ciblées vers les professionnels de santé et les patients. Quant aux données dont il est question, elles peuvent être aussi bien de nature administrative que provenir de dossiers médicaux ou des patients eux-mêmes, à travers les objets connectés par exemple, même si l'article 193 régit en priorité l'accès aux données administratives, celles-ci revêtant une importance croissante pour les chercheurs. Les données de la CNAMTS sont issues des
« recours aux soins » et sont en effet riches d'enseignements sur la manière dont les gens sont soignés et sur leur état de santé.

La multiplication des usages des données de santé comporte plusieurs pré-requis.
Tout d'abord, il faut encore développer la mise à disposition des données anonymes, c'est-à-dire en général des données agrégées, qui sont d'ores et déjà publiées par les ministères et les agences publiques du secteur de la santé. Dans le même temps, la question de l'anonymat des données individuelles n'est pas réglée : dès qu'il y a beaucoup de données individuelles (même dé-identifiées), le risque de ré-identification augmente. Mais entre le zéro risque de l'open data et les données les plus réidentifiantes, il y a des niveaux de risques intermédiaires, qui concernent par exemple les données associées à des échantillons. Il faut donc continuer à travailler pour adapter le système de mise à disposition au niveau de risque des données en évitant la « sur-sécurité ».
Un autre pré-requis a trait à une question juridique : comment faire pour élargir le concept de finalité (concept nécessaire pour traiter des données personnelles), qui suppose souvent de définir des objectifs très précis ? En effet, quand on utilise des big data, on ne sait pas toujours très bien d'emblée ce que l'on cherche. Par ailleurs, l'article 193 recourt à la notion d'intérêt public : pour accéder aux données personnelles, il faut pouvoir démontrer que le traitement comporte un intérêt public. Cela impliquera de définir notamment la manière dont on l'envisage et l'applique pour les acteurs privés.
Troisième pré-requis: fluidifier les relations avec la Cnil pour améliorer le traitement des dossiers portant sur les données en santé. La nouvelle loi devant permettre d'augmenter le nombre de demandes et de recherches en santé, il faudra veiller à ce que cette augmentation ne se fasse pas au détriment des délais. C'est la raison d'être de notre volonté de constituer, en lien avec la Cnil mais aussi avec les utilisateurs, des méthodes de références et des autorisations cadres.
En quatrième instance se pose la question de la sécurisation des données. On le sait, le confinement et la traçabilité des accès aux données de santé n'ont pas toujours été de mise. Notre réflexion dans ce domaine s'inspire beaucoup de la démarche du CASD. Demain, il n'y aura pas que lui, et la réflexion devra notamment se poursuivre à la CNAMTS, qui occupe une position-clé en matière de données de santé. Sur les principes, c'est sans doute la première fois qu'une réflexion aussi approfondie est menée sur le référentiel de sécurité des données de santé pour les études et la recherche, et sur les moyens à mettre en œuvre pour les mettre à disposition plus facilement et évaluer, contrôler davantage a posteriori afin d'alléger les procédures à l'entrée.
Un autre enjeu complexe est celui des appariements. La possibilité de croiser des données de santé et d'autres types de données, notamment des données de l'Insee tel que l'échantillon démographique permanent de plusieurs millions de personnes, représentera une étape importante dans la connaissance et l'évaluation de notre système de santé. Aujourd'hui, les capacités opérationnelles sont limitées puisqu'il faut passer par l'Insee avec des moyens humains et techniques limités.

Dans la nouvelle perspective qui se dessine, le CASD aura un vrai rôle à jouer. Enfin, le modèle économique de l'accès aux données est également sujet à débat. Permettre cet accès dans des conditions sécurisées, traçables et auditables suppose d'y mettre le prix. Les données elles-mêmes ayant une valeur, comment est-il est possible d'extraire une partie de cette valeur pour financer ce coût ? Les financements publics étant sans doute insuffisants, d'autres pistes peuvent être envisagées, comme des « couches » de services de base gratuits, et d'autres, plus individualisés, tarifés.

En ce qui concerne l'ensemble de ces enjeux, nous sommes encore dans un « work in progress »: il faudra encore un an de travail avant qu'on ait des réponses.

De par sa position unique, le CASD ne risque-t-il pas de devenir un monopole marchand de données ?

Kamel Gadouche : il y a quelques années, il n'y avait de système d'accès sécurisé pour l'accès aux données, cela créait une grande insatisfaction des chercheurs qui avait très difficilement accès aux données, voire pas du tout pour la très grande majorité. Aujourd'hui il y en un, le CASD. On lui reproche d'être seul… Mais demain quand il y aura plusieurs centres, les reproches seront encore plus nombreux : Pourquoi ce n'est pas mutualisé ? Comment les chercheurs peuvent s'y retrouver ? Comment font les chercheurs qui veulent travailler conjointement sur des données qui sont hébergées séparément sur plusieurs centres ? Comment les chercheurs peuvent savoir où trouver les données sur quels centres et avec quelle documentation ? De mon point de vue, je le laisse à votre appréciation, c'est une véritable force que la France ait un seul centre pour l'accès aux données, favorisant ainsi la multidisciplinarité et des études plus innovantes que dans les autres pays (grâce notamment aux bases quasi-exhaustives et aux futures nouvelles possibilités d'appariement). Cette position pour le CASD n'est jamais confortable. Il est normal qu'il y ait une forte vigilance de la part de tous les acteurs sur la qualité de service et sur la tarification comme vous l'avez entendu ce matin par la voix du président du comité du secret statistique. La présidente du Cnis y est tout aussi vigilante et ne parlons pas des chercheurs.

Anonymiser les données : un cas d'usage du défi technique de mise en œuvre

Dominique Blum

Médecin responsable de l'information médicale

Praticien hospitalier actuellement en disponibilité, Dominique Blum a exercé les fonctions de médecin de DIM (département de l'information médicale) depuis 1984 dans plusieurs hôpitaux publics et pendant une brève période dans un groupe de cliniques privées. Spécialiste du PMSI dont il a, au Ministère de la Santé, accompagné la mise en oeuvre de 1990 à 2002, il consacre depuis quelques années une partie de ses recherches à la question de la confidentialité des bases nationales de données de santé.

En milieu hospitalier, le PMSI (programme de médicalisation des systèmes d'information), recueil systématique de données de santé, existe depuis 30 ans : c'est un recueil systématique des informations administratives et médicales relatives aux patients hospitalisés. Il est réalisé lors de la sortie d'établissement, qu'il soit public ou privé. Créé en 2005, la T2A (tarification à l'activité) se fonde sur ce recueil pour le financement des établissements. Pour des raisons techniques, tous les établissements doivent centraliser les données dans une agence, l'ATIH qui les traite et les retransmet à la CNAM TS. La base nationale de données a vocation de description d'activité qu'est le PMSI, et celle à vocation de tarification qu'est la T2A constituent certainement l'une des base de données médicales la plus importante au monde, ce qui lui vaut d'être hautement convoitée par les chercheurs en santé.

Si l'accès à ces données était rendu public, on pourrait imaginer un site sur lequel quiconque - un employeur, un assureur, un banquier, un voisin… - rentrerait un certain nombre d'informations « de notoriété publique » (état-civil, âge, sexe, date d'entrée et de sortie de l'hôpital…) concernant une personne ayant séjourné en établissement hospitalier, pour se voir fournir tous les diagnostics et actes le concernant. Et ce, y compris dans le temps, grâce au système de chaînage du PMSI qui crypte les identifiants mais permet de retrouver les informations relatives à toutes les hospitalisations d'un même patient. C'est dire l'importance du risque que représente la mise à disposition de ces informations et la nécessité d'en limiter l'accès.

