Memo historique sur la création du CASD

*Roxane Silberman*

*8 février 2017*

La création du CASD en France s’inscrit dans la continuité d’un effort d’ouverture aux chercheurs des grandes bases de données de la statistique publique et administrative, entrepris dans le milieu des années 80 à un moment où la France est plutôt en retard sur ce plan par rapport à d’autres grands pays. La situation est d’autant plus dommageable pour la recherche française que l’université de son côté investit peu dans la production de grandes enquêtes académiques à la différence par exemple du Royaume-Uni, de l’Allemagne ou des Etats-Unis, tandis que l’INSEE, l’institut national de statistique, institut au caractère scientifique plus marqué qu’ailleurs produit au contraire nombre de grandes enquêtes très riches pour l’analyse que l’institut exploite produisant nombre de travaux au-delà de la simple production de statistiques. Une caractéristique du reste de l’INSEE est l’existence en son sein d’activités de recherche qui se traduisent dès 1982 par des allocations pour des doctorants puis la création d’un département de recherche, le CREST en lien avec les écoles nationales de statistiques regroupées in fine dans le cadre du GENES, Groupe des Ecoles Nationales d’Économie et de statistique. Jusqu’à l’autonomisation récente du GENES comme établissement d’enseignement supérieur, un nombre significatif de membres de l’INSEE peuvent ainsi faire de la recherche mobilisant les sources de la statistique publique très difficilement accessibles aux universitaires autrement qu’au cas par cas, le plus souvent par relations personnelles. Les doctorants ayant obtenu des allocations de recherche bénéficient également de ces accès et découvrent lorsqu’ils obtiennent un poste à l’université ces difficultés et l’inégalité croissante entre statisticiens publics, chargés d’études au sein du système statistique et universitaires. Quelques fichiers de grandes enquêtes et de recensements deviennent progressivement accessibles sous forme payante.

Alors même qu’une première phase d’ouverture se met lentement en place, à l’initiative d’un laboratoire du CNRS, le LASMAS (Laboratoire d’analyse secondaire et de méthodes appliquées à la sociologie) avec la passation à partir de 1986 de conventions entre le CNRS et l’INSEE puis avec quelques départements statistiques des ministères permettant l’accès à l’ensemble des chercheurs du CNRS des recensements et de quelques grandes enquêtes, la fin des années 90 voir surgir de nouvelles restrictions. La CNIL, l’autorité en charge de la protection des données personnelles encadrée par la loi Informatique et Libertés de 1978 restreint l’accès aux données infracommunales en particulier pour les recensements. Ces restrictions qui s’annoncent pour le recensement de 1999 touchent particulièrement les géographes et les sociologues travaillant sur la ségrégation urbaine. Pour accéder à ces niveaux de détail, certains chercheurs s’engagent dans des collaborations avec les aménageurs urbains qui bénéficient de dérogations pour les territoires sur lesquels ils travaillent. Alors que la Loi Informatique et Libertés de 1978, l’une des premières en Europe sur la question de la protection des données personnelles en Europe, votée après une tentative de l’INSEE d’apparier des données administratives (projet SAFARI), avait peu mobilisé les chercheurs, les discussions qui s’engagent sur la transposition dans la loi française de la directive européenne de 1995 qui fait une place à la finalité de recherche, débouchent sur la création d’un intergroupe « Statistique et recherche » qui réunit chercheurs et statisticiens privés et publics.

Les questions posées par l’accès des chercheurs à des données directement ou indirectement identifiantes sont désormais clairement posées. Le rapport de 1999 commandé par le ministre en charge de la recherche « Les sciences sociales et leurs données » (Lettre de mission à Roxane Silberman) pour faire un état des lieux et des recommandations sur ces questions d’accès aux données pour la recherche, pourtant centré sur l’ouverture des fichiers anonymisés qui est loin d’être réalisée, fait ainsi déjà une place aux difficultés grandissantes liées à l’élévation du niveau d’anonymisation (agrégations de modalités, suppressions de variables) des quelques fichiers jusqu’ici disponibles. Le rapport cite notamment l’organisation d’un accès sécurisé à Statistique Canada et l’annexe 6 présente une esquisse d’un accès à distance qui pourrait être utilisé pour le Recensement de 1999.

