Rapport Silberman

Les sciences sociales et leurs données

TABLE DES MATIÈRES

Avertissement

Ce rapport, qui fait suite à la mission qui m'a été confiée par Monsieur Claude Allègre, Ministre de l'Éducation Nationale, de la Recherche et de la Technologie, s'inscrit dans le cadre du travail entrepris il y a une vingtaine d'années au Centre d'Études Sociologiques par le Département d'Analyse Secondaire, et poursuivi avec la création du Lasmas par Alain Degenne. En gérant pour le CNRS une convention avec l'Insee, signée pour la première fois en 1986, ce laboratoire a commencé à organiser un cadre de diffusion aux chercheurs des enquêtes issues de la statistique publique. La réflexion présentée dans ce rapport et les propositions qui y sont faites n'auraient pu exister sans le travail accumulé depuis au sein de ce laboratoire et les relations construites avec l'Insee en tout premier lieu, mais aussi dans d'autres lieux de la statistique publique, le Céreq et la Darés notamment.

Ce travail a cependant trouvé aujourd'hui ses limites. Des quelques enquêtes engrangées au départ, le fonds actuellement disponible est passé à plus d'une centaine d'enquêtes dont plusieurs recensements. Des difficultés au départ peu sensibles ont pris progressivement de l'importance. La diffusion est pour l'instant restreinte aux seuls laboratoires du CNRS alors que la demande des universitaires s'accroît. D'autres apparaissent, liées en particulier à l'inquiétude grandissante sur la protection de la vie privée, menacée par la puissance sans cesse accrue des outils informatiques. L'accès des chercheurs aux données infra-communales du Recensement est devenu de ce fait problématique. Résoudre ces difficultés demande un cadre et une organisation différente, des moyens plus importants qui doivent être inscrits maintenant au niveau d'une politique de recherche nationale. La France est en retard sur ce point de plus de vingt ans sur plusieurs de ses voisins européens qui ont construit, à l'image de ce qui avait été entamé aux États-Unis, de puissants *Data Archives* pour la recherche en sciences sociales. Elle est également absente des grandes enquêtes européennes et internationales universitaires.

Il m'a paru indispensable d'associer à cette réflexion le cidsp-bdsp, qui à l'initiative de Frédéric Bon, quelques années avant le Lasmas, a entrepris des efforts analogues dans le champ de la sociologie politique, archivant et diffusant des données issues des enquêtes universitaires, plus fréquentes dans ce domaine, et celles de certains instituts de sondage. Construire un instrument de diffusion des grandes enquêtes, véritable télescope pour les sciences sociales, n'aurait guère de sens si l'on n'englobait pas d'emblée un champ très large.

Bruno Cautrès, directeur du cidsp-bdsp a donc été associé dès le départ au petit groupe constitué par Alain Degenne, directeur du Lasmas de 1986 à 1998, Annick Kieffer, Ingénieur de recherche au Lasmas et moi-même, pour piloter cette mission. Dans cette tâche, j'ai bénéficié de la collaboration et du soutien de nombreux organismes de la statistique publique. Mes remerciements vont particulièrement à Alain Godinot et Michel Glaude de l'Insee. Ils vont aussi à Hugues Bertrand et Philippe Méhaut du Céreq, à Claude Seibel de la Darés, qui le premier avait permis la signature d'une convention CNRS-Insee. Partout la mission a reçu un accueil favorable et a pu recueillir un état précis de la situation ainsi que des suggestions précieuses.

La mobilisation des milieux de la recherche a également été très forte. Les réponses à des questionnaires pourtant longs ont été nombreuses, qu'il s'agisse des Instituts de recherche ou

des laboratoires du CNRS et des Universités, témoignant de l'intérêt porté par les chercheurs à cette mission. Irène Fournier, Marie-Odile Lebeaux et Alexandre Kych ont eu la lourde tâche d'organiser cette enquête et d'en rendre compte.

Jacques Lautman, Bruno Cautrès, Alain Degenne, Alain Chenu, Michel Forsé ont présidé et rapporté sur les groupes de travail.

Denise Lievesley, Simon Musgrave, Repke De Vries, Margaret Adams, Paul Bernard ont apporté leur expérience de l'archivage et de la diffusion des données au Royaume-Uni, aux Pays-Bas, en Allemagne, aux États-Unis et au Canada.

Bruno Péquignot et Richard Topol pour le CNRS et Antoine Lyon-Caen pour la Direction de la recherche ont suivi le déroulement de cette mission.

Je dois une mention particulière à Annick Kieffer qui a apporté toute sa connaissance des *Data Archives* à l'étranger, à Isabelle de Lamberterie et à son équipe pour l'aide apportée sur les questions juridiques, à René Padieu pour ses éclairages incisifs sur ce point et sa relecture sans concession du rapport, à Jacques Lautman qui m'a soutenue dès le départ dans cette mission, à Benoît Riandey, organisateur à l'Ined du séminaire de Méthodologie d'enquêtes, à Marie-Odile Lebeaux et Jocelyne Léger qui ont assuré la finalisation de ce rapport. Alain Degenne a contribué tout au long de la direction du Lasmas pendant 12 ans à faire mûrir la réflexion. Enfin cette mission n'aurait pu être organisée sans le concours constant de Michèle Amiot.

Très nombreux ont été tous les autres qui m'ont apporté leur aide. On en trouvera la liste plus loin. Ils témoignent du consensus très fort qui s'est dégagé autour de cette mission, sur les principes comme sur les modes d'organisation nécessaires pour mieux réguler le rapport des chercheurs en sciences sociales à leurs données.

Roxane Silberman

La lettre de mission

Le Ministre

Paris, le 14 janvier 1999

Madame la directrice,

La France produit de nombreuses grandes enquêtes intéressant les sciences sociales. Cependant cette production s'est développée essentiellement au sein du monde administratif, de façon extérieure à la recherche et à l'université ou dans quelques grands instituts, fortement liés à l'administration. A la différence de l'étranger, tant le CNRS que les Universités n'y ont que très rarement été directement associés, en partie faute d'investissement financier.

Toutes les disciplines en ressentent les conséquences. En sociologie l'enquête exploratoire de terrain est ainsi restée le domaine privilégié des chercheurs. En même temps la France accuse un assez net retard en matière de sociologie quantitative, tant sur ses voisins européens que par rapport au monde anglo-saxon. En économie, les travaux de conjoncture souffrent par exemple des délais d'accès aux données. D'une manière générale, les chercheurs et les enseignants ont accédé lentement et difficilement aux données.

La mission assignée par le CNRS au Lasmas (upr 320) en 1986 a contribué à faire évoluer la situation, en permettant, à travers des conventions, d'acheter des enquêtes avec droit de diffusion à l'ensemble des laboratoires du CNRS. Des collaborations entre chercheurs et producteurs de données se sont également amorcées et ont contribué à faire évoluer certaines enquêtes.

La circulation des données de même que la participation directe des chercheurs à la production de ces données demeurent cependant insuffisantes. Leur diffusion aux laboratoires universitaires, leur utilisation pour la formation à la recherche, leur circulation entre instituts producteurs eux-mêmes ont rencontré des obstacles institutionnels, juridiques et financiers. Dans le même temps les contraintes de coût tendent à remettre en cause la production même de certaines enquêtes, souvent les plus proches des besoins de la recherche.

Le développement des recherches fondées sur des comparaisons internationales, la construction européenne qui accélère ces travaux de comparaison et d'harmonisation sont un nouveau défi qui rend d'autant plus urgent un rapprochement des instituts de recherches, des universités et des instituts producteurs de données. La question de la circulation des données, dans le cadre de collaborations européennes ou internationales, est d'ores et déjà à l'ordre du jour.

Après avoir :

- donné quelques éléments de comparaison sur la situation dans quelques grands pays européens concernant la production, l'archivage et la diffusion des enquêtes auprès de la communauté académique,
- tracé les évolutions à court et moyen terme qui risquent de peser tant sur la production que sur la diffusion des données (protection des données individuelles, évolutions des moyens informatiques, construction et harmonisation européenne).

Vous devrez :

- faire un état des lieux en France et identifier les principales difficultés,
- proposer, en sondant les partenaires possibles, le cadre d'une collaboration entre le CNRS, les Universités et l'Insee en particulier, permettant d'améliorer et d'accroître la diffusion des données pour les sciences sociales et d'associer plus directement la recherche à la production et l'amélioration des données.

Pour mener à bien cette mission vous bénéficierez du concours des diverses directions du Ministère de l'Éducation nationale, de la Recherche et de la Technologie concernées par le thème de cette mission, en particulier la Direction de la Programmation et du Développement et la Direction de la Recherche.

Je souhaiterais obtenir ce rapport au plus tard à la fin du mois de mai 1999.

En vous remerciant, je vous prie de croire, Madame la directrice, à l'expression de ma considération distinguée.

Claude Allègre

Madame Roxane Silberman
Directrice du LASMAS-IRESCO
59, rue Pouchet
75849 Paris cedex 17
s/c de Madame Catherine Brechignac
Directrice générale du CNRS

I. Les grands enjeux

I.1. Changement social et expertise

Accompagnant la croissance rapide de la demande d'expertise aux sciences sociales, la production de données s'est considérablement développée.

Les sociétés occidentales connaissent des mutations profondes. Pour la France elles prennent place dans le cadre de l'intégration européenne. Face à ces mutations, la demande d'expertise a crû et tout laisse penser qu'elle s'accroîtra encore. Cette demande d'expertise est naturellement celle des instances gouvernementales et gestionnaires. Mais elle est aussi celle des citoyens et en cela fondatrice de la démocratie. Le passage à une société de l'information et du savoir rend plus vive que par le passé cette demande d'expertise qui s'adresse très directement aux sciences sociales. Celles-ci se trouvent ainsi dans la situation paradoxale d'être de plus en plus sollicitées dans le même temps qu'elles se trouvent parfois contestées tant sur leur rigueur que sur leur capacité de cumulativité. C'est par exemple le cas de la sociologie.

Les données pouvant permettre de fonder une expertise sont aujourd'hui très nombreuses. On a assisté depuis la fin de la seconde guerre mondiale à une véritable explosion de leur collecte, qu'il s'agisse des enquêtes diligentées par l'État et ses services, par les chercheurs ou par les instituts de sondage, des enregistrements liés à l'administration ou de ceux que génère l'activité économique. Parce qu'elle a permis de stocker plus facilement des informations (sous condition d'assurer une veille informatique) et parce qu'elle a ouvert des possibilités de traitements rapides et complexes sur des fichiers de taille importante, la révolution informatique a fait croître de façon exponentielle ces données qui constituent aujourd'hui de véritables gisements. On est ainsi passé de la part des utilisateurs, et notamment des chercheurs, d'une demande de données agrégées en grande partie publiées ou de tableaux à façon, à une demande d'accès aux fichiers primaires de micro-données[1] qui ouvrent des possibilités nouvelles et bien plus grandes de traitements. C'est plus particulièrement de ces données qu'il sera question dans ce rapport, même si les problèmes qui se posent à leur propos peuvent être énoncés dans des termes assez proches pour d'autres ensembles de données. Parmi ceux-ci, ceux qu'utilisent par exemple l'histoire ou l'archéologie pour les sciences sociales, l'épidémiologie pour les sciences de la vie, ont donné lieu à des débats, des procédures, voire des institutions qui peuvent alimenter la réflexion. D'autres, comme la

[1]. Données non agrégées, concernant les unités statistiques de base de l'échantillon enquêté, en général des individus ou des ménages.

question des données issues d'entretiens, très utilisés par les sciences sociales, commencent à faire l'objet d'attention à l'étranger.

Gérer et exploiter au mieux ces gisements sont les problèmes des prochaines décennies. Ceci pose à la fois des questions d'organisation (il faut conserver), d'expertise scientifique (on ne peut tout garder) mais aussi des questions juridiques (propriété intellectuelle, protection de la vie privée) qui ont pris aujourd'hui, du fait de la révolution informatique, un relief particulier. Dans ces questions de régulation des gisements de données, la recherche pose des problèmes spécifiques qu'il importe de traiter si l'on veut que les sciences sociales soient à même à la fois de produire en tant que sciences et de répondre à la demande d'expertise sociale que le citoyen, les politiques et les corps sociaux sont en droit d'attendre d'elles. Ces deux questions sont distinctes mais liées.

I.2. Les sciences sociales et leurs données

Dans le domaine de l'archivage et de la mise à disposition pour la recherche en sciences sociales, la France est très en retard par rapport à la Grande-Bretagne, aux États-Unis et à l'Allemagne.

Si l'on compare les sciences de l'homme et de la société à d'autres disciplines, plus anciennement constituées et plus fortement orientées sur l'analyse empirique, on constate immédiatement une particularité des premières : les sciences de l'homme et de la société sont avant tout des sciences d'observation et l'expérimentation au sens strict n'est que rarement possible pour elles. Cette particularité est parfois partagée avec d'autres sciences, l'astronomie par exemple. Elle conditionne néanmoins fortement et sous de multiples aspects le travail de recherche en sciences sociales : la méthode expérimentale des chercheurs de ces disciplines est constituée de procédures de recueil d'observations et de leur analyse. Ces procédures ne trouvent sens que par leur réplication dans le temps ou l'espace. L'accumulation des observations et leur cumulativité constituent pour les chercheurs en sciences sociales des formes de contrôle expérimental.

Disposer d'observations recueillies dans des cadres de recherches, stocker ces observations en vue d'analyses y compris secondaires[2], c'est-à-dire par d'autres, contrôler de manière rigoureuse les procédures de production et de stockage des données, représentent des conditions sine qua non pour que les recherches en sciences sociales puissent articuler, au même titre que les autres sciences, théorie et objectivation. La production, la disponibilité et le traitement des micro-données provenant de fichiers de grande taille constituent de ce point de vue un enjeu de premier plan.

2. On appelle analyse secondaire d'une enquête ou d'une source administrative l'exploitation ultérieure des données soit dans une même visée d'analyse que celle qui avait présidé à la collecte, soit à des fins différentes.

Si les gisements de données, qui peuvent être considérés comme des gisements de connaissance sur la société, sont aujourd'hui très nombreux, les chercheurs ne sont directement à l'origine que d'une très petite fraction d'entre elles. Ils sont en particulier très dépendants des données produites par la statistique publique ou générées par l'activité administrative ou économique, qui apparaissent partout comme un aliment essentiel des sciences sociales. Il n'est que de rappeler ici qu'un ouvrage fondateur comme *Le suicide* d'Émile Durkheim prend appui sur ce type de données. La France dispose sous ce rapport, avec son institut national de statistique, l'Insee, d'un instrument souvent envié à l'étranger.

L'accès à ces données, et surtout leur réutilisation, ne va cependant pas de soi. De même que ne va pas de soi l'accès pour un chercheur à d'autres types de données ; celles produites par d'autres chercheurs le plus souvent avec de l'argent public, celles relevant de la sphère privée des instituts de sondage ou celles, croissantes, générées par l'activité économique et administrative.

Le rapport particulier que les chercheurs en sciences sociales entretiennent avec leur données implique qu'il faut aborder trois points de façon simultanée : leur place dans la production de ces données, la régulation de l'accès à celles produites par d'autres, et la formation aux méthodes d'analyses de ces données. Utiliser de façon rigoureuse des données implique nécessairement de contrôler ou de bien connaître les conditions de leur production. L'attention aux méthodes d'analyse va de pair avec celle accordée à la construction des enquêtes (champs, méthode de collecte, procédure d'échantillonnage etc.).

L'ensemble des dispositifs qui permettent de répondre à ces questions relève d'une politique de la recherche et de moyens de long terme à mettre en regard avec les « grands équipements » aux coûts sans commune mesure, dont disposent d'autres domaines scientifiques. Reprenant l'analogie entre le statut de l'observation en sciences sociales et en astronomie, on peut parler de véritables « télescopes » qui restent à mieux structurer, consolider, voire créer en France. Le retard pris en ce domaine par notre pays est évident si l'on compare la France à d'autres grands pays, les États-Unis, la Grande-Bretagne et l'Allemagne en particulier.

La diversité des disciplines scientifiques relevant des sciences humaines et sociales n'a pas permis à ce jour d'avoir une vision globale permettant de repérer les zones de force et de faiblesse, les retards les plus importants vis-à-vis d'autres pays occidentaux, les actions à mettre en œuvre. Le bilan qui va suivre, les perspectives d'avenir qu'il permet de tracer, devraient alimenter une réflexion collective des sciences sociales françaises. Disposer d'enquêtes, d'observations plus largement, est un enjeu clé au moment où d'importants réseaux de recherche

européens se mettent en place et développent de manière significative des programmes de recherche comparative. Dans ce contexte d'européanisation de la recherche, la collecte d'abord, l'organisation et la conservation ensuite, le traitement et l'analyse des données enfin constituent les trois piliers indispensables d'une politique scientifique audacieuse permettant à la recherche française de tenir sa place. Mais ceci ne peut se faire sans prendre en compte les conditions contemporaines de production des données ainsi que l'évolution du contexte juridique qui conditionnent le recueil comme l'usage des données.

I.3. Structure de la production des données en sciences sociales : une situation variable en fonction des disciplines et des différences entre pays

Dans tous les pays, l'histoire de la statistique, de la production et de la diffusion des données publiques est inséparable du rôle joué par les scientifiques dans la connaissance des faits sociaux. Le lien entre théorie sociologique et statistique publique est fondamental.

La structure de la production des données utilisées par les sciences sociales est une donnée historique et culturelle pour chaque pays, reflétant les rapports particuliers de la recherche à l'État et à l'appareil administratif. Or elle conditionne en partie l'utilisation des données. Elle peut aussi, dans une certaine mesure, induire une attention différente des chercheurs (et une plus grande familiarité) aux conditions de production de leurs données, ce qui est un chaînon important du raisonnement scientifique.

Les liens intrinsèques entre la statistique d'État et les sciences sociales sont nombreux. Un fil continu court des premiers dénombrements aux statistiques sociales d'aujourd'hui. Si la visée est bien administrative, elle implique constamment les scientifiques qui prônent la mise en place d'enquêtes et sont aussi utilisateurs des informations ainsi produites. Cette situation, amplifiée aujourd'hui par la révolution informatique, a des racines historiques très anciennes. Sur la question de la mesure, les sciences sociales entretiennent avec l'État un rapport à la fois étroit et critique.

Le dénombrement est d'abord une opération lourde que les États modernes vont imposer pour asseoir leur autorité et leur fonctionnement, reprenant en cela des pratiques très anciennes. Comme l'a fait remarquer Alain Desrosières[3] ces descriptions ont un caractère secret à l'époque du pouvoir royal. Le passage à un instrument destiné à

[3]. L'ensemble de cette analyse s'appuie en grande partie sur les travaux d'Alain Desrosières, en particulier : Desrosières A., (1993), *La politique des grands nombres. Histoire de la raison statistique*. Éd. La Découverte ; Desrosières A. (1998), L'administrateur et le savant. *Courrier des Statistiques*, n° 87-88. Insee.
Il faut également se référer à : Insee (1987), *Pour une histoire de la statistique*, tomes 1 et 2. Éd. Economica ;Savoye A. (1994), *Les débuts de la sociologie empirique*. Éd. Méridiens-Klinsieck.

éclairer de façon concomitante l'État, une société civile qui en est distincte et une opinion publique autonome change sa nature. La naissance des sciences sociales accompagne cette transformation en même temps qu'elle en est le ferment de façon indissoluble. L'idée d'une mathématique sociale fonde à la fois l'objectivité, l'action et la transparence. Le progrès de la connaissance est fortement impliqué par le développement de l'État moderne, qui va tendre, avec des différences qui renvoient à l'histoire particulière de chaque pays, à s'assimiler les investigations menées hors de lui par des érudits, des médecins, des sociétés savantes pour lesquels la publicité des connaissances est une condition essentielle du progrès de la société. Ce processus d'intégration va s'étaler sur plus d'un siècle et aboutir à la constitution des Instituts de statistiques nationaux. La statistique publique apparaît ainsi partout comme un aliment essentiel des sciences sociales.

Le rapport entre sciences sociales et statistiques administratives n'est cependant ni univoque ni celui d'une subordination totale. Ce rapport étroit est aussi objet d'une tension permanente. Le regard critique porte sur l'activité même de la mesure. Mesurer suppose d'abord de savoir ce que l'on mesure. La question des nomenclatures est naturellement au cœur de ce débat, et apparaît de façon particulièrement vive chaque fois que se dessinent, souvent dans la crise, des mutations sociales importantes. Très tôt, sont exprimées des réserves sur le fait de produire des dénombrements hors d'un cadre théorique, revendiqué comme seul à même de fonder des nomenclatures. La contestation d'une statistique descriptive entraînée par la logique bureaucratique est ainsi au fondement d'une critique plus radicale d'une économie fondée sur des nombres, celle d'un Walras à la recherche de fondements théoriques. Le mouvement n'est cependant jamais à sens unique et les crises sont souvent historiquement des moments d'assimilation, par les instituts nationaux, des universitaires à l'origine de la critique. Il en ira ainsi par exemple aux États-Unis au moment de la grande crise des années 30.

D'autre part, la statistique qui travaille ces données est bien une discipline scientifique mais un corps de statisticiens d'État va se développer partout. Le mode de formation et de recrutement de ce corps, plus ou moins autonome selon les pays, peut induire un éloignement progressif (voire une coupure nette comme en Allemagne) entre les institutions chargées de la production des données publiques et le monde de la recherche. L'inévitable éloignement induit continûment une demande de publicité des statistiques ordonnancées par l'État, seule à même de garantir la validité des raisonnements fondés sur elles.

Le lien tend également à se distendre, inégalement selon les disciplines, les domaines de recherche et les pays. L'économie, fortement utilisatrice d'agrégats, entretient ainsi un lien plus étroit avec la statistique publique que la sociologie. Pour cette dernière, la sociologie politique trouve peu

d'aliments, hormis les statistiques électorales, dans la statistique publique peu productrice de données sur les choix politiques et les valeurs.

L'inégal développement de la statistique publique, les modes d'organisation et de financement de la recherche en sciences sociales induisent également des différences entre pays. Un tableau complet et raisonné des différences entre les pays reste à faire. Des instituts statistiques nationaux tels que l'Insee en France ou Statistique Canada au Canada couvrent un champ très large de données sociales. Pour la France c'est en partie le résultat d'une assimilation précoce et constante d'un milieu universitaire restreint mais actif dans le domaine des sciences sociales quantitatives. C'est moins le cas aux États-Unis où l'organisation des Universités est par ailleurs mieux à même de rassembler des fonds pour financer des enquêtes de grande taille.

Enfin la législation relative à la diffusion des données publiques, jusqu'à présent assez différente, a poussé inégalement les chercheurs à assurer directement la production de données. Si la publicité de la connaissance produite sur la société a bien été historiquement un ferment du développement de la statistique publique, l'accès aux données a évolué très différemment. Ainsi la législation très restrictive en Allemagne a fortement contribué à la mise en œuvre d'une politique de production d'enquêtes dans les années 80 par des instituts de recherche et à la création d'une infrastructure d'aide à leur production, le *Zentrum für Umfragen, Methoden und Analysen* (Zuma). Encore faut-il que les chercheurs mettent en œuvre cette production et que l'organisation de la recherche puisse prendre en compte des financements lourds.

L'exemple des grandes enquêtes européennes et internationales illustre bien ce processus. L'impossibilité de construire des comparaisons internationales fondées sur des données de grande taille à partir des seules données nationales peu disponibles et peu harmonisées a incité les chercheurs à construire des enquêtes internationales. L'absence remarquable de la France dans ce processus relève en partie de l'absence de source de financement institutionnelle pour la production de données en sciences sociales, hormis un cadre très spécialisé par domaine de recherche transitant par quelques instituts de recherche (Ined ou Inra par exemple). D'une manière plus générale, on peut caractériser la situation française par une modeste production de grandes enquêtes par les chercheurs. Même dans un domaine où la statistique publique est peu présente comme la sociologie politique, l'absence de tradition de financement des données s'est traduite par des ruptures dans les séries d'enquêtes pré-et post-électorales. On reviendra plus loin en détail sur cette situation qui induit un éloignement plus fort des chercheurs des conditions de production des enquêtes, même si elle est le reflet paradoxal d'une assimilation précoce par la statistique publique française des recherches produites en dehors d'elle.

On trouvera en annexe des descriptions précises retraçant la situation des chercheurs dans la production et l'utilisation des données pour l'Allemagne, la Grande-Bretagne, le Canada et les États-Unis. À l'évidence, la part occupée par les données produites au sein du monde de la recherche est différente. Elle relève cependant d'une tradition ancienne (Quetelet, Simiand, les hygiénistes anglais, Halbwachs, par exemple) et persistante, complétant et nourrissant la statistique publique, partout prédominante.