Il est toutefois impossible de ne pas centraliser ces données, les objectifs de ce système l'exigeant tant en ce qui concerne le PMSI que le T2A. De surcroît, des données éparpillées perdraient leur intérêt scientifique. Les journalistes s'y intéressent beaucoup et y ont obtenu accès il y a quelques années - sans que l'on réalise le risque de réidentification - ce qui nous vaut le marronnier des palmarès des hôpitaux et cliniques, réalisé à partir des données du PMSI. Mais il s'agit aussi d'une base de connaissance pour divers opérateurs experts.

Certains penseront : « il suffit d'anonymiser les données ! » Mais l'anonymisation de ces données est déjà effective au niveau national, ce que l'on récupère, ce sont des résumés de sortie anonymisés (RSA), ne comportant ni l'état-civil, ni le NIR, ni la date de naissance, l'adresse, la date de séjour. Cependant, les informations disponibles sur chaque séjour associent l'identité de l'établissement d'accueil, le mois de sortie, l'année du séjour, le mode de sortie, l'âge à l'entrée, le sexe et le code de résidence… or une combinaison de ces données renvoie dans 89% des cas à une personne unique. Si l'on ajoute le chaînage anonyme, ce pourcentage de réidentification, monte à 100% pour tous les patients ré-hospitalisés. Il existe cependant une solution technique, consistant à réduire la précision des informations (en utilisant des classes d'âge par exemple) de telle sorte que les combinaisons soient « plus larges » et correspondent à un groupe tel de patients qu'on ne pourrait plus ré-identifier quelqu'un à coup sûr, de sorte qu'on pourrait même les utiliser en open data. Cette approche peut répondre à de nombreux usages, notamment les fameux palmarès ou encore des tableaux de bord rapides, certains travaux de recherche…

Mais ce « floutage » des données ne permettrait plus, en revanche, d'exploiter la base nationale dans le cadre de la T2A ou pour des travaux de recherche pointus, dont on estime que moins de 20% pourraient utiliser ces bases appauvries. Du coup, si on ne veut pas appauvrir ces données… on les met au CASD.

Une entreprise peut-elle recourir aux services du CASD ?

Antoine Frachot : c'est la loi qui s'applique et qui indique si tel individu dans telle entreprise a le droit ou pas de consulter les données conservées au CASD. Pour RTE par exemple, on est dans le cas d'un accès à ses propres données et la société reste souveraine dans la décision d'en permettre l'accès à des personnes extérieures. BNP Paribas, ErDF, Generali… travaillent également avec nous pour le traitement de leurs propres données.

Mais un laboratoire pharmaceutique, par exemple, pourrait-il avoir accès aux données de santé ?

Le projet de loi santé prévoit une ouverture aux acteurs privés, experts, PME, bureaux d'études…. mais aussi industriels des produits de santé et assureurs éventuellement. Il est sain que les premiers y aient accès car ils ont vocation à valoriser les données, et représentent potentiellement une concurrence bénéfique pour les acteurs publics. L'ouverture dans des conditions pas suffisamment maîtrisées aux seconds fait l'objet de certaines réticences, d'où la mise en place de dispositions particulières.

Le nouveau système national des données de santé

André Loth

Administrateur général (retraité), Ministère des affaires sociales et de la santé

Professeur d'économie puis administrateur civil (ENA 1989) au Ministère de la santé, André Loth a été successivement Chef de la mission Programme de médicalisation des systèmes d'information hospitaliers (1989-93), Chargé d'études en économie de la santé (Direction de la Prévision au Ministère de l'économie (1993-95), Responsable du projet SESAM-Vitale au Ministère des affaires sociales (Direction de la sécurité sociale 1995-99) puis à la CNAMTS (1999-2003, comme sous-directeur), Directeur du système d'information au CHRU de Lille (2003-2006), Chef de la mission d'informatisation de la santé (Ministère des affaires sociales 2007-2010), IGAS (2010-2011) puis DREES (2011-2016), chef du projet "accès aux données de santé" de 2012 à mars 2016.

La loi du 26/01/2016 encadre l'ouverture des données de santé à travers une modification du Code de la Santé publique et de la Loi informatique et libertés (LIL). Dans le code de la santé publique, la loi crée un Système national des données de santé regroupant les données de l'assurance maladie (le SNIIRAM), les données hospitalières (le PMSI) et les causes de décès, auxquelles devraient s'ajouter par la suite les données sur le handicap et un échantillon de données de l'assurance maladie complémentaire. Des règles sont définies pour l'open data d'une part et pour l'accès encadré d'autre part. Quant à la gouvernance, elle est partagée entre l'Institut National des Données de Santé, la CNAMTS, l'Etat, et la Cnil, qui reste décisionnaire en matière d'autorisations. Par ailleurs, c'est le numéro de sécurité sociale (NIR) qui est choisi comme Identifiant national de santé.

Du côté de la LIL est prévue une unification dans la mesure du possible des règles d'accès aux données de santé lorsqu'elles font l'objet de traitements à des fins de recherche, étude ou évaluation, la suppression du décret en Conseil d'Etat pour les appariements nécessitant le NIR en vue d'une recherche ou d'une étude en santé, et la possibilité d'alléger les procédures Cnil. Pour revenir au SNDS, qui n'inclut qu'une fraction des données de santé, il faut souligner que c'est un « trésor national ». On y retrouve les données individuelles sur les soins de santé à l'ensemble de la population française sur (potentiellement) 20 ans ainsi que des données sur les prestataires de soins. La richesse de cette base a vocation à se démultiplier grâce à des appariements avec d'autres bases, avec des enjeux considérables à la clé. Ils sont d'abord d'ordre sanitaire et économique, à travers l'évaluation des performances des prestataires de soins, des nouveaux traitements, l'effet des médicaments en vie réelle, l'information du public, etc… Mais ils sont aussi démocratiques via la connaissance des dépenses par pathologie, des effectifs et des revenus des professions de santé, des stratégies thérapeutiques, des inégalités de la prise en charge… Le système a mis du temps à se mettre en place, avec initialement une sous-estimation des usages, le SNIIRAM étant vu comme un simple moyen de renvoyer aux médecins et aux hôpitaux une vision de leur activité, des soupçons pesant par ailleurs sur la qualité des données. D'autre part les risques étaient eux aussi sous-estimés : l'exemple du PMSI montre que ce que l'on croyait anonyme ne l'était pas et que pourtant ni le ministère ni la Cnil ne voyaient d'inconvénient sérieux à diffuser sur cédéroms des données hospitalières individuelles, portant sur la partie la plus malade de la population et aisément réidentifiable (dans 89 % des cas, voire 100 % si on peut chaîner plusieurs hospitalisations) pour qui connaît le lieu et les dates approximatives d'hospitalisation, l'âge, le sexe et le code postal de résidence ; l'engagement de ne pas copier ni rediffuser ces données étant peu respecté, le fait qu'il n'y ait pas eu de fuite est sans doute dû au fait que peu de gens avaient l'expertise nécessaire pour traiter ces données..