Les années qui suivent et qui mettent en œuvre certaines des recommandations du rapport ne voient cependant pas l’émergence d’un tel accès sécurisé alors que de tels systèmes commencent à apparaître, notamment au Danemark et aux Pays-Bas. L’attention se concentre d’abord sur le contenu des nouveaux Fichiers de production et de recherche (FPR), anonymisés mais réservés aux chercheurs, proposés désormais par l’INSEE pour ouvrir plus systématiquement l’accès aux données de la statistique publique aux chercheurs dans des conditions identiques à celles des chargés d’études travaillant au sein du système de statistique publique. Le Centre Quetelet, le « Data archive » français en charge de cette diffusion bataille ainsi très régulièrement pour obtenir plus de détails dans ces fichiers. Les avancées, sans être nulles, sont cependant faibles sur ce point.

Paradoxalement, les données portant sur les entreprises, données qui sont généralement considérées comme difficilement anonymisables sont depuis 1984 accessibles pour les chercheurs après passage devant un Comité du secret. Bien que la procédure soit jugée lourde et complexe, nombre de chercheurs peuvent ainsi disposer de ces données qui leur sont directement transmises sur CDRom.

## Les perspectives ouvertes par la transposition en 2004 dans la loi Informatique et Libertés de la directive européenne sur la protection des données personnelles de 1995, avec plusieurs années de retard, permettent de rouvrir le débat sur l’accès des chercheurs aux données confidentielles. Dès 2003, le Comité de concertation pour les données en sciences humaines et sociales (CCDSHS) qui a été mis en place en 2001 dans la suite du rapport Silberman pour organiser la concertation entre la recherche et les producteurs de données au niveau gouvernemental sous la présidence du ministre en charge de la recherche, se saisit de la question des centres sécurisés. Faisant suite à une recommandation de son conseil scientifique présidé par Roxane Silberman, le secrétariat général du Comité soumet ainsi au Conseil scientifique puis au Comité qui réunit les producteurs de données dont l’INSEE et plusieurs services statistiques ministériels une note sur les centres sécurisés au Canada et aux Etats-Unis, pointant également le retard de la France en la matière. La question de l’accès aux données administratives et celle de l’appariement sont également posées dans cette note. La question est reprise à la séance du Conseil scientifique de décembre 2004 qui recommande la création d’un centre sécurisé et de procédures d’accréditation des chercheurs.

En 2004, en lien avec ces discussions, le Centre Quetelet organise une visite au Centre interuniversitaire québécois de statistiques sociales (CIQSS) de Montréal de même qu’à l’ISQ (Statistique Quebec) à Québec qui a mis en place un accès à distance utilisé par des chercheurs français. Il organise dans la foulée le 27 janvier 2005 une Journée thématique : Secret statistique et protection de la confidentialité des données. La matinée suivante du 28 janvier, réservée à un séminaire fermé avec des représentants de l’INSEE est consacrée à la présentation de centres sécurisés étrangers dont notamment celui de l’accès à distance de Statistique Quebec.

Si l’INSEE se montre intéressé et indique avoir l’intention de créer un accès sécurisé, les discussions avancent cependant lentement alors que la pression des chercheurs s’accroit avec pour seul débouché quelques conventions individuelles, le plus souvent pour les chercheurs les plus reconnus, les associant à des équipes internes et leur permettant avec un statut pour une période déterminée, d’employé du système statistique, de travailler sur des données très détaillées dont ils ont besoin.

La transposition enfin réalisée en 2004 de la directive européenne dans la loi française ouvre la voie à des discussions plus précises. Cependant que le CREST, le département de recherche de l’INSEE, produit en juin 2006 une première note (Stefan Gregoir) résultant d’un groupe de travail interne rassemblant des représentants de différentes directions de l’INSEE et détaillant plusieurs configurations possibles (en référence notamment à l’exemple de Statistics Denmark qui a mis en place un accès à distance) et un premier chiffrage, des discussions s’engagent fin 2006 entre le ministère de la recherche qui préside le Comité de concertation pour les données en sciences humaines et sociales et l’INSEE, associant également le département SHS du CNRS pour examiner comment les instances de la recherche pourraient soutenir la création d’un tel centre sécurisé. Un premier chiffrage est ainsi transmis avec l’aval de la direction de l’INSEE à la DGRI du Ministère en charge de la recherche dans l’idée de la création d’un partenariat recherche/statistique publique pour une telle création et la mise en place d’un groupe de travail à cet effet. Issue de ce processus, en juin 2007, une note du CREST avalisée par la direction générale de l’INSEE et constituant l’offre faite par l’INSEE à la recherche pour monter un partenariat avec un financement recherche (et notamment des postes) est transmise à la direction du département SHS du CNRS.