Y a-t-il lieu aujourd'hui de continuer à opérer une distinction entre données issues de la statistique publique et données académiques, pour adopter la dénomination en vigueur dans le champ anglo-saxon, et que l'on maintiendra dans la suite de ce rapport ? La statistique publique utilise fortement (mais inégalement selon les pays) les résultats des recherches menées en sciences sociales. La finalité de recherche est incontestablement présente dans nombre d'enquêtes issues des instituts nationaux de statistique ou d'agences gouvernementales. C'est le cas notamment de l'Insee en France et de Statistique Canada. Inversement des enquêtes issues du monde universitaire, qui sont largement financées partout sur fonds publics, produisent également de l'information statistique d'intérêt public. Les instituts de sondages ne sont pas non plus complètement absents de cette production, même si la logique en est très différente. C'est également le cas des données administratives, de celles générées par l'activité économique ou par des praticiens, comme les médecins. Il y a donc sens, d'une certaine manière, comme le propose Gert G. Wagner[4] dans un article récent, à considérer que, quel que soit leur mode de production, les données constituent un champ unique où la visée opérationnelle liée à l'intérêt public et à la gestion est de plus en plus difficile à séparer d'une visée de connaissance plus strictement définie. Cela n'est guère étonnant compte tenu du lien historique initial entre sciences sociales et statistique publique. La seule question serait donc d'assurer la cohérence d'ensemble du système. Une instance comme le Cnis en France correspond bien à cette définition. Il faut remarquer que son périmètre d'action est défini de façon très large. Les enquêtes menées par les EPST tels que l'Ined ou le CEE y reçoivent ainsi avis et labels. La notion de statistique publique en France est sans aucun doute fondée sur la notion d'intérêt public, pouvant inclure des enquêtes produites pas des établissements de recherche.

La distinction entre données publiques et données académiques, nous paraît cependant légitime pour deux raisons au moins. La simple comparaison de la France avec d'autres pays montre à l'évidence que l'existence ou non de certains instruments (on pense en particulier aux panels de long terme, à la participation à des enquêtes internationales et aux enquêtes sur les valeurs et les opinions) sont dépendants des mécanismes de financements et d'organisations distincts de la statis-

[4]. Wagner G. (1998), *An Economists Viewpoint of Prospects and Some Theoretical Considerations for a Better Cooperation of Academic and Officials Statistics*. OCDE.

tique publique. Dans certaines conjonctures le recueil de données, sur des sujets jugés sensibles, peut aussi apparaître meilleur s'il se fait dans le cadre d'une institution de recherche. Nous retiendrons en ce sens comme critère le cadre institutionnel de production des données.

La seconde raison qui plaide pour cette distinction tient à ce que l'utilisation des données par un tiers rencontre des conditions légitimement différentes d'accès aux données selon leur origine institutionnelle et la nature de leurs financements. Cette question est celle du droit d'usage des données. Que les chercheurs aient accès à des données issues de la statistique publique pose essentiellement la question du cadre juridique de l'accès aux données publiques et administratives. Cette question se décline dans le monde contemporain sous deux formes : statut des données publiques (droit à l'information, coût d'un bien public) et protection des données personnelles. L'accès des chercheurs aux données produites par d'autres chercheurs pose plus directement la question de la propriété intellectuelle, du droit d'exploitation prioritaire du chercheur qui a constitué la base de données, et trouve des modulations selon l'origine publique ou non de leurs financements. Enfin l'accès à des données produites par des opérateurs privés (instituts de sondages) met plus clairement l'accent sur la question de propriété. Les grandes lignes juridiques sont exposées plus loin, mais on veut souligner ici que la tonalité plus ou moins restrictive des législations a un impact qui varie en fonction de la structure de la production des données dans un pays. À cet égard, on peut considérer qu'en France une restriction sur l'accès aux données de la statistique publique, quelle que soit sa nature et sa forme, a d'autant plus d'impact que celle-ci constitue la source quasi exclusive de micro-données issues de fichiers de grande taille dans certains domaines de recherche.

I.4. Les fondements du partage des données

Partager les données existantes est aussi une manière de ne pas solliciter inutilement les personnes et les entreprises. C'est également l'enjeu du contrôle du caractère

Le débat sur le partage des données naît après guerre. Le terme même de partage des données[5] qu'on utilisera dans ce rapport de préférence à celui d'accès aux données mérite d'être commenté. Ce terme est peu utilisé en France et la question n'a pas donné lieu à des débats alors que nombreuses sont les publications sur ce sujet impliquant au premier chef des chercheurs américains puis très rapidement des chercheurs de plusieurs pays européens (Grande-Bretagne, Allemagne, Pays-Bas, Suède, Norvège) dès les années 50. Le terme partage des données (*Sharing data* est le titre de plusieurs publications) présente l'avantage

[5]. Ce terme de partage des données n'implique pas une circulation sans contrôle des données, en particulier lorsqu'il s'agit de données à caractère personnel. Cette circulation doit obéir à des règles.

scientifique des travaux de recherche que seule la réplication permet d'exercer.

sur celui d'accès aux données de mettre l'accent sur les deux acteurs du partage des données, non seulement le chercheur désireux d'obtenir des données, mais aussi celui (qui peut être parfois chercheur aussi) qui produit les données et est amené à partager. L'un des grands points dans cette question, c'est qu'il faut prendre en compte à égalité le point de vue du producteur initial et celui du chercheur utilisateur secondaire.

Ce n'est donc pas par hasard que la forme initiale du débat sur cette question a pris l'allure d'une discussion formulée en termes de coûts et avantages du partage des données. Une formulation plus récente, reprenant l'un des items formulés dans le cadre de ce premier débat, celui de la validation et de la réplication scientifique, fait de cet argument particulier l'élément central d'une véritable bataille scientifique.

a) Le débat coûts et avantages

Les producteurs de données ont avantage, pour être financés, à faire état d'une large utilisation de leurs matériaux statistiques. Ceci suppose citation et reconnaissance du travail que requiert la mise à disposition des fichiers.

Historiquement, comme on le verra plus loin, l'organisation des premières archives de grandes enquêtes pour les sciences sociales paraît assez fortement liée dans les années 50 à la volonté de plusieurs chercheurs relayés par quelques grandes organisations internationales comme l'Unesco, de travailler dans le cadre de l'après-guerre à des comparaisons internationales. C'est le cadre du premier véritable débat sur la question du partage des données dans la communauté scientifique. La création du *Roper Institute* aux États-Unis à partir des archives de données d'un institut privé, le Gallup, données qui intéressent surtout les sciences politiques, reflète assez bien la première forme du débat, celle des coûts et avantages. Le débat trouve une forme achevée à partir des années 70 alors que sont déjà en place les premiers *Data Archives*, notamment l'ICPSR de l'Université de Michigan aux Etats-Unis et l'ESRC-*Data Archive* de l'Université d'Essex en Grande-Bretagne[6], et que le champ couvert par ces institutions s'est élargi aux grandes enquêtes, publiques et universitaires, intéressant l'ensemble des sciences sociales.

Ce débat est très fortement centré sur l'un des deux acteurs, le producteur, quel qu'il soit, public ou privé, chercheur ou non. Il s'agit de le persuader de partager les données tout en prenant en compte ses intérêts. On verra qu'un autre débat plus récent insiste plus sur le chercheur comme utilisateur de données, qu'il s'agisse des siennes ou non.

Dans l'idée de persuader les producteurs de données, quels qu'ils soient, le débat met d'abord en avant l'intérêt pour eux d'un partage, en regard

[6]. Rokkan S. (ed.), 1966, *Data Archives for the Social Sciences*, Monton, Paris The Hague.
Fienberg S., Martin M., Straf M. (ed.), 1985, *Sharing Research Data*, National Academy Press, Washington D.C.
Sieber J. (ed.), 1991, *Sharing Social Sciences Data. Advantages and Challenges*, Sage Focus Edition.

des inconvénients liés à la prise en compte des intérêts des demandeurs. Certains font ainsi remarquer (Cecil et Griffin[7]) que la grande difficulté du partage des données tient à ce que les charges reposent presque exclusivement sur les producteurs. Les différents auteurs mettent donc en avant au bénéfice des producteurs de données essentiellement trois éléments :
1) la reconnaissance supplémentaire des producteurs via la citation (ce qui suppose bien évidemment que les producteurs de données soient cités dans tout traitement effectué sur la base de ces données),
2) l'accroissement de l'utilisation d'enquêtes coûteuses pour le producteur apportant donc une justification de ces coûts (éventuellement en partie répercutés) d'une part et un retour vers le producteur des éléments de connaissance supplémentaires produits d'autre part,
3) enfin un progrès méthodologique dérivé de ces traitements secondaires utilisable par le producteur pour des enquêtes ultérieures.

Les intérêts reconnus aux demandeurs sont en fait ceux des intérêts généraux de la science : réduction globale des coûts, surtout dans la mesure où le financement des enquêtes est d'origine publique, diminution de la charge pour les répondants aux enquêtes (la mise en commun des données évite la démultiplication des collectes), accroissement substantiel des traitements et donc gains en termes de cumulativité, possibilité de réplication des traitements, seuls garants d'un processus de validation scientifique, enfin possibilité de formation des étudiants à l'analyse des données. En contrepartie toute mise en place de partage des données doit, pour avoir des chances d'être effective, prendre en compte des coûts et des contraintes mis en avant par les producteurs. Les coûts dérivent essentiellement du travail supplémentaire que les producteurs doivent fournir pour rendre les données utilisables par des tiers. Il s'agit de tout ce qui concerne l'archivage et sa maintenance, la documentation indispensable pour comprendre les données, les formats de mise à disposition qui peuvent constituer des obstacles techniques. Le travail supplémentaire important que ce processus de mise à disposition implique pose un problème dans la mesure où il n'est pas nécessaire au moment même du traitement primaire des données par le producteur, car il connaît ses données.

Obtenir ce travail suppose donc soit des instruments de contrainte (obligation de dépôt légal par exemple), soit une prise en compte des coûts éventuellement partagés, soit enfin une négociation sur les avantages obtenus en contrepartie par le producteur. Les contraintes à prendre en compte renvoient aux exigences d'ordre juridique et déontologique que le producteur peut faire valoir. Sur le plan juridique, le producteur est fondé à faire valoir que les données relèvent de la propriété intellectuelle (création originale) et impliquent des coûts de production. La question du droit prioritaire d'exploitation du chercheur

[7]. Cecil J.S., Griffin E. (1985), *The Role of Legal Policies in Data Sharing*, in Fienberg et alii op. cité.

et de ses limites dans le temps est également posée. Sur le plan déontologique, le producteur peut vouloir s'assurer que les utilisateurs ont la compétence nécessaire pour utiliser les données correctement. Dans le cas de données dites personnelles, le producteur doit s'assurer que les contraintes de respect de la vie privée seront prises en compte dans les traitements secondaires.

L'ensemble de ce débat, parfaitement explicité dès les années 60 et réitéré régulièrement depuis, est toujours d'actualité, sous une forme presque inchangée.

b) Le débat validation/réplication

La validation des travaux de recherche et la cumulativité des résultats sont conditionnées par la disponibilité des données sur lesquelles ils sont fondés. Celles-ci deviennent un enjeu incontournable de la crédibilité des sciences sociales.

La question du partage des données comme condition même de la validation scientifique figure bien dans le débat des années 60 et 70 mais ne constitue pas un élément majeur du débat. Le lien entre partage des données et validation scientifique apparaît en revanche plus récemment et de manière très forte dans les sciences politiques américaines. Un important dossier a été récemment consacré à cette question par PS (*Political Science*), revue professionnelle de l'*American Political Science Association (1996)*. Dans ce dossier, véritable table ronde de discussion des questions de vérification, réplication et partage des données, le politiste Gary King défend l'idée que le partage des données, l'accès à toutes les informations sur celles-ci, aux conditions de traitement des données, constituent des conditions sine qua non de vérification par d'autres des résultats obtenus par la recherche en sciences sociales. King situe très clairement la question du partage des données dans un contexte épistémologique de type popérien[8] : la communauté scientifique doit fonctionner selon des normes d'accès aux données, de mise à disposition des programmes de traitement des données et finalement d'espace de discussion critique des résultats des travaux dans une logique de « falsification ». Cette position, et le débat qu'elle suscite entre King et Herrnson aux États-Unis, est particulièrement importante pour les questions traitées dans ce rapport. En effet, derrière les questions d'archivage et de partage des données, apparaît une série d'enjeux scientifiques forts qui touchent au fonctionnement même de la communauté scientifique en sciences sociales : il s'agit bien en fait d'aider les sciences sociales à entrer plus fortement dans la sphère du débat scientifique maîtrisé, fondé sur l'accumulation de données et de résultats, organisant les conditions de la vérification empirique et assurant celles de la validation des résultats.

Les conséquences de cette position, qui lie très étroitement partage des données et vérification/validation des résultats, peuvent être fortes en termes d'organisation et de fonctionnement de la communauté scienti-

[8]. On peut penser que s'ajoute à cela le souci des sciences politiques sur des sujets très sensibles de se prémunir contre le préjugé, l'erreur ou la fraude.

fique. Par exemple, suite à ces débats, quelques revues scientifiques américaines (telle que *Social Science Quaterly* par exemple) ont mis en place des procédures nouvelles de fonctionnement de leur comité de rédaction : on demande aux auteurs d'accompagner leurs textes de leurs données et des programmes de traitement de celles-ci. Si ces débats récents et pratiques nouvelles se sont, pour le moment, essentiellement développés en sciences politiques, il faut néanmoins y voir l'amorce de transformations plus profondes et plus larges du mode de validation scientifique en sciences sociales. Ces débats s'inscrivent en effet dans une controverse plus générale touchant l'ensemble des sciences sur la question de l'erreur voire de la fraude scientifique (voir Le Monde du 26 mars 1999 ou « l'affaire Sokal »), sur fond de concurrence accrue et d'enjeux financiers lourds dans certains domaines.

Les arguments de King plaident en faveur d'une mise à disposition systématique des données utilisées. Ceci suppose cependant de prendre en compte les questions de propriété intellectuelle sur les données et celles relatives à la protection de la vie privée en cas de données à caractère personnel. Herrnson fait valoir que ces positions sont relativement artificielles dans la mesure où les réplications à des fins de validation scientifique sont relativement rares. Il soutient d'autre part que la créativité en sciences sociales s'est historiquement plutôt appuyée sur des enquêtes originales. Les travaux depuis plus de vingt ans autour de la mobilité sociale en France à partir des enquêtes Formation Qualification Professionnelle de l'Insee, le débat britannique autour des effets démocratisants de l'*Education Act*, ou encore les analyses réitérées sur les données de Coleman montrent cependant l'intérêt qu'il peut y avoir à travailler sur les mêmes données.

c) Une solution équilibrée

Le débat actuel a donc plusieurs dimensions. Il paraît impensable qu'une argumentation puisse être validée scientifiquement si les données sont par principe inaccessibles. Ceci suppose une régulation de l'accès aux données, qui prenne en compte les problèmes juridiques, les intérêts respectifs des différents protagonistes producteurs et utilisateurs des données, enfin la spécificité du travail scientifique. D'autre part, la question des coûts de production, qu'il s'agisse de données publiques ou de données académiques lorsqu'il s'agit de financements publics, va prendre une place plus importante et plaide pour une utilisation optimum des données. Inversement, une meilleure régulation du partage des données ne doit pas se traduire par un tarissement des financements pour des enquêtes originales.

Ces débats, qu'il s'agisse de ceux sur les coûts et avantages du partage des données ou de ceux sur la validation scientifique sont totalement absents en France et n'ont donné lieu à aucun ouvrage ni commentaire.

I.5. Les implications juridiques du partage des données

En application de la Directive européenne de 1995 qui prend en compte les besoins spécifiques de la recherche, la révision de la loi Informatique et Liberté de 1978 devrait définir un contexte plus favorable.

Les questions juridiques interviennent, on l'a vu, constamment dans le débat sur le partage des données. Si ces questions sont présentes dès les années 50 lorsqu'apparaissent les premiers éléments du débat sur le partage des données, elles ont pris progressivement depuis une quinzaine d'années, avec la révolution informatique, une toute autre dimension dont il importe de prendre d'emblée la mesure. La multiplication des bases de données personnelles qui accompagnent la vie administrative et économique, puis la possibilité accrue de faire circuler les informations détenues dans ces bases et, dans de nombreux cas, la nécessité administrative comme économique de les faire circuler, en particulier dans les ensembles intégrés comme l'espace européen, ont donné un relief nouveau aux inquiétudes en matière de protection de la personne et de la vie privée, comme aux préoccupations relatives à la définition des droits d'auteur sur ces bases, à leur valeur marchande éventuelle. Une législation visant à encadrer la collecte, l'archivage, l'usage et la transmission des données à caractère personnel a vu le jour partout à partir des années 70. C'est le cas en France avec la loi Informatique et Libertés de 1978. La nécessité de faire circuler des données en Europe pour des raisons tant administratives qu'économiques a conduit à la mise en place d'une Directive européenne sur ces questions en 1995, qui est en vigueur depuis 1998 dans les différents pays de l'Union. On dispose de très nombreux rapports sur ce point, dont le rapport Braibant[9] par exemple pour la France, préparant une modification de la loi de 1978. Il n'était donc pas question dans le cadre de cette mission d'entrer à nouveau dans le détail de ces questions. Il est par contre nécessaire d'en faire apparaître les retombées pour la recherche en sciences sociales et de souligner combien les chercheurs de ces disciplines ont été à la fois peu présents et peu représentés dans ces débats. Or, si des questions d'organisation du partage des données se posent aujourd'hui avec une telle acuité, c'est en grande partie du fait des contraintes imposées par le droit.

Les chercheurs ont été très peu nombreux lors du vote de la loi de 1978 en France à prendre la mesure des retombées possibles des restrictions fortes introduites sur les pratiques des chercheurs en sciences sociales. La loi de 1978 a un caractère très général et ne fait ainsi aucun sort particulier à la recherche, alors que « les traitements automatisés d'informations nominatives opérés pour le compte de l'État, d'un établissement public ou d'une collectivité territoriale, ou d'une personne morale de droit privé gérant un service public » y trouvent place (article 15). La possibilité de ménager des régimes spécifiques apparaît dans la convention du Conseil de l'Europe de 1981. Toutefois, bien que la France l'ait ratifiée, la loi ne prévoyant que des aménagements facultatifs, la loi française n'a pas été modifiée. C'est la recherche

[9]. *Données personnelles et société de l'information*, La documentation française, 1998.

médicale, et l'épidémiologie en particulier, qui se sont mobilisées le plus vite pour faire reconnaître, dans un chapitre additionnel et dérogatoire de 1994, les finalités particulières de la recherche et organiser la collecte et l'utilisation des données personnelles dans ce cadre. Les intérêts de santé publique impliqués par ces disciplines ont été une aide puissante pour faire reconnaître la recherche comme finalité à prendre en compte. Les sciences sociales n'ont mesuré que progressivement l'impact des restrictions introduites sur leurs pratiques. Les limitations fortes à l'usage des données du recensement de 1999 qui ont été introduites par la Cnil ont été un catalyseur d'une prise de conscience plus collective, en particulier pour les géographes. La Directive européenne de 1995 a introduit expressément les finalités de recherche, de statistique et d'histoire et permet donc de faire entrer ces dispositions dans le droit positif. On peut espérer que les modifications de la loi de 1978 qui seront introduites en application de cette directive, reprendront ces dispositions dans leur totalité, créant ainsi un environnement plus favorable pour la recherche.

Il reste que le statut de la recherche occupe une place extrêmement faible, voire quasi inexistante dans des débats complètement centrés sur les questions administratives et économiques, qu'il s'agisse du droit d'auteur en matière de bases de données ou de la protection de la vie privée lorsque des données à caractère personnel sont impliquées. Il est donc utile de parcourir rapidement ces questions du point de vue de la recherche en sciences sociales

a) Bases de données, droit d'auteur, données personnelles et recherches en sciences sociales

Les droits des producteurs sur leurs données ne sont pas clairement définis. Les droits d'usage des utilisateurs secondaires non plus.

Le recueil des données est au fondement, on l'a vu, de la constitution même des sciences sociales, qu'il s'agisse d'enquêtes directes ou d'utilisation de données déjà rassemblées par ailleurs. Pour la monographie comme pour les échantillons de grande taille destinés à une exploitation statistique, il implique nécessairement de recueillir des données à caractère personnel, qu'il s'agisse de personnes privées, physiques ou morales (entreprises ou autres). Les sciences sociales sont donc concernées au premier chef par toutes les questions touchant à la nature du droit d'auteur sur de telles bases et par celles relatives aux données à caractère personnel.

La question du droit sur les données demande à être élucidée[10]. Elle vise à identifier le contenu de ce droit, ce qu'on peut ou non faire, et son titulaire, qui est investi ou non de pouvoir le faire.

Quant au contenu, on distinguerait ce qui concerne la donnée elle-même (son intégrité ou exactitude), la possibilité de l'utiliser et enfin la faculté

[10]. On s'inspire ici de la note de R. Padieu reproduite en annexe.

d'échanger détention ou usage contre des stipulations telles qu'un paiement, une limitation de l'usage ou l'obligation de rendre compte de l'usage fait.

Quant au titulaire du droit, il peut être divers : personne concernée par une donnée individuelle, détenteur actuel des données, tiers pouvant prétendre à cette détention ou à l'usage. Chacun des titulaires possibles peut n'avoir qu'une partie des droits énoncés ci-dessus. Et la transmission des données ne transfère pas de facto tous les droits du détenteur précédent : des dispositions légales ou contractuelles explicitent ce que le bénéficiaire est autorisé à faire. On a ainsi plusieurs personnes qui ont simultanément des droits différents sur la même donnée. En matière de données personnelles, on reconnaît à la personne concernée un droit premier général sur ses propres données : droit de veiller à leur intégrité, d'en concéder l'usage, d'exiger des contreparties. Est-ce un droit souverain et définitif ? Hormis le cas de cession ou autorisation volontaire, une disposition d'ordre public impose souvent à la personne concernée de communiquer ses données (déclarations administratives, procédures judiciaires, enquêtes statistiques obligatoires). Reste en débat de savoir jusqu'à quel point cette reconnaissance du droit d'autrui prive la personne concernée d'une partie de ses droits originels. (Par exemple, doit-elle être informée d'une utilisation ultérieure à des fins scientifiques, lorsque celle-ci ne peut lui nuire ?)

Les grandes collectes de données privées ou publiques sont un gisement souvent de grand intérêt pour la recherche. La directive de 1995 autorise qu'elles soient mobilisées pour celle-ci, moyennant des limitations, précautions et garanties convenables. C'est ce partage à finalité de recherche qui doit maintenant être organisé par la loi et par d'autres dispositions.

La question d'un droit d'auteur se pose à propos de la constitution d'ensembles de données relatives à un plus ou moins grand nombre de personnes. Cette base de données est-elle « une œuvre de l'esprit » ? La réponse donnée par les juristes est habituellement positive (voire en annexe les notes du Cecoji). La conception de l'architecture de base, comme le travail impliqué par le recueil des données, ainsi que parfois des traitements contrôlant les données ou en créant une information originale par combinaison de plusieurs données, invitent à accorder à l'auteur une protection ou un privilège à l'égard de l'utilisation par autrui. Il faut toutefois noter que cette protection ne saurait concerner que ce qui est la création de l'auteur de la base et non les données sous-jacentes en elles-mêmes. En effet, si cet auteur détient effectivement les données en cause, il ne s'est en général pas vu transférer tous les droits premiers des personnes concernées.

Que la base de données soit ou non le support d'un droit d'auteur, son responsable dispose de certains droits : quant à l'exactitude des données[11], l'accès d'utilisateurs tiers et les contreparties qui leurs sont demandées. Or, pour l'exercice de ces droits, le responsable de la base peut se voir aussi soumis à des obligations : tant envers les personnes concernées par les données de base (assurer leur protection) qu'à l'égard des tiers (devoir ouvrir l'accès à des chercheurs, mais aussi devoir exiger des conditions à cet accès). Autrement dit, le détenteur d'un fichier, voire l'auteur d'une base de données disposant d'un droit d'auteur, n'en est pas pour autant propriétaire.

Les considérations qui précèdent visent à savoir qui a certains droits sur les données et ce qu'il peut ou doit en faire. Dans cette perspective, se pose aussi la question du coût d'accès aux données pour les chercheurs. Laissons de côté le cas où le chercheur a collecté par lui-même les données ou a construit une base de données, ainsi que le cas où il est associé à l'organisme qui l'a fait : il a par avance supporté tout ou partie du coût. Sinon, si l'ensemble de données convoité est déjà constitué, si l'on admet légitime l'utilisation par le chercheur et supposant réglées les conditions pour cela, une part des coûts de constitution doit-elle être répercutée sur lui ? La « doctrine » développée par l'Insee au cours des vingt-cinq dernières années considère que le recueil et la production des bases de données sont déjà payées par l'État et confère à celles-ci un caractère de bien public : elles n'ont pas à être à nouveau payées par le bénéficiaire de l'accès. En revanche, l'opération de livrer cet accès peut engendrer des coûts, dits « de mise à disposition » : ceux-ci sont à faire supporter par le demandeur, c'est-à-dire le chercheur pour ce qui nous occupe ici. Lorsque le bénéficiaire entend commercialiser les données ou les intégrer à un produit commercialisé, il est admis d'ajouter au strict coût de mise à disposition une redevance (par exemple, en appliquant au coût de mise à disposition un coefficient : 2 ou 3 ou toute autre valeur). La recherche, étant désintéressée et elle-même d'intérêt public, n'est pas soumise à cette redevance : seulement au coût simple de mise à disposition. D'autres pratiques ont pu être développées ; toutefois, en 1994, une « circulaire Balladur » a unifié les règles sensiblement comme il vient d'être indiqué.