Aujourd'hui, cependant, on approche de la maturité. La loi a repris la doctrine énoncée dans le rapport Bras de 2013, distinguant deux grands modes d'accès aux données de santé : les données anonymes sont en open data, tandis que les données indirectement nominatives donnent lieu à un accès encadré, autorisé par la Cnil, pour la recherche, les études et les évaluations. Entre les deux, il est prévu une procédure homologuée par la Cnil pour les données à faible risque de réidentification (sur le modèle du Réseau Quételet en sciences humaines et sociales). Reste à savoir où sont les frontières ?

En ce qui concerne l'open data, c'est-à-dire l'accès libre et gratuit aux données anonymes, le constat est celui d'une offre insuffisante et d'une sous-utilisation. Plusieurs institutions seront incitées à mettre en ligne davantage de données anonymes, après agrégation, échantillonnage ou floutage des données, ou via des systèmes de requêtes préformatées en libre accès. Mais à une double condition.

D'une part, le futur Institut national des données de santé (INDS) doit jouer son rôle pour créer du consensus, exprimer les besoins et faire œuvre de pédagogie. D'autre part, la Cnil, les chercheurs et statisticiens, les utilisateurs et les gestionnaires des bases de données… devront définir ensemble des éléments de doctrine sur ce qui est anonyme et ce qui ne l'est pas. Pour l'accès aux données « à caractère personnel » comme on dit (nominatives ou indirectement nominatives), les critères sont les bonnes personnes, les bonnes raisons et les bonnes conditions. Les bonnes personnes : public et privé sont à égalité mais des conditions restrictives s'appliquent aux industriels de la santé et aux organismes d'assurance ou de crédit qui dans le cas général ne pourront pas accéder directement aux données. Les bonnes raisons, il appartiendra à la Cnil d'en juger : il faut d'abord justifier que les données demandées sont nécessaires à l'étude ou à la recherche projetée (un comité d'experts instruira la demande et donnera son avis à la Cnil) ;il faut ensuite que la finalité du traitement soit d'intérêt public (des fins privées ne sont pas pour autant exclues, si un bénéfice collectif est aussi attendu par exemple pour une recherche sur le médicament). Là aussi, dans les cas (sans doute rares) où la question se posera, elle sera instruite par l'INDS qui donnera aussi un avis à la Cnil, en toute transparence. Quant aux bonnes conditions, cela vise d'abord la transparence (avec notamment une information sur toutes les études en cours puis la publication de tous les résultats, qu'ils soient positifs ou négatifs) et cela vise ensuite la sécurité, notamment la traçabilité des accès (qui accède à quoi ?) et cela vise enfin le droit d'opposition des personnes, garanti par la loi. Les demandes portant sur des traitements similaires ou sur des procédures conformes à des modèles déjà approuvés pourront faire l'objet de procédures allégées devant la Cnil. Il faut noter que le règlement européen remplacera en 2018 la procédure d'autorisation actuelle par une procédure de simple déclaration avec étude d'impact. Au total, ces mesures accentuent un peu les contraintes en ce qui concerne le PMSI et les extractions du SNIIRAM (où on n'admet plus de « perdre de vue les données »). Inversement, le verrou du décret en Conseil d'Etat pour tout appariement utilisant le NIR est levé, et la procédure pour les demandes d'autorisation à la Cnil est unifiée et clarifiée. Autre progrès notable : la sécurité crée de la confiance… Et on verra à l'usage le résultat des procédures Cnil simplifiées et la rapidité de l'instruction des demandes d'accès. Concernant la traçabilité et le confinement des données à caractère confidentiel, la santé s'aligne sur l'Insee et d'autres secteurs car l'accès des chercheurs aux données a pour contrepartie une exigence de traçabilité des sorties. On entend ainsi concilier l'accès des chercheurs (et assimilés) aux données sans sacrifier la protection de la vie privée. Un arrêté ministériel fixera le référentiel de sécurité du SNDS mais il est déjà acquis que nul ne pourra partir avec des données personnelles de santé sous le bras. Seront concernés la CNAMTS, l'ATIH et l'INSERM notamment, mais aussi d'autres organismes qui voudront eux aussi avoir « leur bulle » pour gérer et mettre à disposition leurs données.

En conséquence, les utilisateurs, contraints de travailler dans des espaces confinés distants mis à leur disposition par les gestionnaires de données souhaiteront disposer dans ces bulles de tous les outils qui leur sont nécessaires. Cela revient à externaliser tout ou partie de l'outil informatique pour un coût pas forcément plus élevé que l'informatique locale, voire moins si on intègre le coût de la sécurité. Ces "espaces de travail" seront donc à géométrie variable, avec des niveaux de service et de performance variés. Pour les organismes gestionnaires de données, au vu des attentes diverses des utilisateurs, il ne suffira pas de fournir un service standard… C'est pourquoi, au delà d'un service de base gratuit, les surcoûts liés aux demandes particulières des utilisateurs devront être payants pour contenir les coûts. Un service informatique distant, sécurisé, adapté et facturé selon les besoins de chaque utilisateur : c'est un métier à part entière, d'où l'intérêt de confier cette mission à un prestataire qui soit responsable de la qualité du service. Le CASD, qui a créé ce métier en France, a certainement des atouts. Et les économies d'échelle, la mutualisation du savoir faire, l'encadrement des tarifs plaident pour un prestataire unique. Cependant le cahier des charges ne sera sans doute pas exactement conforme au modèle actuel du CASD et la concurrence a aussi des mérites.

Le traitement à distance des données du SNDS se fera-t-il par le CASD ou un système propre ?

Je ne sais pas.

Faudra-t-il demander un accord du CPP (Comité de protection des personnes) en plus de la Cnil dans le cadre d'études non interventionnelles et rétrospectives ?

Nous avons essayé de faire simple, mais c'était juste le moment où entre en vigueur la Loi Jardé qui donne à des Comités de protection des personnes (les CPP) le soin d'autoriser les recherches biomédicales notamment et plus largement toutes les études prospectives. Mais comme beaucoup de recherches de ce genre feront appel aussi à des données existantes (par exemple celles du SNDS) cela risque de doublonner la procédure en amont de la Cnil. Le ministère est conscient de ce risque et le projet de décret sur la procédure est rédigé avec le souci de rendre les choses simples pour l'utilisateur.

Si la santé s'aligne sur l'Insee et les données fiscales, pourquoi développer un projet spécifique pour la santé ?

Le dispositif généraliste aujourd'hui est celui de la Cnil. Si vous voulez accéder aux données tirées des DADS, c'est l'article 25 de la LIL qui s'applique. Dans notre cas, nous n'avons rien changé à ces dispositions existantes, même si le dispositif actuel pourrait être simplifié.

Appariements de données et projet de loi sur le numérique

Jean-Pierre Le Gléau
Consultant CASD

À l'Insee, Jean-Pierre Le Gléau a eu en charge les aspects juridiques de la statistique. Il a participé au développement de textes ou de structures relatifs à l'organisation de la statistique publique : loi sur le recensement, création de l'Autorité de la statistique publique, réorganisation du Cnis et du comité du secret statistique, mise en place du CASD.