Les avancées tardant, en mai 2008, le GENES et son directeur Alain Trognon essayent de faire avancer le dossier. Lors d’un Comité de direction, l’INSEE demande au

GENES de conduire une expérimentation d’un tel centre, et de lui en présenter des résultats début 2009. L’équipe en charge de cette expérimentation sous le direction de Kamel Gadouch, après une démarche initiale de s’orienter vers une solution purement logicielle (Microsoft, VMWare, Citrix…), la considère comme inadaptée, peu sûre et très coûteuse (déploiement, maintenance, assistance, intégration, droits d’administration…) et choisi de raisonner en termes de matériel et non plus exclusivement de logiciel. C’est le concept de boîte noire dédiée et verrouillée : la S-Box. Cette phase pilote s’accompagne de la mise en place d’un appel à propositions pour des projets de recherche qui est proposé à trois établissements, Paris School of Economics , Toulouse School of Economics et l’INED. 11 projets de recherche sont sélectionnés. L’expérience de ces quelques mois permet de vérifier que l’environnement de travail remplit bien les deux conditions requises : il donne aux chercheurs le moyen de conduire efficacement leurs recherches, et il protège bien les données confidentielles. Pour ce programme, les chercheurs signent un contrat pour être considérés comme employés par l’INSEE. Le pilote et l’expérimentation jouent un rôle important pour emporter la conviction des différentes autorités, CNIL et Archives notamment et inscrire dans la loi sur les archives de 2008 une modification de la loi sur le secret statistique de 1951 qui permet désormais d’étendre aux données confidentielles relatives aux ménages le système de dérogation et d’autorisation qui fonctionne depuis longtemps pour les données relatives aux entreprises et étendre la mission du Comité du secret statistique en matière d’habilitation des chercheurs aux données ménages .

Dans ce nouveau contexte juridique, le Comité de direction de l’INSEE décide de poursuivre le développement du centre sécurisé distant de l’INSEE, qui sera en charge de tout accès aux données confidentielles de l’INSEE. En 2009, le dispositif expérimental est ainsi transformé en un centre permanent, appelé à accueillir les recherches que le Comité du secret aura approuvées.

Le succès du projet déposé en 2010 auprès du Programme EQUIPEX des investissements d’avenir (lancé par le gouvernement français dans le contexte de la crise financière de 2008) par le GENES, devenu désormais un établissement public indépendant d’enseignement supérieur et de recherche, permet d’apporter, partenariat avec plusieurs autres établissements (HEC, Polytechnique, ENS Cachan) et le Réseau Quetelet qui participent à l’EQUIPEX, les financements nécessaires à la phase de construction (2011- 2019), le modèle économique adopté avec une participation des utilisateurs devant permettre à cette date au centre de date de s’autofinancer.

La montée en puissance du CASD tant en nombre de producteurs utilisant le centre pour mettre leurs données confidentielles à disposition des chercheurs qu’en nombre d’utilisateurs en France et en Europe est soutenue par l’ouverture sur le plan juridique qui se poursuite dans les années suivantes avec en 2014, la modification du Livre des procédures fiscales qui permet l’accès aux données administratives fiscales, en 2016 la nouvelle loi sur la santé qui comporte un article destiné à faciliter l’accès aux données medico-administratives, enfin la même année, la Loi sur pour une République numérique avec un article ouvrant l’accès pour la recherche à l’ensemble des données administratives couvertes par des secrets et un second article facilitant les appariements sur le NIR (numéro d’identification nationale) pour les chercheurs. Si le recours au CASD n’est pas inscrit dans la loi (sauf pour les données fiscales), les producteurs pouvant recourir à d’autres systèmes sécurisés, on peut penser que le développement actuel et l’expérience acquise du CASD attireront nombre d’entre eux comme le montre déjà quelques conventions en cours de signature. Le CASD est également bien placé pour opérer en tant que tiers de confiance pour les appariements sur le NIR à la demande des chercheurs.

Dans cette histoire, le CREST qui a bénéficié initialement d’un statut favorisé pour accéder aux données, du fait de sa position de composante de l’INSEE, a finalement joué un rôle moteur pour la création du CASD. Il est aujourd’hui un utilisateur important du CASD, mais passe par les mêmes procédures que les autres chercheurs et ne bénéficie plus d’aucun statut plus favorable.