Au total, il y a lieu de s'assurer que les budgets des institutions de recherche (qu'elles travaillent sur subvention ou sur contrats) permettent le règlement des coûts de mise à disposition et, éventuellement, de constitution primaire de certains recueils. Qu'ils soient à la charge d'un

[11]. Dans certains traitements statistiques ou à finalité scientifique, l'exactitude des données n'est pas forcément requise comme c'est le cas lorsque les données peuvent fonder des jugements ou décisions qui concernent une personne déterminée. Le chercheur peut ainsi être amené à des « redressements » statistiques, qui améliorent la représentativité d'ensemble de la base de données bien que certaines données particulières soient délibérément inexactes. Dans un but de protection de la confidentialité, il peut aussi être introduit des modifications aléatoires qui conservent les propriétés d'ensemble de l'information mais interdisent toute conclusion particulière à une personne déterminée.

institut de statistique ou des chercheurs, ces recueils sont à considérer, nous l'évoquons par ailleurs, comme des investissements à l'instar des grands instruments de la physique ou de la biologie (accélérateurs, télescopes, souches, etc.).

b) Une place insuffisante de la recherche dans le débat juridique

Les associations de statisticiens ont jeté les bases d'une déontologie. Aux États-Unis le Freedom of Information Act instaure des règles de transparence.

La reconnaissance dans la Directive européenne de 1995 d'une finalité de recherche, de statistique et d'histoire ouvre la voie à une évolution dans ce sens en France. La Société française de statistique par sa commission de déontologie a entamé la mobilisation sur ce plan, en concertation avec les associations d'épidémiologistes et des chercheurs du CNRS. On peut espérer (cf. le rapport Braibant), sans en être certain pour l'instant (cf. l'état actuel de l'avant-projet), que la refonte de la loi de 1978 en France suive la directive sur ce point et prenne en compte la recherche dans le droit positif.

Le cadre européen demeure malgré tout différent de celui qu'établit le *Freedom of Information Act* aux États-Unis, par référence à la Constitution américaine. Des deux idées induites par le *Freedom of Information Act*, seule celle sur la qualité de bien public des données produites par l'État et ses administrations et financées par l'impôt se retrouve dans les débats et les législations européennes et française.

Ce débat se réfère d'abord aux acteurs économiques qui doivent pouvoir accéder (gratuitement) à des informations utiles à leur activité, que l'État n'est pas habilité à exploiter commercialement. C'est une des raisons pour lesquelles les aménageurs locaux qui avaient besoin de disposer de données à des niveaux fins inférieurs aux seuils de protection définis par la Cnil, ont obtenu des droits d'accès qui n'ont pas été reconnus d'emblée aux chercheurs[12].

Ces questions, qui ont deux faces, l'accès à des données utiles économiquement et en même temps leur coût, ont été récemment évoquées dans le cadre du Cnis (Rapport sur la diffusion des données publiques) et du rapport Mandelkern. En France c'est, on l'a vu, une circulaire de 1994, dite « circulaire Balladur », qui fixe certaines règles en matière de coût de mise à disposition des données publiques. Cette circulaire considère bien que les données publiques sont gratuites dans leur principe, mais qu'il convient de prendre en compte un coût de mise à disposition pour des utilisations particulières. La discussion actuelle porte sur la distinction entre un domaine large d'intérêt public où la mise à disposition est gratuite (chaque service administratif définit des informations qu'il convient de faire rentrer dans ce cadre) et un domaine

[12]. Cf. Françoise Moreau (1999), *Distribution des bases de données démographiques locales. Comparaison france-États-Unis*, Ined.

lié à une utilisation spécifique demandant une élaboration supplémentaire. Ce sont surtout les acteurs économiques qui sont actifs dans ce débat. On peut se demander cependant si la mise à disposition pour la recherche relève du périmètre général de l'intérêt public ou de l'utilisation particulière[13].

La prise en compte de la recherche est également faible dans le débat important suscité par la circulation accrue des données du fait de la croissance des réseaux informatiques et de la mondialisation de la vie économique. L'élaboration de la Directive européenne de 1995 est directement issue de la nécessité d'harmoniser les législations pour permettre la circulation des données pour les opérateurs économiques en particulier. La législation américaine est considérée désormais comme moins protectrice et ceci constitue un frein à la circulation des données à caractère personnel en direction des États-Unis. Dans ces débats, des questions très présentes pour les chercheurs en sciences sociales comme la constitution de bases intégrées de micro-données à partir de bases nationales issues notamment de la statistique publique à des fins de comparaison européenne et au-delà internationale sont complètement absentes.

L'autre idée présente dans le *Freedom of Information Act*, le droit à l'information comme fondateur de la démocratie, est beaucoup moins présente en Europe à la différence des États-Unis. Le droit à la protection de la vie privée est reconnu aux États-Unis avec le *Privacy Act* au cours de la même période qui voit les pays européens mettre en place ce type de législation. Mais le *Freedom of Information Act* définit pour les données fédérales aux États-Unis un droit à la transparence qui est proche de l'idée fondatrice de la statistique publique qui émerge dans le courant des Lumières. La mise à disposition des données de la statistique publique pour les chercheurs s'est trouvée grandement facilitée par ce contexte[14]. Les données y trouvent un statut fondateur de bien public, qui ne dérive pas seulement de leur financement au moyen de l'impôt.

Réguler le partage des données, quelle qu'en soit l'origine, pour permettre aux sciences sociales d'utiliser le potentiel dégagé par la révolution informatique dans le sens d'une plus grande cumulativité et jouer tout leur rôle d'expertise, ne pourra se faire dans le contexte contemporain d'inquiétude légitime sur la protection des droits et libertés fondamentaux des individus, sans une implication plus forte et

[13]. Les modifications apportées quant au coût de mise à disposition ne sont pas cependant sans incidence sur le budget des instituts nationaux de statistique. À titre d'exemple, les rentrées brutes incluant recettes de diffusion et de partenariat (co-financement d'enquêtes) représentent 7 % du budget de l'Insee en France, et 30 % de celui de l'institut danois. Elles permettent de financer certains programmes.

[14]. Le débat actuel aux États-Unis porte maintenant sur l'opportunité de faire relever les données de recherches financées sur des fonds publics du régime du FOIA.

organisée de l'ensemble des acteurs de la recherche (enseignants-chercheurs, organisations professionnelles, instituts de recherche, CNRS, Universités, Direction de la recherche) dans le débat juridique, où ils n'ont occupé au mieux jusqu'à présent qu'une place très marginale. On peut remarquer que c'est précisément cette fonction qu'ont assumée les institutions d'archivage et de diffusion des données pour les sciences sociales qui se sont construites aux États-Unis et en Europe à partir des années 60 et dont il n'existe que des jalons pour la France. Cette implication s'est traduite, à l'instar de ce qui a déjà été fait par quelques disciplines (statistique et épidémiologie), par une professionnalisation du milieu et en particulier par l'élaboration de codes professionnels de bonnes pratiques, recherchant un équilibre entre les différents intérêts en même temps que des garanties déontologiques. L'élaboration de tels codes professionnels est précisément encouragée par la Directive de 1995.

II. L'institutionnalisation du partage des données

Face à ces questions centrales pour les sciences sociales, des infrastructures nationales puis internationales se sont constituées à partir des années 50 permettant d'apporter de l'aide aux chercheurs pour accéder aux données, les utiliser, aider éventuellement à en produire, mettre en place des procédures déontologiques prenant en compte les questions juridiques que posent la production et l'utilisation des données à des fins de recherche. La France s'est occupée tardivement et de manière peu structurée de la constitution d'infrastructures dédiées au développement de la recherche empirique en sciences sociales. Les banques de données, l'un des points essentiels de telles infrastructures, se sont mises en place en France avec quinze à vingt ans de retard sur d'autres pays occidentaux. La France est également peu présente dans les grandes enquêtes internationales, faute de politique en matière de financement recherche pour la production de données.

II.1. Les Data Archives et les grandes enquêtes

Les Data Archives se sont développés dès les années 1950. L'objectif initial était de favoriser la comparaison internationale en sciences sociales en mettant à disposition des chercheurs le patrimoine d'enquêtes accumulé.

C'est en effet à partir des années 50 que se sont construits aux États-Unis puis en Europe, à l'initiative de quelques chercheurs (souvent issus de la science politique), des *Data Archives* (selon le terme de Stein Rokkan) destinés à sauvegarder des données d'enquêtes importantes et à les mettre à disposition pour les autres chercheurs. Ces initiatives, appuyées au départ sur des structures de recherche, ont été assez vite relayées dans plusieurs pays par une politique de la recherche qui a mis en place des institutions disposant d'une légitimité et de moyens.

a) Les origines des Data Archives

Les sciences politiques, probablement en raison de leur plus faible lien avec les données publiques, ont eu et continuent d'avoir un rôle pionnier tant sur le plan des infrastructures que sur celui des débats. Deux initiatives sont clairement à l'origine des banques de grandes enquêtes utilisées par les sciences sociales, connues sous le nom de *Data Archives*. Il s'agit d'une part de la création du *Roper Center* aux États-Unis, d'autre part des initiatives prises dans les années 50 par quelques chercheurs s'intéressant aux comparaisons internationales.

La création du *Roper Public Opinion Centre*, dont les prémisses remontent à 1945, bénéficie de la tradition américaine de legs privés aux universités. Elmo Roper, spécialiste de l'enquête par sondage, dépose dans une bibliothèque universitaire dix ans de données d'enquêtes (à l'époque il s'agit de boîtes de cartes IBM), dans l'idée que ces données sont sous-exploitées et peuvent servir plus tard de point de comparaison pour suivre l'évolution des opinions. Il encourage ensuite des collègues et notamment Georges Gallup à suivre son exemple. En 1957, ce fonds prend la forme d'un département distinct, devenant dans les faits la première banque de grandes enquêtes pour les sciences sociales.

La véritable dynamique de l'archivage des données de science sociales renvoie à quelques grandes figures intellectuelles de l'après-guerre, spécialistes de sciences politiques notamment de sociologie politique, acquis à un vaste programme intellectuel : développer une grande banque de données mondiale, véritable infrastructure pour la recherche comparative en sciences sociales. Le politiste norvégien Stein Rokkan est assurément la figure centrale de ce groupe : ouvert à l'international par de multiples séjours dans les universités américaines et européennes, son ambition intellectuelle est de permettre, par la constitution de corpus de données et leur archivage, l'analyse historique de la genèse des systèmes politiques et partisans occidentaux.

La politique de l'Unesco, désireuse d'appuyer les programmes de coopération scientifique internationale au sortir de la guerre, sera un appui fort pour cette idée d'instituts d'archivage. Il suffit pour s'en rendre compte de consulter la Revue Internationale des Sciences Sociales, éditée par l'Unesco, sur la période des années cinquante et soixante. Les références à ces questions y sont nombreuses et l'on voit bien que, dans le contexte de l'après-guerre, l'Unesco appuie par l'organisation de séminaires, tables rondes, colloques, le développement d'une coopération scientifique « interculturelle », fondée sur l'analyse de données empiriques. Il s'agit de doter la communauté scientifique en sciences sociales d'infrastructures matérielles pour la réalisation de ce projet : les *Data Archives*, archives de données, dont la dénomination montre que les données empiriques sont alors conçues comme des éléments du patrimoine scientifique dont il faut assurer la conservation sur le long terme pour une réutilisation à des fins de recherche.

b) Les Data Archives en Amérique du Nord et en Europe

Des centres d'archivage existent dans les pays d'Europe et en Amérique du Nord.

Les *Data Archives* qui se constituent en premier le sont au tout début des années soixante et au sein de grandes universités. Aux États-Unis deux institutions prennent naissance : le *Roper Centre* (Université du Connecticut) dont il a déjà été question plus haut et l'*Inter-University Consortium for Political and Social Research* (ICPSR à l'Université du Michigan) : le premier archive principalement les données produites par

les instituts de sondage au plan mondial et le second se constitue comme un club d'universités dont les membres accèdent à des données d'enquêtes universitaires, de statistiques socio-démographiques et historiques (c'est par exemple l'ICPSR qui réalise alors l'informatisation de la Statistique Générale de la France). En Allemagne se crée le *Zentralarchiv* (Université de Cologne), puis le ESRC *Data Archive* (Université d'Essex) en Grande-Bretagne sur une logique de banque de données d'enquêtes ou de données de statistique sociale. La création de ces deux centres met fin à l'idée initiale de l'ICPSR de jouer le rôle d'une banque mondiale. Le rôle d'Erwin K. Scheuch, qui a séjourné au *Roper Centre*, a été particulièrement important dans cette évolution. À l'origine de la création du *Zentralarchiv* de Cologne, il est aussi l'un de ceux qui insistent particulièrement sur l'infléchissement de ces *Data Archives* vers l'utilisation immédiate à l'inverse d'un processus d'archivage historique pour le futur. Le rôle joué dès lors par les *Data Archives* dans la diffusion des données, l'aide apportée à l'utilisateur, leur rôle en matière de documentation des données et de formation des utilisateurs (avec la création d'écoles d'été dont la plus célèbre est celle de l'ICPSR) s'inscrivent dans ce cadre. C'est ce qui caractérise véritablement les *Data Archives* et explique leur importance.

Dans la décennie qui suit, un véritable mouvement des *Data Archives* se développe : d'autres pays européens se joignent aux trois pères fondateurs (la Norvège bien sûr mais également les Pays-Bas, la Belgique, la Suède, le Danemark, se dotent de *Data Archives* tous constitués sur le modèle allemand ou britannique) et des projets de coopération européenne prennent naissance. L'Europe du sud reste nettement à l'écart de ce mouvement puisque seule l'Italie rejoint au début des années soixante-dix le mouvement, alors que le Portugal, l'Espagne, la Grèce en sont absents pour d'évidentes raisons politiques (ce point permet de souligner que l'accès aux données de sciences sociales ne constitue pas qu'un enjeu de bon fonctionnement de la communauté scientifique. Il en va également du fonctionnement démocratique et de la confiance entre l'État et les citoyens.)

Ces centres ont chacun leurs caractéristiques propres, héritières de leur histoire, mais aussi de la structure de la production des données dans chaque pays, des liens particuliers entretenus avec la statistique publique, de l'organisation des universités et de la recherche, et du fonctionnement de la politique nationale de recherche. Mais le champ des données archivées est désormais très large, et la question des données issues de la statistique publique, très tôt prise en compte par exemple en Grande-Bretagne, prend progressivement plus d'importance.

La France, quant à elle, est absente de ces réseaux de *Data Archives* pendant près de vingt ans : il faut attendre le début des années quatre-vingt pour qu'un pas soit franchi avec la création au sein du CNRS de la Banque de Données Socio-Politiques (BDSP implanté à l'Institut d'Étu-

des Politiques de Grenoble et intégrée aujourd'hui au CIDSP du CNRS) puis du Lasmas à Paris. Cette absence est à la fois la cause et la conséquence d'un certain retard des sciences sociales française vis-à-vis de l'accès et de l'utilisation de données empiriques. Curieuse situation puisqu'un certain nombre de français sont présents aux conférences internationales convoquées au milieu des années cinquante par l'Unesco sur ces questions (on pense notamment au politiste Matteï Dogan ou à Raymond Boudon). Globalement, on peut dire que les efforts entrepris un peu partout en Europe pour la constitution de *Data Archives* n'ont pas été suffisamment relayés en France où la création de la BDSP et du Lasmas est davantage le fruit d'initiatives individuelles que d'une volonté nationale forte : Frédéric Bon pour la BDSP, l'équipe du Département d'analyse secondaire (Das) créé au Centre d'Études Sociologiques par Raymond Boudon, où Jacqueline Frisch joue un rôle important, à l'origine de la création du Lasmas par Alain Degenne.

La comparaison avec le Canada est intéressante. Ce pays se caractérise en effet par l'existence d'un des appareils statistiques les plus performants au monde et une faiblesse de la production d'enquêtes académiques. La proximité avec les États-Unis, liée à l'absence d'une politique nationale d'archivage, a conduit les universités canadiennes à intégrer les réseaux américains, (en particulier l'ICPSR). La question de l'accès des chercheurs aux données de Statistique Canada restait posée. La mobilisation des universitaires et du très dynamique réseau des bibliothécaires universitaires a abouti à l'adoption en 1996 de *l'Initiative de Démocratisation des Données* (IDD). Cette initiative associe la Fondation des Sciences Sociales et des Humanités (HSSFC), les bibliothécaires universitaires, Statistique Canada et les ministères fédéraux sur une logique de partage des compétences, des services et des financements. Elle a permis de poser les premiers jalons pour une politique nationale d'archivage de données d'enquêtes et de mise en réseau des centres existants.

c) Les réseaux de Data Archives

Peu à peu des réseaux internationaux se constituent.

Ces *Data Archives*, constitués sur une période de près de vingt ans, opèrent au milieu des années soixante-dix une mise en réseau américaine d'abord, européenne ensuite, internationale enfin puisque le mouvement s'étend à l'Australie, au Canada et que de nouveaux centres d'archivage se développent dans d'autres universités américaines (mais d'ampleur nettement plus limitée que l'ICPSR, voire même seulement destinés à alimenter l'ICPSR). La mise en réseau européenne se réalise en 1976 par la création du Cessda (*Council of European Social Sciences Data Archives*). Cette mise en réseau retrouve le projet fondateur, celui d'un développement de la recherche comparative européenne notamment. Le Cessda est aujourd'hui un club professionnel (cf. en annexe la liste des représentants nationaux) dont le rôle est officiel-

lement reconnu par des instances de coopération scientifique comme l'Unesco ou la *European Science Foundation*. L'internationalisation des *Data Archives* se concrétise un an après par la création en 1977 de l'Ifdo (*International Federation of Data Organizations*), qui reprend à peu de choses près les grands principes et le mode de fonctionnement du Cessda. Enfin l'Iassist (*International Association for Social Science Information Service and Technology*) rassemble les professionnels de ces *Data Archives* et contribue fortement sur le plan international aux débats et à l'innovation tant sur le plan de l'organisation que sur celui des outils.

Ces organisations réunissent chaque année leurs experts, le Cessda et l'Iassist notamment. Ainsi toute une série de séminaires de travail, d'échanges de savoir-faire, ont permis la réalisation de produits et d'outils destinés à faciliter la recherche comparative européenne : catalogues de données informatisés, disponibles pour chaque pays en langue anglaise, possibilités d'interroger simultanément ces catalogues dans toutes les banques de données membres du Cessda, standards de description des fichiers de données archivées, etc. Le mode de fonctionnement du Cessda – le partage des savoir-faire et expertises, la collaboration dans un esprit de réseau et de club – permet aux *Data Archives* les moins bien dotés en personnel et en ressources de ne pas être complètement laissés de côté par de tels développements et de bénéficier des avancées réalisées par les puissants *Data Archives* qui restent aujourd'hui le *Data Archive* britannique et le *Zentralarchiv* allemand. Le rôle et la contribution du Cessda au développement d'outils facilitant les conditions de la recherche comparative européenne ont été particulièrement reconnus par le rapport « *Social Science in a European context* » rédigé à la demande de la *European Science Foundation* par Howard Newby en 1992. Le CIDSP-BDSP participe activement aux activités du Cessda, ce qui lui donne une visibilité importante au niveau international.

d) Une politique de production de grandes enquêtes universitaires

L'International Social Survey Program réalise chaque année une enquête dans 31 pays à partir du même questionnaire. Les panels sur les conditions de vie se répandent dans les pays de la CEE.

Ces *Data Archives* sont une infrastructure importante pour les sciences sociales. Leur mise en place s'est appuyée au départ sur l'existence de grandes enquêtes académiques que ces centres ont archivées et diffusées avant d'y inclure les données issues de la statistique publique. Dans la plupart de ces pays existent des programmes d'enquêtes régulières de grande ampleur, soutenus par une politique nationale ambitieuse, volontariste qui a imposé un financement important et de long terme dans ces outils d'observation. C'est le cas des États-Unis, de l'Allemagne et de la Grande-Bretagne notamment. Cette politique, énoncée et mise en œuvre par les agences de moyens ou conseils de recherche nationaux (la NFS aux États-Unis, l'ESRC-*Data Archive* en Grande-Bretagne et la DFG en Allemagne) a trouvé le soutien politique

des ministères concernés. Cela a permis la mise en place concertée d'enquêtes dans le cadre de programmes internationaux et la production de grandes enquêtes académiques, conçues d'emblée pour être utilisées très largement par les chercheurs à l'intérieur comme au-delà des frontières. Les coûts de production d'enquêtes sur la base d'un échantillon de taille suffisante pour garantir la qualité des résultats, sont en effet très élevés : outre les coûts directs d'interrogation, il faut inclure également les compétences en statistique, en codage, en documentation. Il faut également des équipements puissants permettant de traiter et d'analyser les informations recueillies. Cela nécessite de la part des chercheurs une formation particulière en techniques quantitatives, souvent insuffisante dans des disciplines encore marquées par leurs liens avec la philosophie sociale. Les universités ou les équipes de recherche, en dehors de quelques puissantes universités américaines, ne peuvent assumer seules de telles dépenses.

Dans le domaine de la recherche socio-politique et plus généralement en sociologie politique, de grandes enquêtes internationales et européennes existent. On peut mentionner les trois principales : l'*International Social Survey Programme (ISSP)* tout d'abord, *les World Values Studies*, et leur partie européenne *European Values Studies*, *les Eurobaromètres* enfin. Ces trois types d'enquêtes sont disponibles pour la communauté des chercheurs mais avec des délais d'embargo différents selon les cas.

Dans le domaine de la recherche sociologique des exemples existent également. À la suite de l'enquête pionnière américaine du *Panel Study of Income Dynamics* (PSID, commencé en 1968) et le plus souvent sous l'impulsion des agences nationales de recherche (telles que la DFG en Allemagne ou l'ESRC en Grande-Bretagne), et grâce à l'inscription de ces productions dans une politique de financement continu de long terme, un grand nombre de pays européens ont entrepris à leur tour de réaliser des enquêtes nationales de suivi (panels) auprès des ménages

Le Panel Communautaire des ménages (ECHP), sous l'égide d'Eurostat, a désormais pour base les panels nationaux produits par des équipes universitaires là où elles existent (alors que la participation française est assurée par l'Insee). Le programme « *Panel Comparability Project* » (Paco), financé dans un premier temps par la *European Science Foundation* (entre 1990 et 1993), relayée ensuite par le programme capital humain et mobilité de la Commission européenne, vise à constituer une base de données à partir des panels de 7 pays (USA, Luxembourg, Allemagne, Lorraine pour la France, Hongrie, Pologne et Grande-Bretagne) pour une réutilisation en vue de travaux comparatifs.

Il apparaît clairement dans tous ces exemples que la condition nécessaire pour produire des enquêtes académiques est l'existence d'une politique nationale qui assure la continuité des financements directs (production proprement dite) et indirects (compétences scientifiques et

techniques, équipements, infrastructure, centre d'archivage, standardisation des classifications, standardisation de la documentation). Les pays qui ont mis en place une telle politique se sont attachés à ce qu'elle soit fortement incitative, cohérente et systématique, allant des conditions de production à celles de réutilisation des données par les chercheurs en passant par celles du dépôt au centre d'archivage.

Faute d'une telle politique, la recherche empirique française en sciences sociales a pris du retard. Ainsi, le caractère scientifique d'une partie des activités de l'Insee, positif par bien des aspects, et qui explique largement l'intérêt des chercheurs pour ses productions, a eu pour contrepartie un éloignement progressif des chercheurs de la pratique quantitative liée à la production des données. Les savoirs liés aux enquêtes sont dans les instituts producteurs tels que l'Insee ou l'Ined, pas dans les universités. Peu de chercheurs la partagent. Le Panel lorrain, produit par une équipe de l'Université de Nancy en 1985, a été abandonné en 1990, faute de soutien financier. Absence de compétences et absence de politique de financements de long terme, se nourrissent l'une de l'autre. Absents d'une grande partie des grands programmes internationaux de production d'enquêtes, les chercheurs français ne possèdent pas les compétences de leurs collègues étrangers dans ces domaines, sont exclus des travaux auxquels ils aboutissent, ne participent pas aux échanges d'expérience, a fortiori n'accumulent pas. De même la possibilité de co-productions de la Recherche avec l'Insee ne peut non plus trouver de support financier.