Lorsqu'on a deux fichiers de données individuelles, leur analyse séparée additionne leur contenu. En revanche, le croisement (appariement) de ces fichiers, individu par individu, multiplie leurs richesses. Pour effectuer cet appariement, il faut pouvoir repérer chaque individu par un identifiant commun aux deux fichiers. Celui-ci pourrait être le nom, mais on dispose rarement de cette information dans les fichiers et même dans ce cas, l'orthographe n'est pas garantie, sans oublier les risques d'homonymie. Quant à des données communes comme l'adresse, la date et le lieu de naissance, elles sont à nouveau source d'erreurs et d'homonymies. Reste l'appariement sur un numéro commun. Le plus universel, c'est le NIR, c'est-à-dire le Numéro d'Inscription au Répertoire national d'identification des personnes physiques, géré par l'Insee. Présent dans les fichiers sociaux et médico-sociaux, il est plus connu sous le nom de « numéro de Sécurité sociale ». Actuellement, l'utilisation du NIR est très protégée par la loi. L'article 27 de la loi Informatique et libertés stipule ainsi que le traitement de données via le NIR doit passer par un décret en Conseil d'État. Les chercheurs, appartenant souvent à un établissement public, devraient donc s'arranger pour qu'un ministre prenne un décret pour eux lorsqu'ils veulent traiter des données contenant le NIR. En pratique, cela n'est jamais arrivé…

Un accès plus facile au NIR demeure néanmoins souhaitable au vu de l'intérêt qu'il présente pour des travaux de recherche, sans renoncer à de fortes garanties sur le respect de la vie privée. Dans cette optique, deux possibilités se présentent.

La première consiste à traiter le NIR lui-même, après une autorisation de la Cnil, longue à obtenir, mais moins inaccessible qu'un décret en Conseil d'État. La deuxième consiste à « hacher » le NIR, par un processus irréversible, validé par la Cnil, en se contentant d'une simple déclaration à la Cnil. La première voie est exigeante sur le mode d'accès, tandis que la seconde apporte une restriction sur l'accès au NIR. Dans tous les cas, l'accès aux données se ferait sur un serveur sécurisé, type CASD.

L'article 18 du projet de loi pour une République numérique, d'ores et déjà adopté par l'Assemblée nationale, traite de l'assouplissement des conditions de traitement du NIR. Les chercheurs ont été partiellement entendus, puisque le projet d'article soumet à une autorisation de la Cnil l'utilisation du NIR à des fins de recherche scientifique ou historique, à la condition que de numéro d'inscription à ce répertoire ait préalablement fait l'objet d'un cryptage, un code spécifique étant dédié à chaque projet de recherche. L'ouverture aux chercheurs en fonction de leurs besoins est donc réelle, avec la fin du décret en Conseil d'État. Le projet d'article cumule cependant les deux inconvénients des possibilités souhaitées : la nécessité d'une autorisation de la Cnil parfois très longue à obtenir… et l'accès au NIR haché seulement, ce qui n'est toutefois qu'un inconvénient mineur. Il faut également souligner que ces règles se cumulent avec celles du reste de la loi, par exemple pour le traitement des données sensibles.

Présentation de la plate-forme de données bigdata sécurisée CASD-Teralab

Alexandre Marty
Responsable Data Science, CASD

Après une formation d'ingénieur en mathématiques appliquées, Alexandre Marty acquiert une solide expérience des technologies big data et de la data science au cours de plusieurs années au sein de startups au Canada. Il est aujourd'hui responsable data science au CASD, où il met en place des clusters big data sécurisés dans le cadre du projet Teralab.

Les big data se caractérisent par de grandes quantités de données qui concernent un nombre important de personnes. Elles requièrent des technologies et des méthodes de traitement puissantes et recèlent des opportunités considérables. Deux chiffres traduisent l'explosion de ces données: 2,5 milliards de giga sont générés chaque jour dans le monde, et 90% des données existantes ont été générées durant les deux dernières années. Parallèlement, leurs usages se multiplient (objets connectés, réseaux sociaux, santé, etc…). Ce phénomène implique aussi de grandes responsabilités liées à l'utilisation de données détaillées, personnelles et confidentielles et le recours à des technologies encore plus sécurisées. Le fort besoin de sécurité qui se fait sentir requiert des infrastructures à la hauteur.

C'est la raison d'être du projet Teralab, lauréat du programme des investissements d'avenir. Il consiste en la construction et l'exploitation d'une plateforme big data pour la recherche, l'innovation et l'enseignement, proposée par un consortium associant l'IMT (Institut Mines-Télécom) et le GENES (CASD), en partenariat avec l'Insee. Démarré en décembre 2013, le projet bénéficie d'un budget de 5,7 millions d'euros pour une durée de 5 ans. Il se compose de deux compartiments. Le premier, porté par l'IMT, garantit une sécurité de niveau industriel et fonctionne avec un accès par Internet. Le second bénéficie des conditions ultra sécurisées du CASD pour l'hébergement de données confidentielles. L'ensemble repose sur une infrastructure puissante et extensible grâce à un environnement Hadoop. Un autre intérêt du dispositif est qu'il est en clés en main, avec de nombreux outils fournis pour les data scientists.

Parmi les projets mis en oeuvre au travers de Teralab, on peut citer un POC (proof of concept) en partenariat avec l'Insee portant sur l'amélioration du calcul de l'indice des prix à la consommation à partir des données de caisse de plusieurs enseignes de grande distribution. L'expérimentation avec les technologies big data en utilisant des données simulées a produit des résultats concluants : le délai d'exécution d'une requête SQL s'est avéré de 40 secondes grâce à la technologie Hadoop contre 1h15 avec une base de données classique. Teralab travaille également sur d'autres projets comme l'optimisation du réseau de RTE, le Datalab BNP Paribas, ou encore l'anonymisation de données agrégées de grande échelle avec l'Insee.

Les projets européens et l'accès aux données sécurisées dans le contexte international

Roxane Silberman

Directrice de recherche CNRS émérite, conseiller scientifique au CASD

Roxane SILBERMAN, sociologue dans le domaine des migrations internationales, ancienne élève de l'École Normale Supérieure et directrice de recherches émérite au CNRS, est actuellement conseiller scientifique du CASD au GENES.

Son rapport « Les sciences sociales et leurs données » (La Documentation française, 1999) est à l'origine du Comité de concertation pour les données en sciences humaines et sociales (CCDSHS) (décret de création du 12 février 2001, abrogé le 15 novembre 2015). Elle en a été la Secrétaire générale et a dirigé le Réseau Quetelet jusqu'en 2014; ces deux entités ont eu un rôle central dans l'ouverture aux chercheurs des données issues de la statistique publique.

Coordinatrice du projet européen du 7e programme cadre Data without Boundaries, 2011-2015, membre du Directoire du CESSDA (Consortium of European Social Sciences Data Archives), l'infrastructure européenne pour les données en sciences sociales dont la TGIR PROGEDO est, avec le Réseau Quetelet, membre, elle oeuvre au niveau européen pour faciliter l'accès transnational en Europe pour la recherche aux données confidentielles.

Ce qui se passe en Europe en matière d'accès sécurisé aux données pour la recherche est important d'au moins trois points de vue. Cela permet évidemment de voir comment la France se situe par rapport aux autres pays. Mais au-delà, les évolutions au niveau européen proprement dit, en particulier sur le plan juridique, peuvent impacter le niveau national et donc dans le cas présent la France, Enfin nombre de projets de recherche ne peuvent se contenter des données d'un pays et utilisent également des données d'autres pays soit qu'il s'agisse d'obtenir l'accès aux données intégrées au niveau européen (données d'Eurostat notamment) soit qu'il s'agisse d'obtenir l'accès à des données sécurisées nationales sans être résident de ce pays, accès transnational donc. Pour comprendre les évolutions comme les difficultés, il faut en ce qui concerne le paysage européen, prendre en compte trois niveaux qui ne sont cependant pas totalement indépendants les uns des autres: le niveau national, le niveau européen et le niveau transnational. Ce qui se passe au niveau national détermine au départ les évolutions au niveau européen et conditionne l'accès transnational. Inversement, le niveau européen une fois consolidé vient impacter les situations nationales et peut faciliter l'accès transnational.

Depuis les années 90, des accès sécurisés permettant aux chercheurs de travailler sur des données confidentielles se multiplient partout dans les pays de l'Union Européenne et les pays associés. C'est un mouvement de fond même si les situations sont encore inégales. On peut observer partout des changements sur le plan juridique avec l'introduction de dispositions prenant en compte les besoins de la recherche mais aussi des interprétations des cadres juridiques qui peuvent varier et être plus ou moins favorable à la recherche. Sur le fond, les débats, notamment sur l'équilibre sécurité/besoin des chercheurs, sont tout à fait similaires. C'est aussi un paysage extrêmement varié du point de vue des modes d'accès, sur site uniquement ou bien à distance mais avec des modes différents permettant un accès aux données elles mêmes (remote access) ou bien uniquement la soumission de programme (remote execution) sans accès direct des chercheurs aux données. Dans ce paysage, la France apparaît tardivement avec le CASD, si l'on prend l'exemple du Danemark (pour les pays à registre, l'anonymisation des donnée plus difficile a pu être un facteur favorisant plus tôt la mise en place d'accès sécurisés) ou bien des Pays-Bas, mais elle connaît ensuite un développement rapide avec une avance technologique que lui a probablement permis son arrivée récente dans le domaine. Les différences existant entre pays même si le mouvement de fond est le même, se traduisent par des compromis nécessaires au niveau européen proprement dit et par une grande variété de situations quant à ce qui est possible en matière d'accès transnational.

Au niveau européen proprement dit, on distinguera deux questions, l'une portant sur l'encadrement juridique de la protection des données personnelles et de la confidentialité, l'autre portant sur l'accès pour les chercheurs aux données européennes intégrées (notamment données Eurostat).

En ce qui concerne l'encadrement juridique européen, on est en train de passer d'une directive à un règlement qui donc va avoir force de loi au plan national également.

La directive de 1995 sur la protection des données personnelles avait fait une place à la recherche et a eu un impact plutôt positif pour les transcriptions au plan national dans les lois sur la protection des données personnelles. Cela a notamment été le cas en France avec la transposition (très tardive) dans la loi Informatique et Libertés en 2004. Cela s'est traduit également de façon collatérale dans le champ de la statistique publique où des dispositions plus favorables à la recherche ont été introduites pour l'accès aux données très détaillées. Le règlement européen acté en 2015 et qui devrait entrer en vigueur en 2018 a suscité inversement des inquiétudes du côté de la recherche, suite à une série d'amendements adoptés par le parlement européen, avec de nombreux débats (recherche vs intérêt public, accord express notamment pour les données sensibles dans le domaine de la santé, durée de conservation des données, des identifiants pour les panels) et des interventions y compris au niveau des Research Councils et ministères de différents pays. A la différence de la directive, il a force de loi au niveau national. Si l'on peut d'un côté estimer qu'un cadre homogène européen devrait faciliter la circulation intra-européenne des données pour la recherche, il subsiste des inquiétudes avec un risque de moins disant pour la recherche. Une des demandes qui avait été faite lors de l'intervention collective en faveur de la recherche était de pouvoir conserver les dispositions nationales pour la recherche si elles étaient plus favorables. Il est pour l'instant difficile d'estimer les impacts comme les marges possibles au niveau national.

Une autre grande question au plan européen proprement dit, est celle de l'accès des chercheurs aux données européennes, notamment celles d'Eurostat, données intégrées au niveau européen qui sont appuyées sur des collectes nationales. Permettant de mener des travaux comparatifs avec des données harmonisées, elles ont bien évidemment une grande importance pour la recherche (Enquêtes Forces de travail, panel européen SILC notamment). Elles étaient jusqu'ici souvent peu exploitables sur bien des sujets du fait d'une agrégation très forte de nombreuses variables avec de surcroît une procédure extrêmement longue pour obtenir ces données.. L'adoption, au terme de longues discussions qui ont vu s'opposer les points de vie nationaux, du nouveau règlement européen de 2013 pour l'accès des chercheurs à ces données, a permis sur le principe, des avancées, notamment une terminologie améliorée et surtout la possibilité de mettre en place un d'accès à distance sécurisé et distribué pour les données détaillées qui serait appuyé sur des points d'accès nationaux en lieu et place d'un accès uniquement sur site à Luxemboug.

En l'état cependant, ces progrès sont pour l'instant faibles. Dans le domaine de la procédure d'accréditation, on peut même parler de plus grande difficulté avec une procédure d'accréditation encore plus longue (y compris pour les fichiers très agrégés) car en deux temps (l'institution d'abord, puis le projet de recherche et l'équipe de chercheurs). Après un projet (ESSnet DARA auquel le CASD avait notamment participé), l'accès sécurisé distribué (qui ne pourrait au départ être appuyé pour les points d'accès que sur les instituts nationaux de statistique) tarde pour sa part à se mettre en place.. Dans l'attente, n'existe pour les chercheurs que la possibilité d'aller à Luxembourg pour disposer d'un accès sur site extrêmement limité. Une part de ces difficultés réside dans les points de vue nationaux différents dont sont porteurs les instituts nationaux de statistiques et dont dépend tout accord au niveau d'Eurostat.

Le troisième plan à considérer est celui de l'accès transnational à des données nationales confidentielles. Il s'agit là des données qui ne sont pas intégrées et harmonisées au niveau européen. Un champ particulièrement intéressant pour la recherche et très large est celui des données administratives. Pour ne prendre qu'un exemple, il existe dans de très nombreux pays en Europe des données permettant de marier les informations sur les employés et celles sur l'entreprise dans laquelle ils travaillent, du type DADS en France, très utilisées par les chercheurs. Qu'il s'agisse d'utiliser des données originales d'un autre pays car non disponibles ailleurs, de comparer les données de plusieurs pays si possible en menant une analyse conjointe sur l'ensemble des données et non des analyses séparées pour chaque pays, ou encore d'enrichir les grandes enquêtes européennes telles que SHARE, GGP (Gender Generation Program) ou des cohortes épidémiologiques avec les sources administratives nationales de l'ensemble des pays impliqués, les intérêts potentiels sont nombreux. Or en l'état, les chercheurs se heurtent à une difficulté centrale mise en avant par les juristes: l'absence d'accès juridique transnational pour un accès transnational à des données nationales. Le cœur du raisonnement avancé est celui de la difficulté à pouvoir poursuivre et appliquer une sanction en cas de rupture de confidentialité : comment et qui peut-on poursuivre ? Observons d'abord que l'accès transnational progresse dans les faits cependant chaque jour bien que selon différents modes. Le CASD pour la France par exemple propose un accès transnational à distance. C'est le cas aussi de CBS, l'institut national de statistique néerlandais. Pour le Royaume-Uni, un chercheur français devra par contre se rendre sur place. Mais la diversité actuelle des procédures d'accréditation (qu'il faut multiplier pour un projet demandant des sources de plusieurs pays avec plusieurs équipes impliquées), celle des équipements, des systèmes d'accès qui opèrent tous en parallèle, est inopérante pour la recherche.