II.2. État des lieux en France : le retard français

Malgré la présence de la Banque de Données Socio-Politiques et du Lasmas, la France reste encore largement absente des grands débats internationaux

En France le CIDSP-BDSP et le Lasmas-Institut du Longitudinal ont fait avancer l'accès aux données et contribué à sauvegarder des données perdues par leur producteur. Ils apportent un travail important en termes de valeur ajoutée en matière de documentation et de « métadonnées », tout en contribuant eux-mêmes et par leur réseau d'utilisateurs à accroître l'utilisation de ces données pour la recherche. Ils ont aussi construit des liens entre utilisateurs et producteurs de données, en particulier avec les organismes publics. Enfin ils assurent une mission de formation à l'utilisation des données des grandes enquêtes. Ceci correspond bien aux différentes missions remplies par les *Data Archives* à l'étranger décrits plus haut. Il existe par ailleurs des bases constituées par des chercheurs autour de données qui suscitent des réticences ou des inquiétudes des intéressés (données fiscales, données pénales, etc.), ou dans des domaines très particuliers (les transports urbains par exemple).

Cependant de nombreuses difficultés restent non résolues. D'autres sont apparues plus récemment. La restriction dans la convention CNRS-INSEE

gérée par le Lasmas à la diffusion aux laboratoires CNRS s'est maintenue alors que la demande des universitaires non rattachés à des laboratoires CNRS s'accroît. Le développement général des coproductions (dans l'idée de répartir des coûts fort élevés) a pour effet de bloquer la diffusion des données, faute de définition en amont des conditions de cession. L'inflexion dans un sens parfois négatif des politiques de certaines administrations, les restrictions imposées par la Cnil sur les données infracommunales, en particulier celles issues du recensement, sont d'autres éléments inquiétants. Du côté de la BDSP, le dépôt des données pour l'archivage ne s'accompagne pas toujours d'un droit à la diffusion.

La croissance du nombre de données archivées nécessite aujourd'hui des moyens techniques et en personnels beaucoup plus importants. Surtout, l'absence d'une structure permettant de donner des garanties déontologiques (conseil d'administration et conseil scientifique notamment) et d'assurer la sécurité des données (zone de sécurisation, personnels et équipements ad hoc) ne permet pas de résoudre les problèmes liés à la diffusion aux universités, à l'utilisation de données sensibles par les chercheurs, et de renforcer les liens entre utilisateurs et producteurs de données.

Pour mesurer ce qui reste à faire pour assurer aux sciences sociales une infrastructure solide et mettre la France en situation de s'insérer avec plus de visibilité dans les réseaux européens et internationaux, la mission a tenté de dresser un état des lieux qui s'est voulu le plus large possible. Une enquête a été effectuée auprès des laboratoires universitaires et CNRS (voir annexe), auprès des grands instituts de recherche, enfin auprès des organismes producteurs de données publiques pouvant intéresser les sciences sociales. Il faut souligner que la mission a bénéficié sur ce point d'une très grande collaboration de ces différents partenaires. Sans être exhaustif[15], ce bilan permet de dessiner les traits généraux de la situation en France sur trois plans : l'accès aux données, l'utilisation des données, la place des chercheurs dans la production des données.

a) L'accès aux données

Trois conditions doivent être réunies pour que des chercheurs puissent accéder à des données déjà existantes, quelle qu'en soit l'origine. Il faut que ces données soient archivées, il faut qu'elles soient correctement documentées pour qu'un tiers sache comment il peut les utiliser, il faut enfin que le producteur décide d'en accorder un droit d'usage. On

[15]. Il existe par ailleurs des organismes privés ou semi-publics produisant des données et des analyses intéressant les sciences sociales (tels le Crédoc, Agoramétrie ou les instituts de sondage), qui n'ont pas fait l'objet d'enquête dans le cadre de cette mission. La BDSP a obtenu le dépôt de quelques enquêtes de BVA.

examinera successivement la situation française sur ces trois plans, pour les données publiques comme pour les données académiques. Les questions portant sur les données privées (instituts de sondage) n'ont pas pu être examinées de façon aussi précise et mériteraient de plus amples développements. En ce qui concerne les données publiques, il faut d'entrée de jeu prendre en compte la différence entre données administratives exhaustives et souvent nominatives et grandes enquêtes sur échantillon.

L'archivage

Conserver les données pour pouvoir les réutiliser implique une mise à niveau régulière des supports informatiques, ce qui n'est pas toujours réalisé en France.

Conserver des données conditionne évidemment la possibilité de mettre à disposition d'un tiers ces données ultérieurement. Il faut souligner que cela conditionne également la réutilisation par le producteur même. S'agissant de données sur support informatique, comme c'est le cas désormais, conserver suppose également une veille informatique consistant en une mise à niveau régulière, nécessaire étant donné le changement continuel des équipements et des logiciels. Tous les organismes ont en mémoire les bandes jetées parce que devenues illisibles.

De très nombreuses données publiques ont été ainsi perdues dans un passé encore proche. On peut citer notamment le cas du recensement de 1954. La situation s'est depuis améliorée, mais reste inégale. Pour décrire cette situation, il faut prendre en compte le rôle des Archives contemporaines qui ont pour mission d'archiver entre autres les données de ce type. Les Archives contemporaines effectuent cependant des choix, qui sont notamment fonction de critères en matière de documentation qui rend seule possible la réutilisation. C'est un point de blocage important en matière d'archivage des données publiques. Certains organismes archivent uniquement dans leurs services. D'autres déposent une copie aux Archives contemporaines. D'autres n'archivent pas ou très inégalement. Un organisme comme l'Insee archive maintenant systématiquement les enquêtes aux Archives contemporaines. Il n'en va pas de même dans plusieurs départements statistiques ministériels ou agences gouvernementales productrices de données.

En matière de données académiques, la situation est là encore marquée par une très grande inégalité. Un institut de recherche comme l'Ined dispose d'un service d'archivage et a commencé à déposer des copies aux Archives contemporaines. Il n'existe par contre aucun protocole pour des laboratoires de recherche où des chercheurs produisent des enquêtes. La BDSP fait un travail d'incitation auprès des chercheurs dans les domaines qu'elle couvre, mais de très nombreuses enquêtes de chercheurs, certaines très importantes pour l'histoire des sciences sociales, ont été physiquement perdues ou sont devenues inutilisables faute de veille technique du point de vue des matériels et des logiciels.

Pour situer le rôle d'organismes comme le Lasmas ou la BDSP, on peut ainsi souligner que le premier a pu rendre au ministère de la Culture, qui les y avait déposées, les premières enquêtes sur les Pratiques culturelles, perdues depuis par ce ministère, et à l'INSEE la première enquête FQP (Formation Qualification Professionnelle) dans une version plus complète. De même, la BDSP rend souvent à des chercheurs des données qu'ils ont eux-mêmes produites et perdues ; il faut noter qu'elle a aussi passé quelques accords pour archiver des données produites par des instituts de sondage privés qui par ailleurs n'ont pas de politique bien affirmée en la matière.

La documentation

Sans documentation les données sont inutilisables par des tiers, mais aussi à plus long terme par les services producteurs eux-mêmes.

C'est le point tout à fait crucial pour que les données puissent être utilisées par des tiers. Mais, on l'a vu, il conditionne également la possibilité pour les organismes de faire un dépôt aux Archives contemporaines. Tous les *Data Archives* ont nécessairement des exigences en ce sens. Le Lasmas-IdL, lorsqu'il acquiert des données, demande la documentation la plus complète possible sur ces données. Le CIDSP-BDSP dispose d'un guide de l'utilisateur, à l'image de celui des Archives contemporaines, avec des recommandations et des exigences minimum en matière de documentation des données.

Les situations sont là encore extrêmement diverses. En ce qui concerne le champ des données publiques, l'Insee apparaît naturellement comme le mieux organisé, même s'il y a encore des difficultés et des priorités. Ceci a conduit à une expérience d'échanges de services avec le Lasmas (aide à la documentation d'une enquête en échange de l'accès aux données). Ailleurs, il faut d'abord distinguer données administratives qui n'ont pas vocation à être diffusées (mais qui constituent des sources intéressantes pour la recherche et qui peuvent servir de base d'échantillonnage), peu ou pas documentées, et les enquêtes sur échantillon. Celles-ci, produites par les administrations dans un but de connaissance immédiate, sont très inégalement documentées. La question de la documentation constitue en fait le point névralgique bloquant la diffusion des données pour les chercheurs, par manque de moyens et de temps. Il faut remarquer que ce peut être également un obstacle à des réexploitations internes par les services concernés. Pour les données académiques, il faut là encore distinguer entre instituts de recherche disposant de protocoles et laboratoires de recherche. La situation est plus favorable dans les premiers. Au niveau des laboratoires les données ne sont le plus souvent pas documentées et sont donc inutilisables. Il faut observer que les chercheurs capitalisent peu le travail d'exploitation des données qui apporte des informations fines et qui manque fréquemment dans la documentation existante.

Le droit d'usage pour les chercheurs

Le droit d'accès aux données est facilité par l'existence de conventions avec les organismes producteurs. Les politiques différentes de diffusion commerciale ou non, les contraintes de secret statistique ou d'exploitation prioritaire, limitent l'usage des données.

La politique de diffusion des fichiers d'enquêtes de l'Insee pour les chercheurs a été plusieurs fois infléchie. Le rôle joué par le Lasmas-IdL est ici un élément déterminant dans le tableau que l'on peut faire de la situation actuelle. Avant la première convention signée en 1986 entre l'Insee et le CNRS pour le Lasmas, les chercheurs accédaient aux fichiers grâce à leurs liens personnels et donc de façon très inégale. Ils avaient également la possibilité de demander des tableaux à façon, souvent très longs à obtenir et relativement coûteux. Des laboratoires ont cependant acquis au fil du temps des fichiers ou bouts de fichiers, en particulier des recensements. La première convention signée par le Lasmas permettait d'acheter à un prix très faible des fichiers pour l'ensemble des chercheurs des laboratoires du CNRS. Elle définissait des délais rapides de mise à disposition et des obligations en retour du Lasmas et des chercheurs. La deuxième convention signée quelques années plus tard accompagne une redéfinition de la politique commerciale de l'Insee qui, tout en maintenant un tarif préférentiel pour la recherche et en conservant au CNRS le bénéfice important d'être considéré comme un seul site, accroît fortement le coût d'acquisition des données. Aucune des deux conventions successives ne couvre le champ des universitaires non rattachés à des laboratoires du CNRS, qui doivent acquérir les données directement à des coûts trop importants eu égard aux moyens dont ils disposent. La pratique grandissante des co-productions conduit à restreindre de fait le champ d'application des conventions, sauf à définir dès l'amont entre les co-producteurs une politique en la matière. Enfin la diffusion tient évidemment compte des contraintes introduites par la Cnil pour les données infracommunales du recensement en particulier. Un groupe de travail Lasmas-Insee a été mis en place pour examiner les problèmes posés de ce fait aux chercheurs.

Pour les autres données publiques provenant des départements statistiques des ministères, des administrations et des agences gouvernementales, il faut là encore distinguer les données administratives, qui posent un problème d'anonymisation et nécessitent le passage par la Cnil (ou pour les données portant sur les entreprises par le Comité du secret statistique), des enquêtes. Il n'existe aucune politique d'ensemble repérable dans ce champ. Il a ainsi existé pendant trois ans une convention DEP (ministère de l'Éducation nationale)-Lasmas, sur le modèle de la convention Insee-Lasmas, mais l'application est restée limitée. Il existe par contre une convention Céreq-Lasmas, dont il faut remarquer que le champ n'est pas restreint aux seuls chercheurs du CNRS. Ailleurs la demande la plus fréquemment exprimée par ces organismes est celle de la définition du chercheur et des garanties de bonnes pratiques (garanties appropriées), ce que l'on peut traduire en termes de demande de professionnalisation et d'organisation du milieu. Est cependant évoquée la peur de voir produire des travaux sur des questions politiquement sensibles. Ce sont les organismes les moins

habitués aux chercheurs qui l'expriment le plus fréquemment. Le conflit d'intérêt est souvent important entre la demande des administrations de conclusions en termes de politique publique et les objectifs des chercheurs. La question de la liberté de publication peut également être un objet de tensions. On observe cependant une demande croissante de ces organismes en direction des chercheurs pour exploiter des données, soit sous forme de demande ciblée, soit sous forme de groupes d'utilisateurs (forme qui se développe également à l'Insee). Le milieu est jugé de ce point de vue trop étroit par rapport aux besoins. Inversement les chercheurs expriment aussi le souhait de pouvoir disposer des données hors étude ciblée pour les besoins du producteur. La constitution de groupes d'utilisateurs est positive pour faciliter l'accès aux données à condition qu'elle ne se traduise pas en fermeture pour les autres utilisateurs.

Les instituts de recherche (Ined, Inra, Cee…) ont pour la plupart des conventions particulières avec les producteurs de données publiques à des coûts et dans des conditions assez différents. Ils expriment des positions diverses quant à la préservation de ces liens directs selon qu'ils leur apparaissent satisfaisants ou pas. Certains souhaitent ainsi pouvoir accéder à la convention CNRS-Insee via le Lasmas, tout en conservant par ailleurs les liens particuliers établis.

L'accès pour les autres chercheurs aux données académiques produites par les chercheurs, soit dans le cadre des instituts de recherche, soit dans le cadre des laboratoires, est encore moins défini. La politique de partage des données des instituts de recherche est très incertaine, y compris à l'intérieur même de ces instituts, mais le débat est en cours sur les limites à mettre au droit d'exploitation prioritaire du chercheur, sur la définition des droits dans le cadre des co-productions ainsi que d'une façon générale avec les bailleurs de fond. Du côté des chercheurs relevant de laboratoires CNRS ou universitaires, le statut des données du point de vue de la propriété et de la gestion du droit d'accès reste extrêmement flou pour les chercheurs eux-mêmes. Le CNRS n'a pas encore mis en place un protocole sur ce point. Dans les faits ces données sont rarement utilisées par d'autres, notamment parce qu'elles ne sont pas documentées. Cependant la BDSP, qui en archive quelques unes, incite à ce partage. En tant que dépositaire, elle gère le droit d'usage, conformément à un protocole établi au cas par cas avec le producteur.

Ce bilan peut être complété par une analyse de la perception et des demandes des chercheurs, telle qu'elle ressort de l'enquête auprès des laboratoires menée dans le cadre de la mission. Dans l'ensemble ceci recoupe en grande partie les observations déjà faites au niveau du Lasmas-IdL à travers le réseau de ses utilisateurs, mais amène cependant à préciser certains aspects :

– La situation varie en fonction des disciplines et des domaines. Ceci tient à la fois à la formation inégale des chercheurs en matière d'utilisation de fichiers d'enquêtes de grande taille, à des politiques différentes des organismes détenteurs de données mais aussi à des caractéristiques particulières de ces données (par exemple données d'entreprises) plus ou moins sensibles. cela tient aussi au rôle et à la position variable des commanditaires de recherches qui peuvent avoir ou non des accès particuliers aux données que les chercheurs n'ont pas pu obtenir pour leur propre compte. Le rôle des relations personnelles dans nombre de cas demeure important.

– Les difficultés les plus souvent pointées par les chercheurs recoupent en grande partie les observations déjà faites au niveau du Lasmas. L'impact négatif des coproductions qui interdit ou limite la diffusion, les questions de coût d'accès, pour les universitaires, des fichiers de l'Insee non couverts par la convention avec le Lasmas, la difficulté que pose l'accès au fichier Sirène (coûts élevés et problèmes d'accès en ligne, nécessaire pour obtenir des appariements de bonne qualité), la question du recensement, les conditions de cession des données par plusieurs administrations qui en limitent l'usage à des contrats d'études spécifiées ou refusent l'accès, mobilisent l'attention. Est également soulignée l'absence de négociation au niveau global avec la Cnil. Chacun mène seul la négociation. Il en va de même avec le Comité du secret qui gère l'accès aux données des entreprises. Les remarques portent également sur le manque de formation pour certains à l'utilisation des données (ceci touche y compris les économistes sur des formations pointues, alors que pour les sociologues il s'agit plus du coût initialement lourd d'entrée dans la maîtrise des logiciels d'analyse), et dans certains cas d'insuffisance des équipements pour traiter des grandes masses de données, des problèmes de réseaux pour accéder aux centres de calcul. Ces derniers points renvoient de façon claire aux problèmes que rencontrent les chercheurs du point de vue de l'utilisation des données.

Enfin la question de l'accès aux données de l'Union européenne, et en tout premier lieu à celles d'Eurostat, est posée par l'ensemble des chercheurs comme par les producteurs de données publiques. L'IRD dont le champ de recherche est tout autre souhaite également la constitution d'une base d'enquêtes au niveau européen sur les pays en développement, tout en soulignant que le problème est ici de préserver des espaces de travail commun avec les pays producteurs.

b) L'utilisation des données

Par comparaison avec d'autres pays, le développement de

Faire des travaux quantitatifs en sciences sociales dépend bien évidemment en grande partie de la possibilité d'accéder aux données des grandes enquêtes. Les difficultés passées et celles qui subsistent

la sociologie quantitative apparaît en retard en France. L'insuffisance de la formation à l'utilisation des données affecte la compétence des chercheurs à traiter des enquêtes. Le passage d'une informatique lourde et centralisée à une informatique répartie induit des besoins en matériel et logiciel mal pris en compte dans les budgets.

expliquent en partie les faiblesses de la recherche en France sur ce plan, en particulier en sociologie. Par comparaison avec quelques pays tels que les États-Unis, le Royaume-Uni ou les Pays-Bas par exemple, le développement de la sociologie quantitative apparaît en retard en France, ce qui se traduit par une très faible présence sur la scène internationale. Une très large partie des travaux effectués se situe de surcroît dans le périmètre des organismes publics producteurs de ces données. Il existe certainement des racines historiques anciennes à cette situation, malgré l'existence d'une tradition précoce mais restée marginale de sociologie empirique. Dès les années 30, alors que se développe dans d'autres pays un fort mouvement d'enquêtes (les *Social Surveys* par exemple), la particularité de la France marquée par des orientations de recherche plus spéculatives est notée à l'étranger (cf. Savoye, 1994).

La croissance d'un appareil statistique au champ très large a contribué à élargir le fossé entre les chercheurs et leurs données à tous les niveaux et à terme à affaiblir l'intérêt des chercheurs pour les données. La compétence même des chercheurs à traiter des données s'en est trouvée affectée. Les organismes producteurs de données publiques soulignent tous l'étroitesse du milieu des chercheurs en la matière et l'absence de visibilité qu'ils ont de ce milieu qu'ils ne connaissent souvent qu'au travers de relations ponctuelles.

Il faut cependant noter que, s'il y a bien un retard significatif de la France sur ce plan, la mission a pu constater dans plusieurs pays des inquiétudes partagées sur l'insuffisance de la recherche empirique en sciences sociales dans l'ensemble des disciplines (voir par exemple le rapport en Allemagne de Richard Hauser, Gert Wagner et Klaus Zimmermann[16] et au Canada celui de Paul Bernard). Face à une richesse croissante des données disponibles pour l'analyse, les rapports s'accordent sur l'insuffisance de la formation des étudiants dès les premières années des universités, comme de la formation continue. Cette faiblesse est patente en France où, par exemple, le nombre de thèses soutenues chaque année en sociologie quantitative est quasiment insignifiant.

La mise à disposition des données facilitée par le Lasmas et la BDSP a contribué à développer ces travaux (cf. *Dix ans d'analyse secondaire au Lasmas-Institut du Longitudinal - Bilan de la convention CNRS-Insee*, 1997). Une expansion plus forte passe cependant par une réflexion sur la formation à l'utilisation des données. À travers la formation continue, le CNRS a contribué à diffuser outils et méthodes, par des stages réguliers ou des Écoles d'été (par exemple celle de Lille). La Formation permanente du CNRS ne permet cependant d'inclure qu'un faible nombre de doctorants et d'universitaires pour des raisons budgétaires. Elle n'a

[16]. R. Hauser, G. Wagner, K. Zimmermann, 1998, Memorandum : Erfolgsbedingungen empirischer Wirtzchaftsforschung und empirisch gestützter wirtschafts- und sozialpolitischer Beratung, *Zuma Nachrichten*, n° 43, Nov. 98, pp 134-144.

par définition aucun impact sur la formation initiale à l'université. Quelques enseignements de troisième cycle (DESS) commencent à se mettre en place.

Le dernier aspect à souligner touche aux aspects matériels qui conditionnent le traitement des données des grandes enquêtes dans l'ensemble des disciplines concernées. On est passé d'une informatique lourde et centralisée à une informatique répartie. Après avoir eu accès aux grands centres de calcul (le Circe à Orsay, le Cnusc à Montpellier, puis maintenant le Criuc à Caen et le CICG à Grenoble), ce qui a entraîné régulièrement des problèmes de migration des données d'un centre à l'autre, la situation informatique a évolué rapidement du fait de la banalisation des micro-ordinateurs et de la mise en place des réseaux de gros débit. Les chercheurs ont dû s'adapter à cette nouvelle donne et s'équiper d'ordinateurs (qui doivent être puissants) et de logiciels (qui doivent être performants), pour lesquels il faut intégrer les mises à niveau, les renouvellements et les licences dans les budgets. Dans ces conditions, un chercheur en sciences sociales coûte plus cher. L'enquête auprès des laboratoires le souligne bien, nombreux sont les laboratoires qui insistent sur leurs problèmes de budget pour cet aspect des choses.

c) La place des chercheurs dans la production des données

La production directe de données sociales par la recherche et l'université reste très rare. La question des financements est un frein essentiel.

Le rapport des chercheurs à la construction de leurs données est un point clé de toute production scientifique à caractère empirique. En sciences sociales, dans la mesure où une part importante, quoiqu'inégale, de la production des données est assurée dans le cadre de la statistique publique, ce rapport s'exerce sous plusieurs formes. De façon complètement extérieure, le chercheur, à condition de disposer d'une documentation suffisante, prend connaissance du minimum d'informations lui permettant de comprendre le contexte de construction des données, leur signification, leurs limites. Ceci suppose une documentation des données par le producteur, évoquée plus haut, et la formation de l'utilisateur à la connaissance d'une enquête particulière. C'est ici la dimension la plus fréquente du rapport des chercheurs à leurs données. Elle est cependant insuffisante sur deux points : elle ne permet au chercheur d'intervenir secondairement sur la construction de ses données que dans les limites déterminées par le recueil initial, elle ne permet pas non plus au chercheur d'avoir une perception fine des problèmes de construction, essentiels pour définir la portée des résultats. Sur ce plan, l'implication des chercheurs dans la production directe des données d'une part, et plus indirectement leur présence en amont de la production des données à travers les consultations préalables à la mise en place d'une enquête, apparaissent comme des chaînons indispensables d'une rigueur scientifique.

La production directe d'enquêtes

La production de données sociales fait en France l'objet d'un découpage en champs bien délimités mais non exhaustifs. À chacun de ces champs correspond un Institut, souvent un EPST (comme l'Inra, l'Inrets, l'Orstom, l'Ined, etc.) ou un service administratif et de recherche (Céreq, Darés, etc.), qui prend en charge de manière régulière la constitution de données nouvelles sur les thèmes relevant de sa compétence.

En dehors de ce contexte institutionnel, quatre situations types ont jusqu'ici permis la production de données par la recherche et l'université.

1. L'enquête annuelle de l'Observatoire Interrégional du Politique (OIP) est un bon exemple de données produites avec un financement décentralisé (régions). La première enquête a été effectuée en 1986. L'intégralité des questionnaires est en ligne et les données sont facilement accessibles via la BDSP après un embargo de six à sept mois. Cette enquête se signale par sa pérennité, qui n'est d'ailleurs sans doute pas étrangère à la décentralisation de son financement.

Cette situation est toutefois bien rare. Le plus souvent, même dans ce domaine socio-politique, les financements sont difficiles à rassembler et le suivi s'en ressent. Par exemple, les enquêtes post-électorales du Cevipof n'ont pas toujours pu être menées (voir annexe).

2. L'émergence d'un problème social important – ou perçu comme tel – représente vraisemblablement et jusqu'à aujourd'hui la source de financement la plus consistante. L'exemple le plus connu est le Sida dont l'expansion a suffisamment inquiété pour que soit financée une des plus grosses enquêtes jamais produite par la recherche en France (effectif d'environ 20 000 personnes).

Contrairement à la situation précédente, il s'agit là d'opérations ponctuelles, situation malheureusement de loin la plus fréquente. Comme les données ne sont pas principalement vues comme devant être réutilisées par d'autres, l'archivage et la documentation ne sont souvent pas menés à leur terme. Vraisemblablement, il existe ainsi beaucoup de données, qui de fait sont perdues, oubliées, non répertoriées et inaccessibles.