Face à ces difficultés, des solutions possibles s'esquissent au travers de projets européens et de discussions menées dans différents cercles : le projet européen du 7ème programme cadre Data without Boundaries 2011- 2015 (coordonné par le Réseau Quetelet), les discussions menées dans le cadre du Working group on statistical confidentiality (Système statistique européen, Eurostat), ou encore l'Expert group on microdata access, initié par l'OCDE. Les conclusions s'accordent sur le fait que la plupart des cadres juridiques n'excluent pas formellement l'accès transnational pour les données très détaillées, le problème étant plutôt celui de leur interprétation, le cœur du problème étant en réalité celui de la sécurité et des sanctions en cas de manquement. Les équivalences en terme d'encadrement juridique dans chaque pays, en ce qui concerne les sanctions notamment, et de sécurité, peuvent permettre des accords sur un transfert de responsabilités (un cercle de confiance) de nature à permettre un « transfert » des données hors frontières nationales. Harmoniser les procédures d'accréditations ou au minimum les formulaires de demandes et s'accorder là encore sur des accords entre parties dans le cas d'accréditation multiples pour un même projet n'est pas non plus hors de portée : toutes reposent sur des principes identiques même si une divergence subsiste sur le point d'accréditer en premier lieu l'institution ou simplement le chercheur.

Enfin il possible et il faut aussi de créer un système sécurisé distribué européen permettant aux chercheurs de travailler avec des équipements et des procédures identiques sur des sources de plusieurs pays en même temps avec d'autres chercheurs et éventuellement de mener des analyses sur l'ensemble des données ce qui implique de pouvoir les rassembler pour le moment de l'analyse. Le CASD a dans le cadre de DwB construit le « proof of concept » pour un tel système sécurisé distribué avec un scénario de recherche qui porterait sur les perceptions différentes du rôle des femmes en Allemagne et en France et qui demanderaient l'utilisation des données détaillées de l'enquête Emploi en France (au CASD), celles du Microcensus à DESTATIS et celles de la European Value disponibles au GESIS en Allemagne... Reste enfin à harmoniser les pratiques sur l'anonymisation, le contrôle des sorties, les pratiques de surveillance des chercheurs et à faciliter un accès unifié aux métadonnées, en anglais et standardisé. Une première réalisation a vu le jour en ce sens, CIMES (Centralizing and Integrating Metadata from European Statistics) dans le cadre de Data without Boundaries, avec une collaboration ADISP/CMH et CASD. Un chantier avec au cœur la construction d'un réseau sécurisé européen distribué pour lequel un projet a été déposé dans le cadre des programmes H2020. Au-delà la question se pose, également, d'un élargissement possible hors UE avec notamment le problème des Etats-Unis sur la protection des données personnelles.

Quelques développements en cours aux USA et au Canada

Lars Vilhuber

Directeur exécutif du Labor Dynamics Institute de Cornell University, Président du conseil scientifique du CASD

Dr. Lars Vilhuber est membre du corps professoral du Department of Economics à la Cornell University et Directeur exécutif du Labor Dynamics Institute au sein de l'ILR School à Cornell University. Il détient un Ph.D. en Sciences économiques de l'Université de Montréal. Ses intérêts de recherche portent sur la limitation de la divulgation statistique, les effets et les causes des licenciements massifs, la mobilité des travailleurs, et la dynamique des marchés de l'emploi local. Au fil des années, il a acquis une grande expertise et appréciation de la nécessité de rendre accessible des données pour analyse par des chercheurs en sciences économiques et autres sciences sociales.

Pour commencer par « ce qui est déjà arrivé », on a pu constater une explosion de l'accès aux données sécurisées. Elle se fait à partir de centres de données de recherche physiques aux USA et au Canada. Au Canada, ce sont 28 centres qui donnent accès aux données démographiques de Statistiques Canada, tandis que l'accès aux données d'entreprise s'effectue par le centre canadien d'élaboration de données et de recherche économique à Ottawa. On peut envisager d'y faire un travail de master ou de doctorat car les procédures d'accès sont très rapides. Plusieurs centres proposent des données fédérales, mais aussi provinciales et municipales.

Aux Etats-Unis, l'accès est plus limité mais progresse de façon plus fluide. Il existe aujourd'hui 23 lieux donnant accès non seulement aux données du Census Bureau, mais aussi désormais aux données de santé du National Center for Health Statistics (NCHS) et de l'Agency for Healthcare Research and Quality (AHRQ). En dépit de la centralisation opérée, une certaine diversité du réseau subsiste, avec des locaux situés dans le Bureau of Labor Statistics, dans le National Center for Health Statistics et à l'IRS, tous les trois situés à Washington. Le Census Research Data Center Census Bureau est devenu aujourd'hui le Federa Statistical Research Data Center qui regroupe les agences déjà citées mais aussi Internal Revenue Service, le Bureau of Justice Statistics, l'Environmental Protection Agency, parmi d'autres à venir… pour constituer un véritable réseau de la statistique nationale. Il reste encore à harmoniser les procédures d'accès et de sortie, en l'absence d'équivalent de la Cnil.

A l'extérieur de cet enclos sécurisé, se sont montées des procédures expérimentales pour accéder à des données synthétiques complexes. Elles bénéficient d'un certain succès, avec un potentiel d'expansion. Une subvention va permettre de travailler sur le même modèle au Canada. Nous travaillons d'ores et déjà sur des données synthétiques allemandes grâce à l'utilisation d'un terminal du Centre d'accès de l'Institute for Employment Research (IAB) allemand, situé sur le campus de Cornell (un parmi 6 en Amérique du Nord).

Cela demande une validation des données confidentielles en arrière-plan, mais la démarche, qui fonctionne sur un principe hybride entre l'accès à distance et l'exécution à distance, est prometteuse. Une chercheuse a ainsi pu mener une étude sur l'impact du revenu de l'homme et de la femme sur le comportement des ménages. « Ce qui va arriver, maintenant…»

Au Canada, le « questionnaire complet » est de retour dans le recensement canadien, ce qui signifie plus de données pour les centres de recherche. La décision a été prise par le nouveau Premier ministre qui l'avait inscrite dans ses promesses électorales…

Aux Etats-Unis, la préparation du recensement de 2020 ouvre de nouvelles opportunités de recherche, avec l'utilisation d'instruments d'enquête dynamique et celle, accrue, des données administratives.

« Quant à ce qui pourrait arriver… » aux USA, ce ne sont pas plus de deux à trois nouveaux centres qui ouvrent chaque année mais l'intégration des données s'améliore avec l'entrée des données IRS, BLS, etc. dans le FSRDC, ce qui va faire augmenter la demande par des chercheurs.

Au Canada, on est en attente d'une subvention en vue de financer une toute nouvelle infrastructure pour un accès plus facile aux données d'entreprise. Des changements des lois (au niveau fédérale ainsi qu'au niveau de plusieurs provinces) sont en outre débattus en faveur d'un accès plus soutenu aux données médicales et provinciales.

« Ce qu'il faudrait qu'il arrive ? » On note dans les articles publiés une utilisation croissante des données administratives généralement accessibles uniquement par les centres de recherche. Ce qui pose un défi sous-jacent : que se passe-t-il si le referee (d'un journal scientifique) demande à vérifier les

résultats en vue d'une publication ou si un chercheur veut répliquer les résultats publiés par un autre ? Par exemple, selon une étude menée par notre institut, sur les 109 articles parus en trois ans dans l'American Economic Journal : Applied Economics, environ 40% utilisent des données confidentielles.
Or il est impossible de vérifier les résultats, ce qui est pourtant nécessaire : même si on veut refaire tourner les programmes qui y ont abouti et que la volonté de les fournir est là, le taux de réussite est loin de 100%. La vérification par la communauté scientifique est donc en jeu. Les articles à données confidentielles sont également davantage cités que les articles à données publiques, ce qui laisse à penser qu'ils sont plus intéressants… mais c'est à vérifier. Or l'accès reste problématique. Le CASD par exemple permet un accès formel. Dès lors qu'on ne peut transmette les données utilisées, faut-il se fier à la réputation des chercheurs, ou est-ce de la naïveté ? Parfois, ceux-ci offrent la possibilité d'aller retrouver le programme utilisé sur un site, mais quelques années après, il risque de ne plus exister. Si on veut reproduire la même chose, on peut extraire le programme, mais comment le donner ?

Donc, que peuvent faire les diffuseurs d'accès ? Tout d'abord, pour éviter des délais trop longs et des démarches complexes, il faudrait une procédure simplifiée quand le but est la validation, avec une finalité déjà validée. Cette piste est discutée aux Etats-Unis. Ensuite, des archives référentiables publiques devraient être disponibles pour les programmes et résultats divulgués, en incluant les résultats qui n'ont pas été utilisés dans la publication. Enfin, des archives référentiables confidentielles pour les données de base, assorties de métadonnées non confidentielles et ouvertes, pourraient faire l'objet d'un accès soumis à approbation.

Roxane Silberman : un cadre européen devrait faciliter les échanges transnationaux. Avec les Etats-Unis, on n'est pas dans le même espace juridique, comment faire ?

Sophie Vulliet-Tavernier : la Commission européenne est en négociation avec les autorités américaines pour rebâtir un accord respectueux des droits et libertés individuels, sachant qu'à l'heure actuelle la politique transfrontalière rend théoriquement impossible la transmission de données. Un avis de la Cnil doit être rendu la semaine prochaine sur l'efficacité de cet accord, mais il n'est pas dit qu'il sera très favorable.

Pouvez-vous nous préciser ce que recouvre les données dites synthétiques et les modes d'accès aux différents types de données ?

LV : les données synthétiques offrent un mode supplémentaire d'accès à des données anonymisées, mais ce sont des micro-données, pas du tout agrégées, visant une certaine fiabilité analytique. Sur tous les modèles mettant en œuvre ces données, on retrouve 10 à 35% des résultats validés de la même façon avec les données confidentielles qu'avec les données synthétiques. C'est trop faible pour être la base d'une publication mais suffisant pour servir comme outil d'exploration de données. On peut y accéder sans se déplacer, grâce à un mode hybride entre l'exécution à distance et l'accès direct.
Aux Etats-Unis, il existe une liste assez longue de données disponibles, surtout d'entreprises, mais il est aussi possible d'accéder aux données confidentielles du recensement et à l'équivalent des DADS, donc on peut faire des jumelages. La documentation sur la disponibilité de ces données reste toutefois toujours le maillon faible. Il faut savoir que les producteurs de données sont attachés à garder un contact avec les chercheurs. Concernant le Census Bureau, la justification de l'accès à ses données par les chercheurs repose légalement sur leur utilité pour l'organisme, donc il est nécessaire de documenter ces bénéfices possibles.
En ce qui concerne l'accès transfrontalier, ce serait très pratique si l'accès par le CASD était possible à partir des Etats-Unis, mais un accord avec ce pays est toujours basé sur la symétrie de l'information, la législation américaine ne permettant pas de sortir des données ou de les visualiser hors des « enclos » situés sur le territoire américain. C'est la même chose au Canada. Cependant, cela n'empêche pas ces deux pays d'échanger des données d'import/export par exemple, sur le principe « je mesure ce qui rentre chez toi et vice versa ».

Le GENES

GENES : la datascience au service de la société

Le Groupe des Écoles nationales d'économie et de statistique (GENES), établissement public d'enseignement supérieur et de recherche rattaché aux ministères économiques et financiers regroupe deux écoles – l'ENSAE ParisTech et l'ENSAI à Rennes –, ENSAE-ENSAI Formation continue (anciennement CEPE), le Centre de recherche en économie et statistique (CREST-ENSAE/ENSAI), une filiale dédiée à la valorisation de la recherche et à la vente de conseil et d'expertise (DATASTORM) ainsi que le Centre d'accès sécurisé aux données (CASD). Toutes les entités du groupe partagent une caractéristique commune : avoir la datascience inscrite dans leur génotype.

ENSAE ParisTech : la Data au cœur de la formation d'ingénieur

Créée il y a plus de 70 ans, l'Ecole nationale de la statistique et de l'administration économique (ENSAE ParisTech) forme des ingénieurs spécialisés en économie, statistique, finance, actuariat et Data Science. Les diplômés de l'ENSAE possèdent des compétences scientifiques, techniques et humaines les rendant aptes à mesurer, analyser et modéliser, en univers incertain et risqué, des phénomènes économiques, financiers et sociaux, à tirer parti des « Big Data » disponibles dans tous les secteurs d'activités, pour évaluer, prévoir et décider. Ils sont donc présents dans tous les domaines où la modélisation économique et statistique est un enjeu stratégique.

L'ENSAE ParisTech délivre annuellement environ 150 diplômes d'ingénieurs, propose une offre de mastères spécialisés en actuariat, Data Science, économie, finance et gestion des risques, et contribue à la formation Master et doctorale de l'Université Paris-Saclay, dont elle est membre fondateur.

Ensai : "Modelling data, creating knowledge"

L'Ecole nationale de la statistique et de l'analyse de l'information (Ensai) est implantée depuis 1996 sur le campus de Ker Lann, aux portes de Rennes. Experts en analyse et traitement de la donnée, les ingénieurs de l'Ensai peuvent capitaliser sur une formation scientifique et opérationnelle qui répond clairement aux besoins des entreprises. L'école forme chaque année 90 ingénieurs statisticiens qui sauront traiter et exploiter les données, les modéliser pour les faire parler afin d'éclairer la décision dans l'entreprise. A l'issue des deux premières années de formation qui leur permettent d'acquérir un solide socle de compétences « statistique-informatique-économétrie», les élèves ingénieurs intègrent l'une des 6 filières de spécialisation : gestion des risques, marketing quantitatif, génie statistique, biostatistique, territoire et santé, data science.

A la rentrée 2015, l'Ensai a par ailleurs ouvert un Master international big data également accessible aux salariés en formation continue.