3. L'insertion dans un dispositif international ou européen a aussi été à l'origine de données nouvelles. Les enquêtes EVS (*European Value Surveys*) en sont un exemple. Plus récemment, l'insertion de la France dans l'*International Social Survey Programme* montre que l'argument de la « chaise vide » (dans ce programme créé en 1984 à l'initiative de 4 pays on trouvait, dix ans plus tard, 25 pays dont tous les pays du G7 sauf la France) peut finir par permettre de trouver un financement public, bien que celui-ci demeure symbolique en regard du coût réel d'une enquête.

4. La coproduction d'enquêtes avec un institut officiel comme l'Insee est quasiment inexistante**.** Il faut noter cependant l'enquête *Modes de vie-Production domestique*. Des collaborations ont toutefois vu le jour notamment pour modifier ou introduire des questions dans les grandes enquêtes (voir plus loin). Mais elles sont ponctuelles et résultent de rapports interpersonnels. Les coproductions sont d'un coût élevé au regard des moyens dont disposent les chercheurs. L'intérêt de la coproduction est pourtant patent lorsque les enquêtes ont servi et servent la recherche scientifique de par leur(s) thème(s) et surtout leur continuité. C'est par exemple le cas de l'enquête FQP dont l'avenir est incertain mais dont le cofinancement pour en assurer la pérennité (de façon significative et non symbolique) devrait être de l'ordre de 2 ou 3 MF. Un tel financement ne peut trouver son sens qu'après la mise en place de mécanismes globaux permettant d'améliorer significativement la situation actuelle, caractérisée principalement par l'éclatement et la non-cumulativité.

À côté de ces enquêtes de grande taille, il existe une production d'enquêtes de petite taille (moins de mille individus). Ce sont en général des enquêtes ponctuelles sur des populations souvent spécifiques (les comédiens professionnels, les magistrats de la Cour des Comptes, les étudiants d'une université parisienne, par exemple). Dans leur très grande majorité, ces enquêtes peuvent être mises à la disposition des autres chercheurs ; quand cela n'est pas possible, la raison en est souvent la spécificité et la perte d'anonymat de la population visée ; sont souvent également évoqués les problèmes de documentation des fichiers. Il est à remarquer que parmi la centaine d'enquêtes décrites, quatre panels ont été mis en œuvre.

La consultation des chercheurs en amont de la production d'enquête

La consultation des chercheurs en amont de la production des enquêtes issues de la statistique publique est significative et doit conduire à nuancer les conclusions. Elle dessine aussi des pistes pour l'avenir.

On ne saurait s'en tenir à ce bilan pour dresser un état des lieux de l'implication des chercheurs dans la production des données en France. Si l'implication directe apparaît significativement plus faible que dans d'autres pays, la consultation des chercheurs en amont de la production des enquêtes issues de la statistique publique est significative et doit conduire à nuancer les conclusions. Elle dessine aussi des pistes pour l'avenir.

En assurant, de par sa convention avec l'Insee, une organisation du retour des travaux effectués par les chercheurs vers l'organisme producteur des enquêtes mises à disposition, le Lasmas a pu contribuer à placer les chercheurs plus près de la production même de ces données. Le simple retour des publications, en mettant en valeur le rôle d'une information ou ses limites dans le cadre de l'enquête, a un impact significatif sur l'évolution des enquêtes. Surtout le Lasmas a pu par exemple organiser, en mobilisant son réseau d'utilisateurs autour de l'enquête FQP, importante pour les travaux sur la mobilité sociale, une

expertise pour faire évoluer l'enquête de 1993. Il est à nouveau fortement impliqué dans les discussions actuelles sur l'opportunité de maintenir ou pas cette enquête.

Un organisme tel que le Lasmas permet d'accroître la participation des chercheurs en amont. Mais il existe aussi tout un faisceau de relations directes entre chercheurs et producteurs de données publiques, inégales mais significatives. Ces relations prennent des formes diverses et ont un caractère plus ou moins organisé. Les relations personnelles des chercheurs développées sur la base de leurs travaux dans un domaine particulier avec tel ou tel producteur, qu'il s'agisse d'un département statistique d'un ministère ou d'un département de l'Insee sur une enquête particulière, sont anciennes. Elles ont permis et permettent encore une intervention directe pour modifier, infléchir une enquête, parfois sur une question, parfois plus largement. Leur stabilité dans le temps est cependant tributaire du caractère personnel de cette relation. L'intérêt reconnu par les producteurs de cette présence a conduit ceux-ci à l'organiser plus systématiquement, d'une part en mettant en place des conseils scientifiques (c'est le cas de la Darés par exemple), d'autre part en généralisant les groupes d'utilisateurs qui permettent, au-delà de l'exploitation immédiate de l'enquête, d'engranger pour l'avenir des remarques utiles. Il s'agit là cependant de pratiques qui sont loin d'être égales d'un organisme à un autre, d'un département statistique à un autre, mais dont beaucoup de nos interlocuteurs ont souligné l'intérêt.

Enfin on ne saurait terminer ce tableau sans prendre en compte le rôle tout à fait important et original que joue le Cnis (Conseil national de l'information statistique) en la matière. Instrument de cohérence de la programmation statistique, dans l'esprit de la planification à la française, avec un périmètre d'action définissant de façon très large les enquêtes à prendre en compte (production d'intérêt public non limitée aux seules institutions chargées de la statistique publique au sens strict), le Cnis organise une consultation où l'ensemble des partenaires sociaux, dont la recherche et l'enseignement supérieur, sont présents. Ces derniers sont représentés au Conseil, mais ils sont également présents là où s'élaborent les avis et les propositions de modifications, dans les différentes formations du Cnis (et leur présidence) et les groupes de travail de ces formations. La mission a pu constater avec l'aide du Cnis que cette présence, quoiqu'inégale selon les formations, était significative et proportionnelle à la représentation institutionnelle au sein du Conseil. Elle apparaît plus ou moins étendue selon qu'on inclut ou non les chercheurs et statisticiens d'Instituts de recherche tels que l'Ined, très attentif aux discussions menées dans ce cadre. Le Lasmas s'est efforcé, dans la mesure de ses moyens actuels, d'assurer une présence régulière dans quelques formations. Enfin dans le cadre des groupes de travail, il est fait appel à des présentations de travaux des chercheurs pour éclairer l'avis des formations sur les enquêtes. D'une manière générale, l'impression qui prévaut est que lorsque les

chercheurs sont présents ils ont une incidence significative. Cependant cette présence n'est ni systématique ni toujours forcément représentative des travaux, assez dépendante des connaissances des personnes présentes initialement dans les formations. Il faut aussi souligner que la présence dans les formations représente un investissement en temps que les chercheurs ne consentent pas toujours. Il est vraisemblable également que nombre de chercheurs ignorent l'existence du Cnis ou mesurent peu le rôle qu'ils peuvent y jouer.

d) Un premier bilan de cet état des lieux

Les conclusions que l'on peut tirer de cet état des lieux sont donc nuancées. Si des jalons significatifs ont été posés par deux laboratoires du CNRS, le Lasmas et le CIDSP-BDSP, en matière de partage des données, la France ne dispose pas d'un instrument et d'une politique en la matière, équivalents à ceux qui ont été construits il y a une vingtaine d'années déjà aux États-Unis et en Europe. Elle est en retard pour prendre une place plus importante dans les réseaux déjà constitués dans le cadre de la construction européenne. La faiblesse de la recherche empirique, en particulier en sociologie, souffre particulièrement de cette situation, mais résulte également de l'insuffisance de la formation initiale à l'utilisation des données d'enquêtes et de l'absence de politique de production de grandes enquêtes universitaires. En revanche le caractère scientifique du travail sur les données qu'effectue l'Institut national de statistique, sa proximité avec la recherche, qui est une caractéristique française (que l'on retrouve au Canada), a permis de construire des liens en amont de la production des données, qui se sont étendus à d'autre producteurs de la statistique publique, qu'il importe de prendre en compte dans le bilan et de préserver, voire accroître, dans les propositions qui seront faites.

II.3. Nouveaux contextes, nouveaux enjeux

La politique à mener en France en matière d'accès aux données des grandes enquêtes devra aussi prendre en compte la nouvelle donne créée par l'accélération de la circulation des informations du fait des réseaux, les implications de l'intégration européenne, l'évolution actuelle des centres de diffusion des données .

a) Une accélération de la circulation des informations

Le web et les réseaux à gros débit permettent aujourd'hui d'accéder plus facilement et très rapidement à des informations partout dans le monde.

Transporter des fichiers de données de grande taille, accéder à des fichiers en ligne, soumettre des programmes à distance et obtenir des résultats sans transfert matériel des fichiers sont des possibilités nouvelles qui supposent seulement de disposer du matériel nécessaire.

On peut voir là le point de départ d'un véritable saut qualitatif pour les sciences sociales (William Sims Bainbridge de la *National Science Foundation*[17]). La possibilité de trouver des données adaptées au problème posé, permettant de vérifier les résultats dans des contextes différents et de faciliter les comparaisons, va nécessairement impliquer une demande croissante des chercheurs d'accès aux données hors du cadre strictement national, ce qui n'est pas toujours prévu par le cadre juridique définissant le partage des données, en particulier les fichiers issus de la statistique publique. Parallèlement, la difficulté à contrôler la circulation de l'information sur les réseaux s'est accrue et pose des problèmes nouveaux de sécurisation des données ainsi que de contrôle du droit d'usage sur celles-ci.

b) La construction européenne

De grandes questions pour la communauté européenne comme l'inégalité et la mobilité sociale relancent les efforts pour rendre comparables les données existantes et pour produire des données comparables.

Si d'incontestables progrès ont été réalisés grâce à la mise en réseau des *Data Archives* européennes, l'européanisation liée aux développements politiques de l'Union européenne pose de nouveaux enjeux. Depuis près de dix ans la recherche comparative européenne a pris une signification nouvelle. Quelques indicateurs sont à cet égard parlants : la structuration de grands réseaux européens de recherche (rôle du IV^{ème} puis V^{ème} PCRD), le développement de revues scientifiques européennes (par exemple *European Sociological Review* mais aussi les très nombreuses nouvelles revues de politique comparée européenne et d'études européennes), l'impact d'un certain nombre de programmes de recherche européens (du type *Beliefs in Governement* de la *European Science Foundation* ou *Whitehall project* de l'ESRC).

Faire le constat de l'européanisation de la recherche est sans doute un lieu commun aujourd'hui. Néanmoins, les conséquences en sont fortes pour les grandes questions traitées dans ce rapport : de nouvelles exigences portent sur l'analyse comparative et demandent donc d'accéder directement aux fichiers d'enquêtes.

Cette question se pose à trois niveaux :
1) Les demandes d'accès des chercheurs européens à des données des différents pays de l'Union européenne se sont multipliées. Ceci pose à la fois le problème des conventions diverses réglant l'accès aux fichiers, de la diversité des coûts d'accès et des métadonnées absolument nécessaires aux chercheurs moins à même de comprendre les contextes

[17]. Bainbridge W. S., 1999, *International Network for Integrated Social Science*. OCDE

nationaux. À titre d'exemple, le *Data Archives* d'Essex diffuse les données issues de la statistique publique aux chercheurs étrangers à des conditions très avantageuses, proches de celles consenties aux chercheurs britanniques. La convention CNRS-INSEE, gérée par le Lasmas, ne prend pas en compte actuellement cette diffusion et par ailleurs le Lasmas reçoit des demandes de documentation des données de la part de chercheurs étrangers dont les organismes de tutelle ont acquis des fichiers de données françaises directement à l'Insee.

2) La recherche comparative européenne, impulsée fortement par les programmes du PCRD et la Commission, génère de la part des chercheurs la demande d'autorisation de créer des bases intégrées à partir de fichiers nationaux. Ceci est pour l'instant difficile, voire impossible, mais va devenir absolument nécessaire.

3) Dans le même temps, s'organise au niveau européen la production de données directement comparatives, conçues dès leur production dans un plan d'observation commun. Ce processus est double. Il a d'abord été le fait des chercheurs qui ont mis sur pied des grandes enquêtes européennes. Il va devenir progressivement le fait de la statistique publique au niveau de l'Union Européenne (ce qui repose à nouveau la question de l'accès à ces données). Il existe ainsi un début de programmation de la statistique européenne, un peu sur le modèle du Cnis français, associant les différents partenaires et où sont représentés les chercheurs, dont les indications sont désormais prises en compte par les programmations nationales. C'est le cas notamment au niveau du Cnis qui les intègre dans les avis qu'il est amené à formuler sur les projets. Les enquêtes Emploi comportent une partie commune depuis longtemps (conformes aux recommandations du BIT et d'Eurostat).

Deux questions sont dès lors posées, celle de l'harmonisation a priori des nomenclatures et celle du partage des données produites sur un niveau européen. L'harmonisation des nomenclatures est discutée à la fois au sein de la communauté scientifique qui a mis en place des travaux de recherche en commun, et au niveau des instances européennes avec les instituts nationaux. Un des lieux d'articulation de ces deux processus est, pour la France, le Cnis. La question est en tout cas posée de la place des chercheurs dans le processus d'harmonisation des enquêtes, qui doit être prise en compte si l'on veut trouver un équilibre entre l'intégration commandée par les nécessités des politiques publiques et la prise en compte de la diversité historique des contextes nationaux. L'autre question est celle de l'accès aux données européennes pour les chercheurs. La mise à disposition est prévue pour toutes les enquêtes produites par les chercheurs, et déposées à cet effet au Cessda. Celle des fichiers issus de la statistique publique, et pour l'heure des données d'Eurostat, va se poser de façon croissante.

c) La multiplication des centres d'archivage et de diffusion des données

Alors que se sont créés il y a une vingtaine d'années des centres d'archivage et de diffusion des données à vocation nationale et même internationale, on assiste actuellement à une multiplication de centres à compétences thématiques. L'exemple des États-Unis est sur ce point tout à fait exemplaire. Il existe aujourd'hui plus d'une vingtaine de centres, dont certains étaient initialement des relais de l'ICPSR de Michigan, dans le cadre d'une politique de diffusion maximale des données. Le développement de ces centres liés à des Universités, l'émergence d'autres plus autonomes sur des thématiques de recherche particulière ont conduit depuis 1995 l'ICPSR à engager une réflexion sur l'évolution des modèles d'organisation institutionnelle du partage des données et sa propre évolution. La réflexion de la *National Science Foundation* va dans le même sens. La question centrale devient aujourd'hui celle des réseaux liant les centres de diffusion des données, celle de la navigation sur ces réseaux et de la recherche par les utilisateurs des données adéquates, celle de l'échange des données, enfin celle indispensable de l'harmonisation des outils de documentation et de diffusion qui conditionnent ce processus. Ces préoccupations rejoignent celles du réseau européen du Cessda. La France aura naturellement à tenir compte de ces évolutions.

II.4. Pour une politique de la recherche et des moyens sur le long terme

La question d'une véritable structure d'archivage et de diffusion des données pour la recherche en sciences sociales, de son insertion dans les réseaux européens et internationaux est à l'ordre du jour en France. Elle pose en même temps celle de la formation à l'utilisation des données et celle de la place des chercheurs dans la production de

Sur le plan des institutions d'archivage et de diffusion des données pour les sciences sociales comme sur le plan du débat, le constat assez général est que la France est en retard, et du coup peu à même de s'insérer dans les réseaux en train de se mettre en place. Il existe cependant des jalons posés. Par certains aspects aussi le retard français a pu se traduire positivement. C'est le cas par exemple des coopérations entre les chercheurs, l'Insee et quelques grandes administrations publiques produisant des données. La BDSP et le Lasmas-IdL créés au CNRS à quelques années de distance ont apporté des éléments de réponse. Il importe aujourd'hui que ce qui a été, comme dans d'autres pays, fait à l'initiative de chercheurs et de laboratoires de recherche soit relayé au niveau d'une véritable politique de la recherche permettant de résoudre les difficultés et de disposer des moyens nécessaires pour pérenniser ce qui a été construit. La question d'une véritable structure d'archivage et de diffusion des données pour la recherche en sciences sociales, de son insertion dans les réseaux européens et internationaux est à l'ordre du jour en France. Elle pose en même temps celle de la formation à l'utilisation des données et celle de la place des chercheurs

données.

dans la production de données, dans un contexte international et européen en forte évolution. Il faut utiliser le retard français pour prendre en compte d'emblée l'ensemble des problèmes qui se posent et des tendances qui se dessinent, tout en gardant les aspects originaux et riches de potentialités qui tiennent à la proximité des statisticiens de l'Institut national de statistique et des chercheurs.

III. Les principes d'une mise en œuvre du partage des données dans le contexte français

Les besoins pour la France identifiés à partir de cet état des lieux sont au nombre de trois.

1) Accroître la diffusion des données, ce qui pose deux types de problèmes, des questions d'ordre déontologique et juridique (propriété des données) et des questions d'ordre plus technique mais qu'il ne faut pas séparer de la recherche : archivage pour l'utilisation, documentation et outils de diffusion.

2) Accroître l'utilisation des données, ce qui passe par l'amélioration de la formation. La formation à l'utilisation des données est aussi l'une des réponses aux garanties de bonnes pratiques demandées par les producteurs de données.

3) Mieux associer les chercheurs à la production des données, ce qui passe par une meilleure organisation du milieu, sa reconnaissance à un niveau plus institutionnel et un financement spécifique complémentaire pour la production et la coproduction des enquêtes.

Ces besoins appellent des réponses d'ordre et de niveaux différents. Avant de faire des propositions sur ces points, il faut d'abord examiner les principes de traitement dans le contexte français des différents problèmes que toute proposition d'organisation et de structure rencontrera inévitablement. Il s'agit ici, en retenant l'expérience étrangère et en particulier de quelques pays européens, de prendre en compte les contours particuliers de la situation française tant institutionnelle que juridique, ses faiblesses mais aussi ses points forts, et d'anticiper sur les évolutions à venir.

On s'est appuyé ici sur la réflexion des groupes de travail (voir liste en annexe) qui ont associé les différents partenaires concernés. On se situe dans l'hypothèse d'une structure d'archivage et de diffusion des données pour les chercheurs dont les jalons existent déjà mais qu'il faut rendre plus efficace. Il s'agit de définir des principes et d'en tirer les implications que toute structure, quels qu'en soient les contours (examinés au chapitre suivant), devra prendre en compte.

III.1. Les principes d'une mise à disposition des données

a) Archivage historique ou archivage vivant ?

Il convient de distinguer archivage historique dont la mission principale est la conservation et archivage vivant centré sur la diffusion et le partage des données et qui crée de la valeur ajoutée (en particulier sur la documentation).

Le projet doit avoir une ***visée opérationnelle*** pour des utilisations sinon programmables immédiatement, du moins envisageables de façon réaliste et probable. L'objectif premier est de faciliter des travaux approfondis d'analyse de données existantes convenablement archivées et documentées.

On ne se situe donc pas dans le cadre d'une mission d'archivage au sens du dépôt légal ou des Archives nationales. Une coordination avec ce que fait la section Archives contemporaines des Archives nationales est, d'évidence, nécessaire, mais la séparation des objectifs est claire. Pour mieux comprendre ce partage des rôles mais aussi cette nécessaire articulation, il convient de distinguer un archivage « historique » pour le temps long et un archivage « vivant » pour l'utilisation. Cette distinction permet de souligner que, si la sauvegarde sur le temps long est essentielle (patrimoine de la recherche scientifique), l'archivage dans une finalité d'utilisation plus immédiate relève d'une autre logique, celle d'une mise à disposition plus rapide et à destination de la recherche.

Cette utilisation immédiate (moyennant réserve d'une primeur pour le producteur) n'est d'ailleurs pas l'objet de la loi sur les Archives, où il s'agit essentiellement de sauvegarder dans une visée de long terme ce dont les détenteurs ont besoin. Il faut remarquer que l'actuel avant-projet de la loi Informatique et Libertés en application de la Directive européenne en reste à cette logique pour l'autorisation d'archivage des données à caractère personnel. L'archivage pour une finalité de recherche immédiate n'est pas complètement prise en compte et il faut espérer qu'elle le sera comme dans la directive européenne.

Par ailleurs, cette logique de mise à disposition plus rapide des données repose sur une véritable politique scientifique de l'archivage : il ne s'agit pas de tout archiver et il faut des mécanismes de choix. Ceci implique l'existence d'un conseil scientifique et de conseils d'utilisateurs par discipline et par domaine, mais également une démarche active et permanente de recensement des fichiers existants. Cette démarche peut s'appuyer en France sur quelques canaux comme le Cnis et bien entendu les Archives contemporaines. On trouve de nombreux exemples à l'étranger qui montrent le bon fonctionnement de ces conseils et la clarté de leur rôle par rapport à celui des services des Archives nationales dans le cadre d'une coopération. La mission a pu ainsi examiner l'articulation du Département électronique des Archives nationales américaines et des centres tels que le *Roper Centre* et l'ICPSR. Cet exemple est intéressant car il se situe dans un cadre très libéral d'ouverture systématique des fichiers d'enquêtes issus des services fédéraux, et d'une politique d'archivage nationale dotée de moyens

importants pour l'investigation et le recueil de ce type de fichiers. Il apparaît cependant qu'il n'existe aucune concurrence entre les centres qui ont au contraire établi un système de référencement des uns sur les autres.

Clairement les services de l'archivage national ne peuvent répondre à une demande massive pour ce type de fichiers, même si sur le principe ils doivent être ouverts. Sur ce point, il faut noter que l'évolution vers une ouverture plus rapide en France des fichiers soumis à délais, en application de la Directive européenne, va déjà accroître la demande en direction des Archives. En ce qui concerne les grandes enquêtes, les utilisateurs ont des besoins précis qui nécessitent une aide appropriée à l'utilisation des données à des fins de recherche ; c'est à cela que répondent les centres d'archivages pour la recherche. Enfin les *Data Archives* jouent un rôle de négociation avec les producteurs de données afin d'obtenir les données pour les utilisateurs, en assurer une exploitation dans le respect de la déontologie en vigueur ; ils organisent un retour vers les producteurs, toutes activités qui ne sont pas du ressort quotidien des Archives nationales. Par contre la coopération entre ces deux structures d'archivage permet notamment l'harmonisation des outils de documentation et assure une bonne sauvegarde des données.

b) Champ des données

Le champ qu'il s'agit de consolider dans le cadre de cette mission est celui des fichiers de grandes enquêtes permettant le retraitement statistique de données individuelles.

Compte tenu de la croissance exponentielle des données susceptibles d'intéresser les chercheurs en sciences sociales, la question du champ concerné par la consolidation en France d'une structure de diffusion des données est à l'évidence posée. Elle doit l'être par référence d'une part aux jalons existants, particuliers dans chaque pays, d'autre part aux besoins prioritaires. Les données qu'utilisent les chercheurs en sciences sociales sont extrêmement diverses. Il existe des politiques pour certains ensembles de données, par exemple les données d'archives utilisées par les historiens, les matériaux pour les archéologues ou les images utilisées par plusieurs disciplines dont les géographes. Il y a par ailleurs nombre de bases de données gérant les statistiques publiées (agrégats) qu'utilisent notamment les économistes.

Le champ qu'il s'agit de consolider dans le cadre de cette mission est celui des fichiers de grandes enquêtes permettant le retraitement statistique de données individuelles. Les données à recueillir seront très majoritairement des données individuelles sur des personnes, des entreprises ou, plus rarement, d'autres unités statistiques. En particulier, il apparaît souhaitable que des données spatialisées à un niveau fin puissent être concernées. La question que posent actuellement les données du recensement est traitée plus loin de façon plus détaillée, mais il est clair qu'a priori ces données sont bien dans le champ. Le Lasmas archive et diffuse d'ailleurs déjà les données de quelques recensements antérieurs.

La pratique actuelle d'autorisation d'une enquête par la Cnil pour un propos déclaré et précis rend problématique son archivage dans un centre de diffusion des données et sa réutilisation pour un autre objectif scientifique. Ces conditions restrictives sont en contradiction avec l'éthique scientifique de la réplication possible qui doit être au principe d'un tel centre (Voir ci-dessous). Cette question ne devrait en principe plus se poser avec les modifications en cours de la loi de 1978, en application de la Directive européenne.

Les données d'origine administrative sont de deux types. La cession des fichiers d'enquêtes ne pose pas de problèmes particuliers. L'Insee et le Céreq déposent déjà certaines enquêtes au Lasmas-IdL. La DEP (devenue DPD) du Ministère de l'Éducation Nationale s'était un temps inscrite dans cette optique. Il est éminemment souhaitable que d'autres conventions puissent être passées. En ce qui concerne les données administratives, il est souhaitable que certaines, qui peuvent être particulièrement riches pour la recherche, au besoin rendues anonymes pour les utilisateurs, puissent devenir des données statistiques (par exemple, le fichier historique de l'ANPE). Ces fichiers peuvent également servir de bases d'échantillonnage. Cette possibilité de fourniture d'échantillon est prévue dans la recommandation du Conseil de l'Europe sur les statistiques. Une structure d'archivage pourrait assurer en ce sens un service d'échantillonnage pour la recherche.

Une attention devra être portée à tout un ensemble mal définissable de banques de données spécialisées dont la mise en commun pourrait être profitable. Il conviendra de prospecter notamment auprès des régions et de divers organismes qui ont des données urbaines. S'agissant de centres organisés autour de données spécifiques et qui, tout en étant pas sensibles au sens de la loi, peuvent susciter des inquiétudes particulières, comme les données fiscales utilisées par les économistes, il n'apparaît pas utile de retenir l'idée d'une centralisation dont toute l'évolution actuelle des *Data Archives* montre qu'elle n'est plus tenable. Le référencement dans un centre généraliste est par contre utile pour des centres plus spécifiques.

Les données d'opinion recueillies par la BDSP sont bien dans le champ généraliste que l'on cherche à prendre en compte. Elles sont, comparées à celles provenant de la statistique publique diffusées par le Lasmas, largement d'origine académique mais aussi privée et commerciale, d'où des restrictions de communication définies contractuellement. Les restrictions ne tiennent cependant pas à la nature privée de production puisqu'elles apparaissent dans d'autres contextes ; le principe du **contrat** doit être la règle partout. Ces contrats seraient avantageusement encadrés par des dispositions d'ordre public, qui peuvent notamment définir des obligations à la charge du bénéficiaire de la cession ou de l'usage, rassurant ainsi les détenteurs quant aux abus que ce bénéficiaire

ferait. La publicité donnée à des codes de bonne conduite et l'adhésion des bénéficiaires à ces codes sont aussi de nature à faciliter les choses.

Les matériaux textuels commencent à être pris en compte par les centres d'archivage dans le monde anglo-saxon et en Allemagne où des études quantitatives avec des logiciels performants se développent, dans l'idée de soumettre à validation et réplication la base de travaux plus qualitatifs. Compte tenu du retard actuel de la France en matière de fichiers d'enquêtes de grande taille, il ne paraît pas raisonnable d'envisager d'emblée cette question, qui devra cependant être soumise à réflexion à moyen terme. Il s'agit au demeurant de données qui posent des problèmes très différents sur tous les plans (anonymisation, documentation, traitement). Il ne faut donc pas préjuger pour l'instant du cadre de traitement de ce type de données.

D'une manière plus générale, il paraît nécessaire d'avoir une position pragmatique en partant de ce qui existe sans pour autant fermer le champ a priori. L'extension du champ est du ressort d'un conseil scientifique de la structure d'archivage et de diffusion des données. On peut imaginer aussi que pourront être archivées également sur proposition du conseil scientifique des traitements de données dans un objectif de validation scientifique.

c) Champ des utilisateurs

Les données archivées sont diffusées à des fins de recherche. L'ouverture aux universitaires est une question centrale. Il faut prévoir pour chaque fichier les conditions juridiques et financières de cession du droit d'usage pour la France et l'Étranger.

La question se pose également de savoir quel doit être le champ des utilisateurs d'une telle structure. A priori on vise une utilisation à des fins de recherche, et c'est bien cette finalité qui est à l'origine du dépôt des données par les producteurs. À regarder les pratiques des *Data Archives* à l'étranger, qui ont bien cette même visée, on observe cependant que dans certains cas il existe une diffusion à des fins commerciales. Dans ce cas, l'organisme de diffusion n'est que le relais des producteurs de données, exerce pour eux leurs droits et répercute leurs conditions de cession des données. Sauf exception demandée par les producteurs et qui devrait être examinée au cas par cas, ce n'est pas l'objectif visé ici. Lorsqu'il s'agit de données à caractère individuel protégées dans le cadre du respect de la vie privée, et dont l'accès est autorisé à des fins de recherche, toute autre diffusion doit être exclue. Lorsque la question sous-jacente est celle de l'intérêt économique du producteur de données, avec lequel la structure d'archivage ne doit pas entrer en concurrence, une question souvent complexe est posée par la multiplication des travaux de recherche à finalités mixtes. Parmi ceux-ci on trouve souvent des organismes de gestion locale et des aménageurs urbains. L'Insee par exemple n'applique pas le tarif réduit des chercheurs à ce type d'utilisateurs, mais elle n'est pas, de son propre avis, toujours en mesure de distinguer clairement l'utilisateur final.

Tout laisse penser que ces utilisations mixtes vont se multiplier et qu'il faudra nécessairement les prévoir.

L'autre question concerne le périmètre même des utilisateurs définis comme chercheurs. L'actuelle convention Insee-CNRS gérée par le Lasmas limite le champ des utilisateurs aux laboratoires du CNRS. Il est clair que ce qui compte est la finalité de la recherche. On vise donc l'ensemble des chercheurs, indépendamment de leur statut et de leur appartenance. À ce titre, ***il faut absolument pouvoir y inclure les universitaires***, dont on ne voit pas pourquoi sur le principe ils seraient traités différemment des chercheurs du CNRS. Leur demande va croissant. Il faut également prendre en compte la demande des étudiants dès la maîtrise si l'on veut susciter leur intérêt pour les données. C'est un point tout à fait central. L'exclusion actuelle des universitaires du champ de la convention avec l'Insee renvoie en fait à deux problèmes différents.

La question de la contrepartie financière est la plus facile à résoudre. Dans l'état actuel des choses, la contrepartie financière de la diffusion aux laboratoires du CNRS, considérés comme un seul site, est assurée par le département SHS (Sciences humaines et sociales) du CNRS. Les universitaires non associés à des laboratoires du CNRS assument seuls le coût de cession des données le plus souvent hors de leur portée. Il est clair que leur inclusion dans une convention générale doit impliquer une participation financière globale des universités dans la mesure où l'Insee pourrait voir dans une extension de la convention sans contrepartie financière un manque à gagner.

Toute autre est la question des garanties de bonnes pratiques (responsabilité sur la sécurité des données, respect des engagements contractuels, de la déontologie). La question posée par l'Insee est celle de la définition d'un chercheur dans le cas des universités. Dans le cas des laboratoires du CNRS, la responsabilité est clairement assumée par l'organisme. Il s'agit de définir qui, pour les universités, a vocation juridique à assumer cette responsabilité, et si un organisme de diffusion peut le faire et dans quelles conditions. Cette question est examinée plus loin, mais il est clair qu'elle doit trouver une solution. Dans tous les pays c'est l'ensemble des chercheurs qui accèdent aux données. Le rapprochement des universités et de la recherche rend d'autant plus urgent une solution. Le passage d'une gestion de la diffusion des données par des laboratoires du CNRS, à un organisme disposant d'un conseil scientifique et d'un pouvoir de contrôle est un chaînon indispensable pour résoudre cette question. C'est clairement une fonction centrale des *Data Archives* à l'étranger.

La question de l'accès pour les chercheurs étrangers, en particulier ceux de l'Union européenne, est d'ores et déjà posée et devra être prise en compte. Dans l'espace européen, la question de la protection des

données privées étant réglée, celle qui se pose est désormais qu'il faut pouvoir considérer l'organisme de diffusion des données comme garant de la responsabilité quant à une utilisation par des chercheurs étrangers. L'autre point soulevé est à nouveau le risque de concurrence déloyale qui pourrait résulter d'une mise à disposition gratuite à des clients potentiels, pour l'organisme producteur. Il n'est pas impossible d'imaginer une tarification modique rétrocédée à l'Insee à l'image de ce que pratique l'ESRC-*Data Archive* d'Essex. Cette tarification modique devrait en effet tenir compte de la nécessité pour assurer une bonne utilisation des données par un chercheur étranger, de fournir des méta-données, travail assuré par le centre d'archivage[18].

Une dernière question est celle des actuels instituts de recherche, souvent EPST, Ined, Inra, etc. Ils ont inégalement signé des conventions avec les producteurs de données, en particulier avec l'Insee. Lorsque ces conventions existent et qu'ils les jugent favorables, ils souhaitent les maintenir. Lorsque ce n'est pas le cas ils souhaitent pouvoir bénéficier d'un accès aux données via un organisme de diffusion des données pour la recherche. On peut penser que l'existence d'une telle structure, qui archivera des données d'origines très diverses et produira de la valeur ajoutée sur les données, en termes de documentation et de variables nouvelles, intéressera nécessairement les instituts de recherche à très court terme. Il faut donc prévoir d'emblée cette possibilité, somme toute conforme à la visée fondamentale. On peut penser que cela se traduira par une implication des Ministères de tutelle au niveau des moyens, et une négociation avec les producteurs de données, pour ceux qui ne mettent pas les données à disposition gratuitement, prenant en compte ces utilisateurs.

Par contre l'idée d'une mise à disposition étendue aux autres organismes producteurs de la statistique publique doit, sous toute réserve, être exclue. La recherche bénéficie aux yeux des producteurs de données, quels qu'ils soient, d'un statut de neutralité qui doit être préservé. Au reste la mise à disposition des données à des fins statistiques entre services de l'État, qui est de son seul ressort sous le contrôle de la Cnil, est prévue par la loi du 23 décembre 1986 (portant création d'un article 7^{bis} de la loi de 1951). Cette loi rend possible la transmission des données des administrations versus l'Insee et les services statistiques des ministères.

d) Obligation de dépôt ou incitation ?

Le dépôt légal des fichiers d'enquêtes, qu'il s'agisse de ceux issus de la statistique publique ou du monde universitaire, existe dans quelques pays. C'est une idée a priori attrayante, surtout lorsque les chercheurs se

[18]. Le Lasmas-IdL répond souvent aux demandes de documentation ou d'explications de chercheurs étrangers qui ont par ailleurs acheté leurs données.

Plus que l'obligation, c'est l'incitation qui semble être la bonne formule pour obtenir le dépôt de leurs données par les producteurs.

trouvent confrontés à une situation de fermeture forte. L'examen de la situation en France et surtout des exemples à l'étranger montre qu'en réalité c'est l'incitation qui est le véritable moteur du dépôt. Le dépôt implique en effet que les données soient documentées pour pouvoir être utilisées, et l'effet de l'obligation paraît ici très limité. La proposition est donc d'exclure pour la France toute idée d'obligation de déposer, de même que toute idée d'obligation pour la structure envisagée d'avoir à accepter et maintenir n'importe quel fichier de données. L'objectif reste cependant clairement d'obtenir le dépôt des données.

L'obligation devrait cependant être le cas pour des données collectées sur fonds publics, pour des finalités de recherche qui devraient être systématiquement rendues par les chercheurs qui les ont créées, archivées et mises à disposition, pour une éventuelle analyse secondaire. Cette position paraît raisonnable et est pratiquée par l'ensemble des pays disposant de mécanismes de financement d'enquêtes universitaires. Ce point soulève cependant une difficulté pour les enquêtes qui doivent être autorisées par la Cnil (données à caractère personnel), dans le cadre de la loi de 1978, dans la mesure où actuellement l'autorisation n'est accordée que pour une recherche définie. Ce point est examiné plus loin.

L'idée centrale à retenir est que la création d'un archivage opérationnel devrait avoir un effet incitatif à terme grâce aux services offerts et à l'amélioration démonstrative des exploitations. C'est bien l'idée qui a présidé à la politique de conventionnement avec l'Insee et le Céreq mise en œuvre par le Lasmas ou par la BDSP avec les chercheurs et elle a effectivement entraîné un intérêt de la part des producteurs de données. La question de la visibilité nationale d'un tel centre est importante. En l'absence d'un archivage connu, l'accès à certaines données est possible, mais au prix d'un parcours mal balisé et donc décourageant a priori.

Retenir le principe de l'incitation implique que le centre ne peut pas se contenter de ne faire que de l'archivage, mais doit être producteur de valeur ajoutée, qui prend plusieurs formes examinées plus loin (documentation, incitation à l'exploitation, responsabilité déontologique et retour vers les producteurs).

e) Gestion du droit d'usage

Le centre d'archivage doit être en mesure d'assurer la sécurité des données et le respect des règles concernant le droit d'usage à des fins de recherche.

Dans tous les cas un tel organisme assure la gestion du droit d'usage des données. Ce droit est d'abord encadré par les textes et la jurisprudence générale sur la propriété intellectuelle. On trouvera en annexe une revue détaillée des problèmes posés par l'application de ce droit à des bases de données. Les bases législatives concernant les données qui nous intéressent sont limitées. La loi du 7 juin 1951 relative à la seule statistique publique dispose, d'une part, que les données recueillies dans ce cadre et relatives à la vie personnelle et familiale ne peuvent être

communiquées et, d'autre part, que celles d'ordre économique ne doivent pas servir au contrôle fiscal ou à la répression économique ; à cela près, elles peuvent être transmises. Un ajout de 1986 à cette loi autorise la transmission des fichiers des administrations à l'Insee et aux services statistiques de l'administration ; strictement, les instituts de recherche publics ne peuvent bénéficier de cette disposition. La loi du 8 janvier 1978, dite « Informatique et Libertés », ne prévoit pas, on l'a déjà souligné, de régime particulier pour la statistique ni la recherche. Le « principe de finalité » développé par la Cnil a entravé les transmissions aux fins de recherche. La loi « bioéthique » de 1994 a cependant introduit dans la loi de 1978 un chapitre qui permet à la recherche en santé de mobiliser les données correspondantes. On doit aussi signaler que la convention n° 108 du Conseil de l'Europe (1981) a introduit la possibilité d'un régime spécifique, que la loi de 1978 ne comportait pas. Cette convention, ratifiée par la France, a un caractère contraignant ; toutefois, les spécificités de la recherche et de la statistique n'y apparaissent que comme faculté ouverte aux pays membres de déroger aux règles générales ; or la France n'a pas amendé dans ce sens sa législation. Néanmoins, pour l'application de cette convention à la statistique et à la recherche, le Conseil de l'Europe a adopté une recommandation (non contraignante) en 1993, qui vient d'être amendée et complétée pour la statistique en 1997 : on y trouve des propositions tout à fait appropriées, dont l'évolution du droit positif comme les pratiques pourraient s'inspirer. Quant à la loi d'archives de 1979, elle prévoit une possibilité d'accès des chercheurs pendant la période durant laquelle les données ne sont pas encore publiques, mais rien n'y distingue l'accès nominatif du traitement anonyme d'un ensemble de données personnelles. On relève du reste une opposition de logiques : la loi Informatique et Libertés demande que les données ne soient conservées qu'autant qu'elles sont nécessaires à la finalité de leur collecte (ensuite elles doivent être détruites ou anonymisées) tandis que la loi d'archives demande qu'elles soient conservées indéfiniment et que, passé un certain délai (30 ou 100 ans), elles soient librement accessibles. Quant au droit d'auteur, nous avons vu qu'il n'est pas bien défini et qu'il ne fonde en particulier pas pleinement les conditions de rétribution d'un usage partagé des données. Seule la « circulaire Balladur » organise ceci, mais pour les seules données administratives.

Ce cadre juridique, on le voit, est lacunaire et pas totalement cohérent. Des retouches ont été apportées au fil des années, encore insuffisantes : la transposition de la Directive européenne va être l'occasion d'une mise en ordre et d'une meilleure prise en considération des particularités de la recherche, de ses besoins mais aussi des garanties que constitue pour la protection des données la nature même du travail scientifique.

Il n'entrait pas dans le cadre de ce rapport d'examiner dans le détail ces dispositions qui font l'objet actuellement de très nombreux débats et travaux. Dans le cadre de cette mission, la liaison avec le groupe

Déontologie de la Société française de statistique, qui intervient activement pour que la finalité de recherche et de statistique soit prise en compte, a été bien assurée. Il est clair cependant que le degré auquel la loi prendra en compte la finalité de recherche est d'une importance cruciale pour un organisme de diffusion des données dont une partie est à caractère personnel.

Propriété des données ou droit de les gérer ?

La loi de 1951 sur laquelle s'appuient les enquêtes Insee est muette sur la propriété. L'Insee fait comme s'il était gérant de cette propriété pour la collectivité nationale. Les difficultés apparaissent avec les enquêtes cofinancées avec d'autres partenaires qui se développent de plus en plus. Assez souvent, les cofinanceurs publics sont très réticents, pour diverses raisons (exclusivité des traitements, peur d'analyses publiées gênantes ou simplement affirmation du droit supposé de propriété).

Parler de droit de propriété est une mauvaise façon d'aborder le problème. Il faut plutôt s'intéresser à trois questions :
– Qui a le droit de décider de l'usage ou de la cession d'une source statistique ?
– En vertu de quoi : le statut de l'organisme, la souveraineté publique, la propriété commerciale privée ? le fait d'être le premier collecteur ? le fait d'avoir payé en tout ou en partie ?
– Doit-il y avoir une contrepartie à la cession ? immédiate ? différée ? en argent ?

Un point est clair. En fait, la question de la propriété, s'agissant en particulier de celles des données publiques, n'intéresse pas un organisme de diffusion des données. Ce qu'il faut, c'est une doctrine de la cession d'usage, des droits et des obligations afférentes.

Coûts de mise à disposition

L'Insee, sans qu'il soit pour autant nécessaire de le considérer comme propriétaire des données, dispose de ce droit de cession. Il considère que le coût de mise à disposition justifie un paiement du demandeur. Ce coût comprend une quote-part d'un coût général de mise à disposition pour l'ensemble des utilisateurs et peut inclure un coût supplémentaire induit par une utilisation particulière. Dans ce cas, l'Insee établit des devis. Les principes de tarification relèvent actuellement d'un décret relatif à l'Insee (Décret 95-171, 17 février 1995). L'Insee pratique un tarif recherche, et il faut noter que l'intérêt de la convention Insee-CNRS est d'avoir considéré le CNRS comme un seul site.

L'Ined comme le Céreq ont une approche différente, ils considèrent que les données sont un bien public et ne facturent pas les cessions qu'ils

pratiquent libéralement. À titre de comparaison encore, dans un domaine différent, l'IGN pratique des tarifs jugés extrêmement élevés par les chercheurs.

La question du coût de la mise à disposition des données publiques est en cours de réexamen en application des directives européennes. Ceci va en fait se traduire par la redéfinition dans tous les services de l'État du périmètre des données mises gratuitement à disposition du public, notamment par les sites Web (cf. les Actes de la rencontre du Cnis du 28 septembre 1998, L'avenir de la diffusion de l'information statis-tique : impact des nouvelles technologies de l'information et de la communication. Cf. également sur ce sujet le rapport Mandelkern). La plus grande partie des débats a été consacrée à la demande des acteurs économiques et il est vraisemblable que ceci touchera peu le domaine des fichiers d'enquêtes dont il est question ici. Du point de vue des services producteurs de données publiques, la question est celle de l'impact des modifications éventuelles sur leur budget. Celui-ci est très inégal selon l'ampleur de la diffusion des données et les pratiques en vigueur pour l'instant. À titre d'exemple, l'ensemble de la diffusion au public contribue pour environ 5 % au budget de l'Insee, ce qui n'est pas négligeable compte tenu de son budget global. La moitié de ces recettes provient de l'accès au fichier des entreprises Sirène. L'estimation du montant des recettes de diffusion en direction de la recherche est de 1,1 million, sans que l'on puisse distinguer ce qui provient de l'achat de fichiers, de l'accès à Sirène ou de travaux à façon (tableaux ad hoc). Dans ce montant figure l'achat des fichiers pour le CNRS par le Lasmas pour un montant annuel qui s'est progressivement élevé à 160 KF, sauf cas exceptionnel (500 KF pour achat des recensements).

Une première conclusion s'impose : l'organisme d'archivage devra tenir compte des pratiques, différentes, des déposants. Ceci n'est guère différent de ce qui se passe à l'étranger, qui apparaît très variable. Dans le cas de la France, si l'on souhaite maintenir la répercussion d'un coût de mise à disposition des données publiques, notamment pour les organismes de recherche, il importe de prévoir le financement pour la recherche de l'acquisition des données. La question de savoir dans quelle mesure le coût doit être entièrement assuré par l'organisme de diffusion ou répercuté sur les utilisateurs finaux se pose. Au *Data Archive* d'Essex et au *Zentralarchiv* de Cologne, les données sont gratuites pour les utilisateurs et leur coût est pris en charge par l'organisme de tutelle. Un dispositif différent prévaut à l'ICPSR qui est un club d'utilisateurs où les moyens sont apportés par chacune des universités, moyennant gratuité de l'ensemble des données.

Données à caractère personnel et protection des données

En ce qui concerne les données à caractère personnel, le cadre juridique général visant à protéger de retombées dommageables pour l'individu,

Le principe qui consiste à n'autoriser la constitution d'un fichier que pour un usage précis est contradictoire avec le principe de réplication des études et de partage des données. Il est par ailleurs impossible à mettre en œuvre.

la détention et l'usage d'informations à caractère privé, autorise le recueil à des fins administratives ou de gestion (par les opérateurs économiques), autorise le traitement à des fins statistiques par les détenteurs de ces fichiers, interdit l'accès à des tiers, et donne un droit d'accès individuel et de rectification à la personne concernée. Il règle également la conservation sur le long terme de tels fichiers.

De tels fichiers peuvent constituer une source intéressante pour la recherche, ils peuvent également être utilisés comme base d'échantillonnage. Par ailleurs dans le cas de panels constitués par les chercheurs, la conservation sur des temps longs de données à caractère personnel est indispensable. La prise en compte de la finalité de recherche dans la loi est ici cruciale. Elle peut s'accompagner de conditions d'assermentation des chercheurs, via un organisme d'archivage pour la recherche, qui mettrait les chercheurs dans les mêmes conditions que les statisticiens assermentés dans le cadre des administrations. On peut remarquer que la pratique actuelle de la Cnil tend en fait à reporter sur l'administration concernée la responsabilité de l'accès des chercheurs à ses fichiers.

Les deux problèmes liés que posaient ce type de fichiers devraient être résolus par l'application en France de la Directive européenne.
1° L'impossibilité d'archiver, pour être réutilisée, une collecte autorisée pour un objectif précis, déclarée à la Cnil, posait une difficulté sérieuse. Elle enveloppe en effet implicitement une subdivision potentielle entre usages scientifiques, ceux qui sont autorisés et ceux qui, tout autant scientifiques et sur les mêmes données, pourraient ne pas l'être.
2° De nombreuses conventions incluant la cession d'un fichier comprennent une clause de « restitution » et de « destruction de fichiers ». Si les fichiers sont restitués, il n'est plus question d'archivage pour la recherche et, s'ils sont détruits, il n'est plus question de réplication éventuelle d'analyses ou de contrôle de validité scientifique. Ces clauses apparaissent comme largement dépourvues de sens et, au demeurant, invérifiables. Les deux vraies questions sont celle des possibilités techniques de protection de l'archivage et celle de la déontologie des chercheurs, donc de savoir qui les cautionne ou comment on leur fait confiance.

Dans le champ concerné, les données d'entreprises posent des problèmes spécifiques. Les enquêtes contiennent des informations pouvant intéresser la concurrence et l'anonymat n'est pas une protection suffisante dès lors que, de par la taille, la spécialisation ou la localisation, l'identification est possible, sinon facile.

Il convient cependant de noter que le caractère délicat de ces données se périme vite et donc que le traitement, après un délai à fixer, ne devrait pas rencontrer cet obstacle. Cinq ans semblent raisonnablement plaidables. On peut imaginer que, passé un délai à examiner, certaines

données particulièrement intéressantes pour la recherche pourraient être déposées, à charge pour l'organisme de diffusion d'en gérer l'accès sous les conditions prévues. S'agissant des données pouvant poser de vraies difficultés, l'exigence d'accord préalable du Comité du secret du Cnis n'est pas critiquée.

Un organisme d'archivage doit donc être en mesure d'assurer une protection convenable des fichiers qui y sont placés en dépôt (voir plus loin). On ne saurait trop souligner l'importance de ce point et le coût en établissement et en maintien de procédures. Dans le système britannique, les chercheurs qui reçoivent des données du *Data Archive* s'engagent soit à assurer la protection des fichiers soit à les confier tous au Centre.

Le cas particulier du recensement

Des zones de secret doivent être créées pour l'accès aux données qui permettent d'identifier des personnes.