CREST - Centre de Recherche en Economie et Statistique

Le Centre de Recherche en Economie et Statistique est depuis janvier 2016 l'UMR joignant les forces de recherche du GENES (ENSAE et ENSAI) en économie, statistique, sociologie quantitative et finance avec celles de l'Ecole Polytechnique en économie avec le soutien du CNRS. Les recherches menées au CREST combinent théories statistiques, théories économiques, questions sociologiques, données innovantes et méthodes économétriques pour produire des travaux appliqués permettant par exemple d'évaluer des politiques publiques ou d'analyser de grands ensembles de données. Le Labex ECODEC "réguler l'économie au service de la société" est le lieu de collaboration entre le CREST (ENSAE et Polytechnique) et HEC.

Ensae-Ensai Formation Continue (Cepe)

L'Ensae-Ensai Formation Continue (Cepe) est l'entité de formation continue du Groupe des Ecoles Nationales d'économie et Statistique (Genes). L'objectif principal du Cepe est de délivrer des formations exigeantes et de qualité via des contenus scientifiques innovants et des formateurs de premier plan, toujours experts dans leur domaine d'intervention.

Être compétent, aujourd'hui, dans son travail nécessite d'acquérir des compétences techniques, comportementales, et également sectorielles. L'Ensae-Ensai Formation Continue enrichit chaque année ses domaines d'intervention afin de répondre au mieux aux préoccupations du marché. Le catalogue du Cepe couvre ainsi un grand nombre de thématiques liées à l'entreprise : statistique, économie, prospective, finance-actuariat, corporate finance, techniques de communication, marketing quantitatif et data science…

Datastorm : expertise et consulting en Datascience et en Economie

Filiale de droit privé détenue à 100% par le GENES, Datastorm a été créée en 2013 pour porter toute l'expertise des chercheurs et ingénieurs du groupe après des entreprises. DataStorm a ainsi permis à des entreprises de multiples secteurs (énergie, transport, banque, assurances, secteur public, etc.) de bénéficier de solutions pointues d'accompagnement (Big Data, DataScience, Econométrie, Evaluation, DataViz, conseil en architecture de calcul, etc.) sur de larges domaines d'application : Marketing quantitatif et CRM, Big Data & Open Data, Finance et actuariat, Revenue Management, Optimisation des processus et des organisations, Maintenance prédictive et sûreté de fonctionnement, etc.

LES PARTENAIRES

Insee

L'Institut national de la statistique et des études économiques est une direction générale du ministère de l'Économie et des finances. Il a pour mission de collecter, analyser et diffuser des informations sur l'économie et la société française sur l'ensemble de son territoire.
Il conduit ses travaux en toute indépendance professionnelle.
Pour mener à bien ses missions, il mobilise des compétences variées et recrute chaque année pour de nombreux métiers des fonctionnaires et des contractuels.

PROGEDO

Ayant pour but la production et la gestion des données en sciences humaines et sociales, PROGEDO réunit les acteurs concernés par les enquêtes quantitatives autour d'une politique nationale commune animée par deux grandes dimensions : la production et la mise à disposition de données pour les SHS. Le CASD participe au volet mise à disposition des données confidentielles.
Destinée à doter la France d'une infrastructure comparable à ses équivalents européens, PROGEDO est impliqué dans trois consortium européens constitués ou en cours de constitution comme ERIC autour des banques de données, (CESSDA - Consortium of European Social Science Data Archives) et des enquêtes européennes (ESS et SHARE).

l'ENS Cachan

L'École normale supérieure de Cachan, école des métiers de la recherche et de l'enseignement supérieur, s'inscrit dans la tradition d'excellence des Écoles normales supérieures et offre une formation culturelle et scientifique de très haut niveau.
A la fois une école et un centre de recherche, l'ENS Cachan rassemble treize laboratoires de recherche reconnus au niveau national et international, en sciences fondamentales, en sciences humaines et sociales et en sciences pour l'ingénieur. Trois instituts fédératifs promeuvent les interactions entre ces laboratoires, les soutiennent dans les actions interdisciplinaires et font l'originalité et la force de l'ENS Cachan.
Ainsi, la singularité de l'ENS Cachan est de rassembler des disciplines qu'aucun autre établissement d'enseignement supérieur ne rapproche de cette manière et à ce niveau. C'est dans cet environnement qu'élèves et étudiants de l'ENS Cachan reçoivent une formation disciplinaire renforcée «à la recherche et par la recherche», ouverte sur l'international et la pluridisciplinarité, qui les mène au master et au doctorat.
L'ENS Cachan est membre fondateur de l'Université Paris-Saclay, qui fédère les forces en formation et en recherche de 19 grandes écoles, universités et organismes de recherche en un ensemble scientifique de dimension internationale.

HEC PARIS

Spécialisée dans le domaine de l'enseignement et de la recherche en management, HEC Paris offre une gamme complète et unique de formations aux décideurs de demain : le programme de la Grande Ecole, les Mastères Spécialisés, les MSc, l'université d'été, le MBA, l'Executive MBA, TRIUM Global Executive MBA, le Doctorat et une large gamme de programmes pour cadres et dirigeants.

Créée en 1881 par la Chambre de commerce et d'industrie de Paris, HEC Paris, membre fondateur de l'Université Paris-Saclay, rassemble 138 professeurs à temps plein, plus de 4 400 étudiants et 8 000 cadres et dirigeants en formation chaque année.

HEC Paris est classée 2ème business school dans le classement général des business schools européennes, publié par le Financial Times en décembre 2015.

ÉCOLE POLYTECHNIQUE

Largement internationalisée (30% de ses étudiants, 39% de son corps d'enseignants), l'École polytechnique associe recherche, enseignement et innovation au meilleur niveau scientifique et technologique. Sa formation promeut une culture d'excellence à forte dominante scientifique, ouverte sur une grande tradition humaniste.

À travers son offre de formation – cycle ingénieur polytechnicien, master, programme doctoral, doctorat, formation continue – l'École polytechnique forme des décideurs à forte culture scientifique pluridisciplinaire en les exposant à la fois au monde de la recherche et à celui de l'entreprise. Avec ses 22 laboratoires, dont 21 sont unités mixtes de recherche avec le CNRS, le centre de recherche de l'X travaille aux frontières de la connaissance sur les grands enjeux interdisciplinaires scientifiques, technologiques et sociétaux. L'École polytechnique est membre fondateur de l'Université Paris-Saclay.

Références juridiques évoquées durant la conférence

LOI n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques, sa modification en 2008 par la LOI n° 2008-696 du 15 juillet 2008 relative aux archives.

LOI n° 2013-660 du 22 juillet 2013 relative à l'enseignement supérieur et à la recherche, modifiant le livre de procédures fiscales pour l'accès aux données fiscales pour la recherche.

LOI n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé.

Décret n°84-628 du 17 juillet 1984 relatif au Conseil national de l'information statistique et portant application de la loi n° 51-711 du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistique.

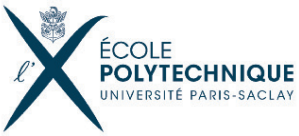

Insee


INVESTISSEMENTS
D'AVENIR


PRO
GE
DO
DATA
INFRASTRUCTURE


ens
CACHAN
université
PARIS-SACLAY


HEC
PARIS


ÉCOLE
POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY


GENES
Groupe des écoles
nationales d'économie
et statistique


ENSAE
ParisTech
École nationale
de la statistique
et de l'administration
économique
université
PARIS-SACLAY


ENSAI
École nationale
de la statistique
et de l'analyse
de l'information


CEPE
ENSAE-ENSAI
Formation continue


DataStorm
EXPERTISE & CONSULTING