Les contraintes fortes imposées par la Cnil pour les données infra-communales ont suscité de façon très large les protestations des chercheurs qui ont besoin de passer par l'utilisation des données fines du recensement à des fins de reconstruction statistique. Ces contraintes sont d'autant moins acceptées que les aménageurs urbains se sont vus d'emblée reconnaître des droits dérogatoires compte tenu de leurs besoins particuliers, et que des chercheurs sous contrat avec ces aménageurs peuvent accéder ainsi à des données qu'ils ne peuvent utiliser dans le cadre de leurs recherches propres (cf. J.-P. Damais et Y. Guermond dans le Monde du 28 janvier 1999). Un groupe de travail a été mis en place associant l'Insee, le Lasmas et des chercheurs pour examiner comment les chercheurs pourraient accéder à ces données. L'une des solutions passe par la création d'une zone du secret accessible sur accréditation aux chercheurs, comme il en existe au Canada par exemple. Cette zone peut être gérée soit par l'Insee (au Canada, Statistique Canada a implanté des zones dans plusieurs universités, gérées sous son contrôle), soit par une structure de diffusion pour la recherche. Il va de soi que la modification de la loi de 1978 d'une part (prise en compte de la finalité de recherche, possibilité de faire enregistrer à la Cnil des codes professionnels), la consolidation, d'autre part, d'une structure disposant d'un conseil scientifique et d'un tel code sont de nature à favoriser la solution du problème. Elle passe dans tous les cas par une négociation avec la Cnil.

L'avantage pour les producteurs de données, l'Insee dans le cas présent, de confier cette gestion à un organisme de diffusion est de ne pas avoir à gérer les demandes individuelles, inévitablement au cas par cas. S'il en est ainsi, la création d'une zone du secret (éventuellement sous contrôle de l'Insee) implique nécessairement des moyens de sécurisation des données et d'accueil des chercheurs dans une zone particulière, qu'en l'état actuel des choses ni le Lasmas ni la BDSP ne peuvent assurer. La possibilité pour les chercheurs de soumettre des programmes, à charge

pour le centre de vérifier qu'ils ne conduisent pas à des identifications trop fines, trouve aujourd'hui des solutions techniques qui permettent de réduire les temps d'attente. Ce type de solution qui évite l'accès aux données implique cependant des moyens en personnel.

Le schéma selon lequel l'organisme d'archivage répond à des demandes par des cessions de données n'est pas en effet le seul possible et sera, dans un avenir proche, en partie complété par d'autres procédures, notamment la commande de traitements effectués par le centre d'archivage. Le demandeur a accès à un dictionnaire des données archivées concernant son sujet, établit un premier programme de traitements que le centre réalise et lui cède, moyennant facturation. Il n'a pas contact avec les données individuelles.

Dans le cas d'appariement de fichiers (enquêtes sur les comportements patrimoniaux par exemple), la procédure du double aveugle est bien rodée et elle peut être étendue. La question est alors seulement celle de l'accréditation des personnels du centre d'archivage. Elle est plus facile à régler que celle d'une accréditation de tous les demandeurs « recherche » dont les statuts sont variés.

Le Centre d'archivage garant du respect de la déontologie : objectifs et conditions

Toutes les difficultés passées en revue renvoient à l'évidence à deux questions. Sur le plan juridique, il faut que la finalité de recherche soit prise en compte. Mais la contrepartie de cette prise en compte est nécessairement la responsabilisation du milieu de la recherche en sciences sociales. Elle est à la fois évidemment du ressort de chaque chercheur qui est engagé personnellement, mais aussi de l'organisation institutionnelle de cette responsabilité. La création d'un institut d'archivage pour la recherche en sciences sociales doté de structures et de moyens garants de cette responsabilité est un élément important de la professionnalisation du milieu et de la résolution de ces problèmes, comme cela a été le cas à l'étranger. Les différents principes énoncés impliquent quel que soit le projet de Centre retenu et sa structure quatre points tout à fait incontournables :

1° La finalité de l'archivage est scientifique. En conséquence ***le Centre est un organisme scientifique de recherche et de services*** qui a une mission d'interface entre des producteurs de données et des chercheurs de statuts divers. Il a un Conseil scientifique actif. Il inscrit son fonctionnement dans le cadre légal et a des relations définies avec la Cnil.

2° ***Sa pratique respecte la déontologie des communautés scientifiques. Il établit à cet effet un code professionnel***. Ses relations avec les dépositaires et les demandeurs sont définies contractuellement, selon

plusieurs contrats types, dûment approuvés par un Conseil mais dont les formulations doivent pouvoir être révisées sans trop de lourdeur.

Le Centre peut ou non se voir reconnaître le pouvoir d'accréditer les demandeurs ou de se porter garant pour eux, moyennant signature d'un engagement individuel ou de l'institution de référence du chercheur. Actuellement, dans le cas du CNRS, une convention engage le CNRS et le Lasmas, exécutant de cette convention, fait signer un engagement individuel aux chercheurs des laboratoires ou, dans le cas de doctorants, à leur directeur de thèse ou à celui du laboratoire d'accueil. Dans le cas des Universités, un engagement devra être recherché soit via la Conférence des Présidents d'Universités, soit par des conventions générales engageant chaque université, les engagements individuels étant ensuite signés par les universitaires ou les directeurs des laboratoires universitaires, ou à défaut par le directeur de Département. Dans tous les cas, il faut souligner que la responsabilité pénale du chercheur est engagée indépendamment de l'institution de référence, en cas de manquement grave. Un classement des utilisateurs et des données pourra être élaboré par le conseil scientifique, à l'instar de ce qui se pratique dans d'autres *Data Archives*, définissant différentes procédures selon le statut du demandeur et la nature des fichiers.

3° ***Le statut du Centre doit lui permettre de facturer***, selon des tarifs partiellement définis par les exigences de reversement des dépositaires et négociés contractuellement, ou de ne pas facturer dans des cas définis.

4° ***Une procédure doit être définie pour traiter des conflits***, mauvaises pratiques d'un chercheur, plainte d'un dépositaire pour conditions non respectées, etc. La question de la valeur juridique des codes professionnels est actuellement en cours de discussion au niveau européen. Ces codes, dont il existe quelques exemples dans les milieux des sciences sociales (statisticiens, psychologues) à l'image d'autres disciplines (épidémiologistes) ou d'autres professions, n'ont pas de valeur juridique au sens strict. On peut observer cependant que lorsque des cas viennent au pénal, la jurisprudence prend effectivement en compte le manquement à ces codes professionnels. La possibilité de faire enregistrer de tels codes professionnels auprès de la Cnil semble pouvoir être ouverte dans le cadre de la révision de la loi de 1978. Ceci faciliterait certainement le rôle et le fonctionnement d'un centre d'archivage pour la recherche. La stigmatisation dans le milieu et le refus d'accréditation ultérieure pour obtenir des fichiers sont des sanctions qui ont fait leurs preuves à l'étranger.

III.2. Valeur ajoutée par le centre d'archivage

Le caractère opératoire de l'incitation à déposer les données, retenu comme principe, tient en grande partie à la valeur ajoutée par le centre d'archivage et de diffusion. C'est là que réside sa différenciation des Archives nationales.

Une charte constituera le cadre de la transaction entre le Centre d'archivage et les utilisateurs de données.

a) La relation entre les producteurs et les utilisateurs de données

Le rôle d'intermédiaire du Centre entre les producteurs et les utilisateurs est naturellement sa première mission. La mise à disposition des données ne va jamais de soi. Dans le contexte actuel où le souci de protéger la vie privée va grandissant, les administrations pourraient se montrer de plus en plus réticentes. Le Centre devra donc en permanence jouer un rôle de négociation pour obtenir les données, pour obtenir également qu'elles soient documentées, ce qui est long et donc coûteux pour le producteur.

Il ne peut le faire qu'en garantissant que le rôle du producteur sera préservé et reconnu. Il ne faut pas en effet sous-estimer l'effet de territoire. Exploiter des données et publier les résultats est valorisant et les producteurs de données, quels qu'ils soient, ne s'en dessaisissent pas volontiers. Une réponse à cette question peut être apportée par l'instauration d'un délai (l'Insee met désormais à disposition dès la publication d'un *Insee-Première*) pendant lequel les analyses sont réservées aux producteurs et à leurs associés mais ceci ne suffit pas.

Il faut que les utilisateurs de leur côté prennent aussi en compte les producteurs. Une structure de diffusion des données est en position d'intermédiaire et de ce fait peut favoriser la circulation de l'information et de la reconnaissance mutuelle de la valeur ajoutée dans les deux sens. Ce doit donc être l'une de ses préoccupations essentielles. Quand on parle de partager les données statistiques, il faut y inclure les producteurs.

Un engagement des chercheurs utilisant des fichiers de données devra figurer dans une charte des utilisateurs que le Centre élaborera sur le modèle de ceux des *Data Archives* à l'étranger. Il doit :
– faire référence au producteur dans toute utilisation des fichiers,
– faire remonter au Centre d'archivage un exemplaire de toute publication réalisée au moyens de ces données,
– adresser au Centre toute remarque sur le fichier mis à sa disposition, de nature à compléter la documentation sur ce fichier et à en faciliter l'usage,
– communiquer au Centre, dans le cas où il effectuerait une opération de nettoyage d'un fichier, les modalités et une copie du fichier nettoyé,
– apporter son concours pour la préparation des enquêtes sur des thèmes voisins ou sur le même thème.

Une partie de ces propositions figure déjà dans les engagements que font signer le Lasmas et la BDSP.

Inversement le Centre doit jouer un rôle auprès des producteurs de données. Les administrations ont des préoccupations de politique publique et souhaitent prioritairement donner l'accès à leurs données pour en obtenir des indications sur ce plan. Leur demande est celle d'une meilleure interface leur permettant d'accroître leur visibilité du champ de la recherche et de repérer des interlocuteurs potentiels. En sens inverse le centre accroît, pour les chercheurs, la visibilité de l'ensemble des sources accessibles. Les sources de conflit potentiel ne sont cependant pas négligeables entre des administrations qui souhaitent obtenir des conclusions visant à l'action sur des sujets parfois sensibles et des chercheurs, de par les exigences de l'évaluation scientifique, prioritairement soucieux de publications à caractère scientifique. S'il existe un conflit d'intérêt inévitable entre le « savant et le politique », il peut être négocié et le centre peut y aider.

b) Documentation et outils de diffusion

La documentation des données est la principale valeur ajoutée par le centre d'archivage aux matériaux qui lui sont confiés. Il coordonne le recueil des informations venant des producteurs et des utilisateurs.

Pour être archivées, les données, on l'a vu, doivent être déposées dans des conditions minimum d'état afin de pouvoir être utilisées dans les meilleures conditions de validité et d'information sur leurs contenus. Dans le cas contraire, la qualité des données archivées peut s'en trouver fortement affectée.

La question de la documentation des données est, de ce point de vue, fondamentale. C'est l'obstacle le plus important à la diffusion des données par les producteurs. Mais c'est en même temps ce qui peut servir pour un Centre de diffusion des données de valeur d'échange pour inciter le producteur à déposer ses données.

Souvent peu valorisée par les organismes ou les individus producteurs de données (ou peu valorisante pour eux), la documentation est en fait une activité essentielle. Le coût de travail que représente la fabrication d'une bonne documentation est très souvent un élément de frein du côté des agences gouvernementales ou des administrations. Les producteurs sont soumis à des demandes de leur tutelle et à des délais souvent très courts, ce qui entre en contradiction avec l'investissement que représente la préparation des données pour l'exploitation secondaire. L'aide à la documentation est en conséquence une demande que l'on a retrouvée chez tous les producteurs de données (voir état des lieux plus haut).

Du côté des chercheurs individuels, les données sont rarement bien documentées et donc en état d'être partagées. Les chercheurs individuels sont naturellement poussés à ne documenter que selon leurs

propres normes (non réutilisables par d'autres) ou selon leurs intérêts de recherche immédiats. Le centre d'archivage a un rôle important à jouer dans la standardisation des procédures de codage, de classification et de documentation des fichiers, comme le montre l'exemple du Gesis en Allemagne.

Il faut donc que la communauté des chercheurs ne se contente pas d'envisager de récupérer les données mais se sente responsable de leur mise à disposition.

Proximité avec la recherche, partenariat avec les producteurs sont deux conditions propres à favoriser la documentation des données.

À cet égard, il faut sans doute jouer sur deux mécanismes : à la fois inciter fortement à faire cette documentation (une enquête universitaire sur fonds publics de la recherche devrait obligatoirement être archivée et documentée) et mettre en place des mécanismes divers d'aide à la documentation. La structure de diffusion pourrait aider à réaliser cette documentation en favorisant, voire organisant, des mécanismes d'échanges et de mobilités fondés sur la réciprocité avec les organismes producteurs de données. Ces échanges pourraient se traduire en termes de détachements et/ou mises à disposition de personnels. Bien entendu, ces « outils » d'une politique de la documentation sont brossés ici à grands traits généraux. Il faut sans doute retenir surtout le principe sous-jacent, celui d'un échange équilibré entre la structure de diffusion des données et les producteurs de données, en vue d'améliorer la qualité de la documentation des données archivées par cette structure et utilisées par les producteurs. Enfin toute utilisation des données devrait normalement donner lieu à des vérifications complémentaires contribuant ainsi à l'amélioration de leur documentation. On pourrait donc s'attendre, dans le cadre d'un échange fondé sur la réciprocité, à ce que l'utilisateur des données s'engage à faire revenir vers le diffuseur et le producteur les améliorations de documentation auxquelles il aurait procédé. Cette incitation pourrait être plus ou moins fortement suggérée (il pourrait s'agir d'une condition d'accès aux données, voir plus haut charte des utilisateurs).

Comme on le voit, il s'agit, à travers des mécanismes variés et adaptés à des situations spécifiques, de conduire une véritable politique de documentation des données, considérée comme une valeur ajoutée à des enquêtes, souvent financées sur fonds publics, reconnue scientifiquement comme une activité « noble » au service de la communauté scientifique. Cette reconnaissance est une condition importante à remplir si l'on souhaite que les chercheurs s'impliquent davantage dans cet échange réciproque entre une structure d'archivage des données et eux-mêmes.

Dans cet esprit, les chercheurs pourraient consacrer plus de temps à un travail en commun avec les organismes producteurs. La Darés par exemple serait intéressée à accueillir des stagiaires ou des doctorants qui

travailleraient sur ses propres fichiers. On rejoint ici la question de la formation car des bourses sont envisageables, à condition de leur trouver un support institutionnel.

Une condition est essentielle pour que ces mécanismes fonctionnent. Il faut construire une structure d'archivage vivante, ne pas la couper de la recherche. Il ne s'agit pas d'archiver pour archiver. Aider à documenter les données, assurer un retour vers les producteurs, ce qui est le meilleur moyen d'inciter à déposer les données, ne peut se faire qu'avec l'aide des chercheurs, dont toute structure d'archivage devra être proche.

Fichiers élaborés et veille informatique

Une autre question est celle du développement par les centres producteurs de données et donc par les centres d'archivage de la mise à disposition de fichiers « élaborés » et non plus de « fichiers sources ». Cette évolution pose le problème de l'indépendance du fichier des données vis-à-vis des logiciels de traitement des données et des supports informatiques. Or, si l'on ne prend pas garde à cette question et compte tenu de l'évolution des matériels informatiques, de très graves problèmes de « migration » des fichiers vont se poser. Les supports informatiques vont encore évoluer fortement dans les prochaines années et, même si l'on ne doit plus craindre ce qui s'est passé avec les bandes archivées dans les centres de calcul, on peut avoir des doutes sur la transmission intergénérationnelle des fichiers. Une veille doit être assurée sur ce point et une banque de donnée doit être capable de procéder régulièrement à des vérifications de son archivage. La sécurité de celui-ci passe par une politique d'archivage double ou d'archives miroir.

Outils de diffusion

Par ailleurs, des standards d'archivage des données se sont développés (Insee et CAC en partenariat, BDSP, Cessda). Quel que soit le standard retenu, se pose la question de la « documentation rétrospective ». Y a-t-il intérêt à mettre aux normes d'anciens, voire très vieux, fichiers ? Quels coûts cela représente-t-il pour quels avantages ? Il paraît raisonnable de dire que cela relève clairement d'un mécanisme de choix du point de vue de la communauté scientifique. Des standards se sont également développés en termes de supports de diffusion et d'échanges des données. Le support cd-rom et l'échange électronique des données (du type FTP) se sont imposés au cours des années récentes comme les standards les plus reconnus et utilisés. Ce mode de diffusion pose néanmoins le problème de l'envoi des documents papiers. La solution la plus appropriée semble être de procéder au « scanning » de ceux-ci dans une logique « image » plutôt que de reconnaissance de texte. L'utilisateur peut alors très aisément consulter la documentation papier des données. La mise sur le Web de documents plus élaborés (du type

tris à plat ou autres résultats d'enquêtes) constitue également une piste d'avenir. De plus en plus d'utilisateurs sont habitués à « naviguer sur le Web » et à récupérer de tels documents. À terme se posera donc la question du niveau d'élaboration des documents d'enquêtes disponibles sur le Web. Le Lasmas s'engage actuellement dans cette voie afin de faciliter le travail des utilisateurs à la recherche des fichiers les plus adaptés à leur recherche. Pour cela il est évident qu'un centre d'archivage devrait s'engager dans une coopération accrue avec l'Insee d'une part et le réseau européen de *Data Archives*.

Il ressort de toutes ces considérations que :
– il faut instaurer entre les structures de diffusion pour la recherche, les producteurs de données et les institutions d'archivage, un véritable partenariat dont certains éléments existent déjà et sur lesquels on peut s'appuyer pour partir de l'existant ;
– ce développement passe par l'affectation de personnels et l'accès à des réseaux à gros débits.
On trouvera en annexe quelques spécifications technique relatives à ces moyens.

III.3. Formation à l'utilisation des données

Les méthodes quantitatives permettant l'exploitation des enquêtes sont mal connues et peu enseignées.
Le centre d'archivage, comme tous ceux qui existent à l'étranger, doit développer une politique dynamique de formation tant initiale que continue.

La formation est une réponse centrale à la fois en termes de garanties de bonnes pratiques et en termes d'incitation à utiliser les données. Le milieu formé à l'utilisation des données, en particulier en sociologie, est trop étroit. Tous les producteurs de donnés s'en plaignent. En même temps il y a des besoins à la formation sur les enquêtes et des besoins en formation continue aux nouveaux outils.

Les futurs chercheurs reçoivent, de l'avis général, une formation insuffisante pour utiliser les données de grandes enquêtes dans le cadre de l'enseignement initial. L'enseignement statistique est très cantonné au DEUG et sans relation avec une utilisation concrète d'une enquête autour d'une question de recherche. Les Mass qui ont tenté d'orienter vers les sciences sociales des étudiants provenant des filières de mathématique se sont révélés peu opératoires dans la mesure où ces étudiants n'avaient pas une formation initiale en sciences sociales. Sans une réflexion sur l'organisation de l'enseignement statistique dans les sciences sociales, en particulier en sociologie, au plus tard au niveau de la licence, il est vain d'espérer une croissance forte de l'utilisation des données dans les thèses. C'est le point le plus difficile. Il suppose à la fois une politique des départements concernés en ce sens et des moyens à disposition des étudiants. Il s'agit en effet de dispenser un enseignement statistique en situation d'exploitation de données qui nécessite la disposition de micro-ordinateurs en nombre suffisant et

régulièrement renouvelés. La question de la disponibilité de données n'est plus un problème. L'Insee est en train d'élaborer des fichiers simplifiés peu coûteux qui pourraient être utilisés. Ce pourrait être également une tâche d'un centre de diffusion des données.

Il faut donc une gamme de réponses, où la structure de diffusion des données peut jouer un rôle important mais pas unique. Les *Data Archives* jouent partout un rôle de formation, en organisant des écoles d'été très importantes. L'école d'été organisée par l'ICPSR de Michigan, largement ouverte à l'ensemble des utilisateurs, et organisée pour des niveaux de compétences très différents, est la plus connue. Une mission de formation doit être assignée au Centre de diffusion des données. Cette mission doit être assurée par l'existence d'un département assurant une veille et un transfert de compétences en méthodologie, et éventuellement du développement en la matière. Le principe n'est pas celui du travail à façon (sauf exception) par le Centre mais celui de l'aide aux utilisateurs et l'organisation de formations.

D'autres réponses passent par une politique plus générale. Cette période dans laquelle on renégocie les écoles doctorales est favorable aux propositions que peuvent faire les EPST en direction des universités. De très nombreux universitaires sont conscients de la nécessité de relancer une culture scientifique et technique pour les sciences sociales. Même dans les disciplines où l'on fait des choses très pointues, les doctorants manquent souvent d'une culture générale dans le domaine de l'analyse des données quantitatives. Cette question doit être prise en compte dans une politique d'allocations de recherche.

Il convient donc d'intervenir à la fois dans l'enseignement initial et dans la formation continue, dans une logique de formation par la recherche.

III.4. La place des chercheurs dans la production des données

L'attention aux conditions de production des données est pour la recherche en sciences sociales une condition nécessaire de rigueur. Elle passe, on l'a dit, par une implication plus forte des chercheurs dans la production des données. Inversement la statistique publique a intérêt à ce que sur des champs nouveaux, où elle ne peut se mouvoir qu'avec lenteur du fait de l'inévitable lourdeur de son système d'enquêtes, ou sur des sujets sensibles, des enquêtes dont les universitaires sont maîtres d'œuvre puissent avoir lieu. Elle marque régulièrement son intérêt à impliquer les chercheurs en amont de la production de ses propres enquêtes, assimilant ainsi plus rapidement les résultats de recherches et contribuant à développer un milieu plus à même d'utiliser les données.

Enfin dans un contexte de maîtrise des coûts, la politique de coproduction qui se développe du côté de la statistique publique doit naturellement pouvoir inclure la recherche, sous réserve de prévoir les financements nécessaires. La définition très large en France de la notion de statistique publique, fortement liée à celle de données d'intérêt public, constitue un cadre favorable à une implication plus forte de la recherche dans la production de données. On peut y inclure sans difficulté la nécessité de participer aux grandes enquêtes européennes et internationales.

a) Production d'enquêtes universitaires en France

Production directe d'enquêtes à l'initiative des universitaires et des chercheurs, participation à l'élaboration des enquêtes nationales réalisées par les grands instituts ou coproduction et co-financement sont trois voies à explorer.

Parmi les nombreuses formes de réalisations nouvelles possibles, quatre méritent de retenir l'attention.

– Les coproductions, évoquées ci-dessus, sont une bonne occasion de développer des collaborations avec l'Insee ou d'autres producteurs de données publiques, qui y sont par ailleurs ouverts, notamment dans le cas d'enquêtes plus spécifiquement recherche, comme c'est le cas de l'enquête FQP par exemple qui permet d'étudier la mobilité sociale en France (voir Annexe IV).
– La réinterrogation. Les enquêtes sur les grands échantillons abordent plusieurs thèmes mais souvent l'impression est qu'il faudrait aller plus loin sur un sujet précis. Il serait donc intéressant que la recherche puisse réinterroger des sous-échantillons de façon plus approfondie et en collaboration avec l'institut ayant effectué l'interrogation primaire, comme cela a été fait avec l'enquête Conditions de vie de l'Insee (1986). Plusieurs enquêtes (telle l'enquête sur la santé de l'Inserm) ménagent cette possibilité en demandant aux enquêtés s'ils accepteraient d'être réinterrogés. Il existe ainsi des réserves de sous-échantillons.
– Des collaborations plus étroites avec les administrations productrices de données qui les inciteraient vraisemblablement à aller davantage dans le sens de la continuité. Par exemple cela permettrait de suivre le panel d'élèves interrogés en 1995 par le MEN, après leur sortie du système scolaire (voir Annexe IV).
– En termes de production proprement dite, il pourrait être utile que la recherche et l'université lancent une enquête sociale annuelle. Les questions pourraient ainsi toucher des domaines que n'aborde pas l'Insee (religion, politique, valeurs, etc.) et selon des problématiques peu développées en France jusqu'à présent (réseau par exemple).

Enfin, il faut absolument que les chercheurs puissent trouver le financement nécessaire à une participation à des enquêtes européennes ou internationales (ISSP, ESS, par exemple, voir annexe).

Dans tous les cas, il faut que le pilotage, l'archivage et l'accès aux données trouvent leurs cadres institutionnels. La production de données

sociales, en matière d'opinions ou de pratiques, est l'équivalent d'un grand équipement en sciences exactes. Que ce soit pour développer les productions existantes, en créer de nouvelles ou engager des coproductions, il faut mettre en place un dispositif public d'appel d'offre et donc de sélection par un comité scientifique. Ce comité jouerait un peu le rôle que joue la NSF aux USA pour le financement de données. Il apparaît naturel que l'Insee, tant pour des raisons de cohérence du dispositif d'ensemble que de compétences en matière de production d'enquêtes, y soit représentée, à côté de l'université et de la recherche.

Le financement de données par la recherche pose, et posera d'autant plus qu'il s'amplifiera, tout un ensemble de problèmes qu'il ne faut pas mésestimer. L'accès de tout chercheur qui le souhaite à ces données est un principe de déontologie scientifique intangible. Une publication reposant sur des données totalement inaccessibles ne peut en aucun cas être considérée comme scientifique. Une fois ce principe réaffirmé, il appelle des aménagements sous forme de délais de carence (raisonnables), car il faut aussi tenir compte d'un droit d'exploitation des données par les chercheurs qui ont été à la source de leur production.

L'exigence de mise à disposition ne va pas sans une exigence de documentation. Or elle ne peut être imposée aux chercheurs. Il faudrait que le temps passé à la documentation soit reconnu comme un temps scientifique par les instances d'évaluation. Il faut surtout qu'une institution d'archivage soit en position de garantir la qualité de la documentation, en même temps d'ailleurs qu'elle serait garante de la sécurité des données. En sens inverse, il n'est pas nécessaire de tout archiver. Il faut un noyau dur et ensuite procéder par cercles concentriques déterminés par la communauté scientifique elle-même dans l'usage qu'elle fait des données. Le minimum est toutefois le questionnaire et le plan de codage. Il faut rappeler que faute d'avoir jusqu'ici considéré qu'il existe un véritable patrimoine historique des données sociales produites par la recherche, certains « trésors » ont été irrémédiablement perdus (exemple : les enquêtes sur le niveau intellectuel des enfants en âge scolaire, de 1944 et 1965).

Les avantages d'une telle politique sont multiples. Il est évident que l'on gagne en cohérence et en récurrence. Cela devrait sans doute aussi réduire la redondance. Une politique d'archivage (même si les centres sont en fait multiples et mis en réseau) permettrait en amont de mieux focaliser les financements, vraisemblablement comme en Grande-Bretagne ou aux États-Unis, vers des instruments lourds, multidisciplinaires et continus.

Elle permettrait aussi d'améliorer encore la relation entre instituts producteurs de données et recherche. De ce point de vue, il est clair qu'à peu près partout il existe une stricte dichotomie entre données universitaires (ALLBUS, *General Social Survey*, *British Social Attitudes*

Surveys, etc.) et données officielles. L'accès des universitaires à ces dernières est très difficile et la France est ici relativement mieux placée. Mais en même temps partout la tendance est au rapprochement, un peu à l'instar de la situation en Grande-Bretagne où le monde académique a su construire le relais entre les instituts de statistiques officielles et la communauté scientifique (y compris internationale). On peut penser que les évolutions des législations nationales, en Europe, vont permettre d'aller assez vite plus loin dans l'accès aux micro-données officielles. Il faut donc que la France maintienne sa situation favorable tout en facilitant les accès à des publics plus variés : étudiants et chercheurs étrangers notamment.

b) De la présence des chercheurs en amont de la production des données

La présence des chercheurs en amont des enquêtes, sans pour autant qu'ils soient directement impliqués dans leur production, quoiqu'inégale, est un des points positifs qui ressort de la situation française. Il importe de la préserver et de la faire croître. Ce doit nécessairement être l'une des missions du Centre qui doit faire remonter aux producteurs les résultats de recherche des utilisateurs, et jouer un rôle d'information entre les deux milieux, sans pour autant s'assurer un monopole de l'organisation de ces relations. Une coopération avec le Cnis pourrait être envisagée ici.

Dans le contexte d'intégration européenne, la question de la place des chercheurs dans le processus d'harmonisation des systèmes d'enquêtes et des nomenclatures est posée. Là encore le Centre, qui doit accroître ses relations avec les autres centres européens, est bien placé pour jouer un rôle pilote.

En conclusion

Les principes dégagés dans ce chapitre, qui sont de nature à recueillir l'assentiment de la plupart des partenaires concernés, utilisateurs ou producteurs de données, ont été mis en œuvre efficacement à l'étranger. Ils ont reçu un début d'application en France à travers le Lasmas et la BDSP. Pour résoudre les difficultés qui demeurent, on ne peut aller plus loin sans répondre en termes de professionnalisation du milieu, propres à apporter les garanties nécessaires aux producteurs, et de moyens permettant par l'échanges de services d'avoir une véritable politique d'incitation à déposer les données. Ceci ne peut se faire que dans le cadre d'une véritable politique d'instrumentation pour la recherche en sciences sociales.

IV. Propositions

Préambule

Trois objectifs doivent être poursuivis de front, avec des mécanismes distincts : faciliter l'accès aux données, former à l'utilisation des données, rapprocher les chercheurs de la production des données.

La situation de la France en ce qui concerne la relation que les chercheurs en sciences sociales entretiennent avec leurs données, s'agissant ici des fichiers d'enquêtes sur grands échantillons, doit être décrite de façon nuancée. L'attention portée par les chercheurs à leurs données et aux conditions de production de ces données – ce qui est, dans les autres sciences, considéré comme un élément central – est insuffisante. La faible implication des chercheurs dans la production directe des données quantitatives, le manque de formation dans certaines disciplines à l'utilisation des données qui explique aussi une trop faible utilisation de celles-ci, figurent parmi les éléments négatifs de la situation. Par contre, les liens qui ont été tissés entre utilisateurs et producteurs de données publiques, même s'ils sont loin d'être généralisés, sont des aspects positifs qu'on ne retrouve pas toujours à l'étranger et qu'il convient de ne pas perdre. Parmi ces jalons posés, figure le travail effectué par le Lasmas-IdL et la BDSP mais aussi les liens tissés entre les différents instituts de recherche et les chercheurs à des niveaux plus individuels.

Un contexte plus favorable constitue évidemment un atout pour mettre la France au niveau d'autres pays qui ont depuis longtemps des structures plus institutionnalisées et disposent de moyens plus conséquents pour traiter la question de l'accès aux données pour les sciences sociales. De ce point de vue, la mise en réseau de ces structures tant sur le plan européen qu'international est une incitation puissante à mettre en place une politique en la matière. Le fait de considérer les données publiques comme un bien public est éminemment favorable pour la recherche. La Directive européenne de 1995 qui fait place à la recherche, la statistique et l'histoire, si sa traduction dans la loi française se fait complètement, constitue désormais un contexte favorable. Enfin l'émergence d'un débat sur la réplication et de la validation scientifique qui est au cœur du principe du dépôt des données, qu'il s'agisse des données publiques produites par les agences gouvernementales ou des données universitaires, est également un élément très positif.

Une politique d'ensemble doit impérativement associer une politique de partage des données pour la recherche, une politique de formation à l'utilisation des données et une politique de rapprochement des chercheurs de la production des données. Si ces trois objectifs doivent être poursuivis de front, les mécanismes à mettre en œuvre dans ce but ne sont pas nécessairement les mêmes.

Une première question est celle de ***l'archivage et de la diffusion des données***. Il s'agit là de consolider une structure dans les conditions définies ci-dessous : partenariat avec les producteurs de données dans le cadre d'un projet scientifique assignant des missions, préservation d'une position neutre de la recherche, gestion par un conseil scientifique. Il faut ici partir de ce qui a été construit par la BDSP et le Lasmas en engageant dans ce nouveau cadre des moyens assortis éventuellement de ***possibilités de mobilité des personnels entre les partenaires***.

Une seconde question est celle de ***la formation à l'utilisation des données***. Sur ce point, une telle structure peut jouer un rôle moteur, à l'instar de ce qui se fait dans d'autres pays, d'une part en matière de formation ponctuelle sur telle ou telle enquête, mais aussi en matière de formation aux outils d'analyses. Ceci pourrait se faire dans le cadre du partenariat défini plus haut mais en s'appuyant sur les services de la formation permanente du CNRS et des universités. Mais il ne s'agit là que de l'une des pièces d'un dispositif plus large à mettre en place. Il est nécessaire en particulier de définir une politique sur ce point au niveau des écoles doctorales.

Une troisième question est celle de ***la place des chercheurs dans la production des données***. Là encore une structure d'archivage et de diffusion des données, articulée à la recherche, peut jouer un rôle d'organisation du milieu et de lien entre les utilisateurs et les producteurs de données. Mais il faut aussi chercher à systématiser la présence des chercheurs au Cnis, à organiser les relations des chercheurs avec la Cnil, faire une place plus importante dans une politique de la recherche au financement de la production des données. Ceci passe par des mécanismes distincts visant à faire des chercheurs des partenaires à part entière dans la production des enquêtes sur vastes échantillons intéressant les sciences sociales.

IV.1. L'accès aux données

Le champ visé est celui des grandes enquêtes intéressant les sciences sociales (données sociales au sens large) et permettant le retraitement statistique de données individuelles, le plus souvent sur des personnes, mais qui peuvent être également, après un délai variable, des données sur les entreprises. Les données spatialisées à un niveau fin sont concernées. Le dépôt de certains fichiers administratifs, éventuellement anonymisés, pourra être envisagé.

Dix principes à respecter pour organiser l'accès aux données.

Principes retenus

1) En ce qui concerne le partage des données, le principe retenu est celui de la création d'une structure à vocation nationale d'archivage et de diffusion des données pour la recherche en sciences sociales au niveau de celles qui existent à l'étranger depuis plus de 20 ans, appuyée sur le principe de l'incitation forte et non de l'obligation (dépôt légal). Cette incitation s'appuie sur l'intérêt du producteur lui-même : obtenir la reconnaissance de son travail par la citation, accroître l'utilisation des données, obtenir de la valeur ajoutée en termes de documentation, récupérer le travail des utilisateurs en vue de nouvelles enquêtes, sauvegarder éventuellement des données.

Le principe de ce dépôt implique :
– de distinguer dans les différents droits sur les données. Il ne s'agit pas de la propriété des données mais de leur droit d'usage.
– de faire valoir et reconnaître la finalité de recherche.
– d'inciter à définir les conditions et les délais dans lesquels les données seront disponibles, dès le montage des enquêtes, et particulier en cas de coproductions, qu'il s'agisse des données publiques ou des données académiques. En ce qui concerne ces dernières, il est proposé d'inclure au moment du financement l'obligation de déposer au centre d'archivage et de documenter les données relevant de fonds publics de la recherche.
– de reconnaître une valeur ajoutée à la documentation résultant du travail d'une structure d'archivage et/ou des exploitations secondaires, sous condition d'un retour vers le producteur. Ceci implique de définir des obligations de retour vers le producteur, quel qu'il soit.

2) La professionnalisation du milieu apparaît comme la contrepartie et la condition sine qua non du partage des données. À la question posée « qu'est-ce qu'un chercheur ? » il faut répondre par une professionnalisation du milieu qui garantisse les bonnes pratiques en matière d'utilisation des données. Il n'y a pas en effet lieu d'estimer que les statisticiens des agences gouvernementales seraient plus susceptibles de donner des garanties sur ces bonnes pratiques que les chercheurs, si ce n'est parce que les mécanismes de ces garanties ne sont pas aussi institutionnalisés. Ceci a pour conséquence qu'un centre d'archivage doit :
– être doté d'un conseil scientifique où les producteurs de données peuvent être représentés,
– disposer ou adhérer à un code professionnel qui pourrait être enregistré à la Cnil (voir avant-projet de révision de la loi de 1978),
– établir une charte des déposants et des utilisateurs (droits et obligations),
– établir par la voie de son conseil scientifique un classement des données et des utilisateurs qui définisse des procédures distinctes d'accès aux données et des sanctions propres à engendrer la confiance

des producteurs de données et à favoriser le dialogue avec la Cnil. Ceci implique d'identifier pour l'ensemble des chercheurs, CNRS, Universités, Instituts de recherche, les instances pouvant engager dans chaque cas leur responsabilité par la voie de ***conventions générales*** (Direction du CNRS, Conférence des Présidents des Universités, ou Présidents d'Université).
– définir des engagements signés par les utilisateurs et des sanctions possibles (plus de cessions de données, pénalisation possible). Leur responsabilité individuelle est directement engagée, indépendamment des conventions générales, en cas de manquement grave.
De telles procédures sont de nature à être facilitées par la définition ou l'adhésion à des codes professionnels de la part des EPST et des Universités.

3) Une telle structure doit disposer de la ***personnalité juridique*** pour signer des conventions avec les producteurs de données.

4) À moins d'un changement radical de la politique de mise à disposition des données publiques, le coût d'accès aux données pour la recherche doit être pris en compte au niveau de la politique de la recherche, s'agissant du centre d'archivage naturellement mais aussi des EPST ou des Universités pour des données qui ne seraient pas disponibles au centre. ***Une structure d'archivage et de diffusion doit inciter à la mise à disposition à prix coûtant, mais doit aussi pouvoir acquérir les données, si nécessaire*.** Elle peut aussi être habilitée à répercuter les conditions du producteur.

Que les producteurs pratiquent la facturation du coût marginal de mise à disposition ou qu'ils répercutent une part des frais de collecte et d'élaboration, est une question qui a une incidence sur un centre d'archivage qui doit disposer des moyens pour acquérir les données. Mais c'est aussi une question de politique générale des données. Une telle politique pourrait se proposer d'harmoniser les règles et pratiques ; par exemple, de promouvoir le principe de la seule facturation du coût de mise à disposition.

5) Les moyens techniques propres à assurer l'archivage et la sécurisation des données, la diffusion dans de bonnes conditions, l'aide à la documentation des données sont des conditions indispensables. Ceci implique des moyens informatiques, l'accès à des réseaux à gros débits (Renater 2), des espaces sécurisés, des moyens en personnel. À l'autre bout de la chaîne, les utilisateurs, qu'ils soient dans les EPST ou les Universités, doivent eux aussi disposer de moyens informatiques incluant, outre l'équipement initial, sa mise à niveau régulière, son renouvellement et les licences annuelles de logiciels de traitement des données (SAS, SPSS). Ils doivent eux aussi pouvoir accéder à des réseaux à gros débits. C'est donc d'une politique informatique d'ensemble qu'il s'agit.

6) L'objectif est de créer un archivage vivant, centré sur l'utilisation des données. L'intérêt porté aux méthodes d'enquêtes et aux méthodes d'analyse des données fait partie des missions du centre. Dans le même esprit, le centre peut développer des tests sur des méthodes d'enquêtes et des outils d'analyse dans des domaines particuliers de recherche. Il apporte son aide aux utilisateurs sur tous ces points. Il peut être amené, en collaboration avec l'Insee, à aider à la production d'enquêtes universitaires.

7) Le centre sert toutes les disciplines. Il apparaît cependant nécessaire qu'il soit particulièrement actif dans le domaine de la sociologie quantitative, trop peu présente en France. Le développement de ces travaux fait partie intégrante de la construction de relations de confiance avec les producteurs de données publiques. Le centre doit jouer un rôle d'impulsion dans ce domaine, qu'il faut par ailleurs chercher à développer. Il doit avoir un département recherche. Une façon de faire est également que le centre puisse disposer d'allocations (éventuellement dans le cadre d'un partenariat avec les producteurs de données publiques pour encourager doctorants et post-doc à des travaux sur les données) et de possibilités d'accueil de chercheurs français ou étrangers.

.

8) La réponse en termes de formation est une garantie de bonnes pratiques que l'on peut apporter aux producteurs. C'est aussi une réponse à la demande de travaux « à façon » qu'on ne peut exclure mais qui n'est pas l'objectif visé par le centre, qui est de mettre les données à disposition des chercheurs. Le centre doit organiser en partenariat avec les EPST et les producteurs de données une formation régulière autour des nouvelles enquêtes et une formation aux méthodes d'analyses, ouvertes largement aux utilisateurs.

9) Le centre doit permettre aux producteurs de données publiques comme aux chercheurs qui veulent utiliser ces données d'avoir une meilleure visibilité respective de l'ensemble du champ. Il organise par exemple des sessions thématiques autour des enquêtes.

10) La question de la circulation internationale des données, en particulier en Europe, est posée et reste difficile. Ce sera un sujet majeur de réflexion dans les années à venir. Une structure d'archivage et de diffusion des données pour la recherche doit être insérée dans les réseaux européens et internationaux d'archivage des données. Elle doit jouer un rôle dans l'organisation de la réflexion sur ce point. Elle doit au minimum pouvoir disposer de postes d'accueil sur place des chercheurs étrangers.

Propositions

Un institut doté d'un conseil scientifique où sont représentés les producteurs de la statistique publique et doté d'une charte des utilisateurs

Si les missions centrales se retrouvent dans tous les *Data Archives* à l'étranger, leur structure particulière et leur articulation à d'autres activités, en particulier de recherche, est chaque fois particulière et fondée sur l'histoire. On ne peut donc prétendre en la matière copier complètement l'un ou l'autre. Il est en particulier nécessaire de partir de l'existant pour assurer une montée en puissance et de tenir compte des points forts comme des points faibles à développer. Deux points doivent attirer l'attention. La structure de type consortium comme celle de l'ICPSR a montré les difficultés d'une structure d'accès aux données par association des partenaires. Lorsque pour une raison quelconque l'un des partenaires se retire, la question de ses données pose problèmes. La structure du Gesis en Allemagne qui assure l'intégration de plusieurs institutions, le *Zentralarchiv* (Cologne), le *Zuma* (Mannheim) et l'*Informationszentrum* (Bonn), ayant des compétences différentes et collaborant ensemble, pourrait inspirer la France. Les missions à développer doivent cependant tenir compte des spécificités de la situation française. Du fait de la situation allemande de très forte coupure entre la statistique publique et les chercheurs, le *Zuma* s'est beaucoup attaché à la question de la production d'enquêtes universitaires. En France, ce type de compétences devrait être développé en collaboration avec l'Insee. Il est par contre nécessaire en France de promouvoir la recherche dans le domaine de la sociologie quantitative.

En partenariat avec des organismes publics produisant des données et en particulier avec l'Insee, ***il est proposé de créer un Institut*** ayant pour fonction d'archiver et diffuser les données et d'en promouvoir l'utilisation pour les sciences sociales. Dans le cadre de ce partenariat, la position neutre de l'Institut doit être préservée. L'Institut diffuse les données pour la recherche, il n'intervient pas dans la circulation de ces mêmes données entre organismes producteurs de données publiques.

Les objectifs de cette politique à vocation nationale et de création de cet Institut sont définis dans une Convention cadre associant au départ le ministère de l'Éducation nationale, de la Recherche et de la Technologie, le ministère de l'Économie, des Finances et de l'Industrie (pour l'Insee) et le ministère de l'Emploi et de la Solidarité. Ceci permet d'engager notamment la Direction de la Recherche, la Mission Scientifique Universitaire, le CNRS et les Universités, éventuellement les autres EPST ainsi que la Conférence des Présidents d'Universités si nécessaire, et d'autre part l'Insee, le Céreq, la Darés, éventuellement le CEE et la DREES. Cet accord cadre peut naturellement être ouvert à d'autres ministères (en particulier Justice ainsi que Culture et communication).

Les partenaires sont naturellement présents dans le Conseil d'administration qui définit les moyens. Ils sont également représentés dans le Conseil scientifique qui définit la politique d'archivage, les

priorités d'aide à la documentation (appuyée sur les conseils des utilisateurs), la politique de diffusion, et assure le contrôle des bonnes pratiques des utilisateurs.

La structure à créer a vocation à être stable dans le temps. Idéalement il s'agit d'un EPST. Ce peut être provisoirement un GIP, mais cette structure n'a pas vocation à être permanente. On peut également envisager que le conseil d'administration et le Conseil scientifique confient dans un premier temps la mission d'exécution aux deux laboratoires du CNRS, le Lasmas-Institut du Longitudinal et le CIDSP-BDSP qui ont posé les jalons de cette politique.

Les deux laboratoires sont, quelle que soit la structure retenue, le point d'appui pour préfigurer, par une réorganisation et une collaboration de leurs moyens et de leurs personnels, les différents départements de l'Institut autour des missions définies : archivage, documentation, diffusion, méthodes d'enquêtes et méthodes d'analyse des données, gestion des données longitudinales, département recherche. D'autres associations avec des centres plus thématiques et des laboratoires de recherche sont possibles autour de ces missions.

L'Institut (qui doit pour cela disposer de la personnalité juridique) passe des conventions particulières avec les organismes producteurs de données et gère les engagements des utilisateurs, dans le cadre de conventions générales signées par les EPST (Direction du CNRS, Présidents des Universités).

Les différents partenaires définissent les moyens qu'ils apportent. Les moyens provenant de l'Enseignement supérieur et de la Recherche doivent être inscrits dans le cadre d'une politique nationale de la recherche. Les moyens doivent assurer :
1° le fonctionnement général du centre (personnel, locaux, équipement),
2° le coût d'acquisition des données lorsque nécessaire.

Une politique facilitant la mobilité des personnels entre les différents partenaires (CNRS, Universités, Insee, Darés, CEE, Céreq, ...) constituerait un contexte favorable de mobilisation des moyens en personnels et de transfert des compétences. Elle n'est pas du ressort de cette mission. On peut cependant imaginer qu'une politique de mobilité ciblée soit inscrite dans la convention signée. L'attribution de bourses est également un élément important de ce partenariat.

Des moyens significatifs en personnels et en équipement doivent être assurés, quelle que soit la structure mise en place qui s'appuie sur trois sites, Paris et Caen pour le Lamas-IdL et Grenoble pour le CIDSP-BDSP. Le Conseil d'administration et le Conseil scientifique devront définir une politique des sites qui doit impérativement prendre en compte l'accès à des réseaux à gros débits, la création d'espaces de sécurisation

des données, d'espaces d'accueil pour des postes d'allocataires et de chercheurs étrangers, et la collaboration nécessaire et fréquente avec l'Insee et les principaux organismes producteurs de données publiques. Il faut tenir compte de l'éventuelle création d'une zone sécurisée dans un lieu facile d'accès pour les chercheurs français. La montée en puissance doit assurer, sur au moins un des sites, une visibilité internationale.

IV.2. La formation à l'utilisation des données

Une impulsion diversifiée pour améliorer la formation, appuyée sur l'Institut à créer, une collaboration CNRS-Universités et un partenariat avec les producteurs de données.

À côté de la mission de formation confiée à l'Institut, il faut impulser par d'autres canaux la formation à l'utilisation des données. Mieux préparer les étudiants à utiliser et traiter les données de grandes enquêtes dès les premiers cycles, relève de la réflexion des universités. L'impulsion extérieure peut être donnée sous trois formes :

1. Une politique d'allocations ciblées au niveau des Écoles doctorales. La période actuelle de rénégociation s'y prête bien. Il faut parallèlement prévoir une politique d'équipement en postes de travail pour les laboratoires qui encadrent ces doctorants. Ceux-ci pourraient également être accueillis à l'Institut.

2. Il serait souhaitable de mettre en place l'équivalent des bourses Cifre pour les entreprises. Ceci permettrait à des doctorants d'aller travailler sur des données (qu'ils pourraient en même temps contribuer à documenter) dans le cadre des organismes producteurs de données publiques. Dans l'état actuel des choses, il n'existe pas de support institutionnel pour ce type de situation, qui est cependant facile à créer. Ce support doit prendre en compte la nécessité de concilier le temps long de la thèse avec le temps plus court auquel sont habitués les producteurs de données publiques.

.

3. La Formation Permanente du CNRS devrait pouvoir être ouverte plus fortement sur l'université et en particulier les doctorants. Dans le cadre du rapprochement actuel entre le CNRS et les Universités, on pourrait envisager que la Formation permanente du CNRS contribue également à la formation des futurs jeunes chercheurs. La formation à l'utilisation des grandes enquêtes, à ses outils d'analyse se prête particulièrement bien à ce type d'opération, qui pourrait être montée à titre d'expérience.

4. À l'instar de ce qui se développe en Grande-Bretagne, l'Institut pourrait développer des relais dans les Universités, implantés par exemple dans les bibliothèques universitaires, disposant d'une personne ressource, diffusant l'information sur les données disponibles et apte à aider ou orienter les étudiants vers les canaux les plus appropriés.

IV.3. Un financement pour la production d'enquêtes universitaires

L'inscription au Fonds National de la Science d'un financement d'enquêtes universitaires ou de coproductions, appuyée sur un Conseil scientifique

Afin de rendre possible, lorsque cela apparaît nécessaire, le financement d'enquêtes universitaires ou des coproductions université ou CNRS avec l'Insee ou d'autres organismes producteurs de données publiques, un financement devrait être inscrit au Fonds National de la Science.

Un Conseil scientifique serait alors créé pour examiner la légitimité et l'utilité de ces enquêtes, ainsi que leur faisabilité. Il pourrait allier une procédure d'appel d'offres à une procédure prenant en compte les propositions des chercheurs. L'Insee devrait notamment y disposer d'une représentation.

Ce mécanisme doit rester distinct d'un institut de diffusion de données. Cependant, avant tout accord, une vérification auprès de l'Institut (sur le modèle de ce qui est fait au Royaume-Uni) pourrait être exigée afin de s'assurer que des données identiques ne sont pas déjà disponibles. Une coordination avec le Cnis, qui assure la cohérence du système statistique français, devrait être organisée, ne serait-ce qu'à titre d'information. Les producteurs seraient libres de lui demander éventuellement un avis, ce qui constituerait une aide à la qualité de l'enquête. Dans le cas où il s'agirait de la participation française à une enquête européenne ou internationale, l'intérêt de l'information demeure pour le Cnis, dans la mesure où des données concernant la France sont produites. D'autre part, ces enquêtes peuvent constituer un maillon dans l'articulation des enquêtes au niveau européen que le Cnis doit désormais prendre en compte, et sur lequel il est amené à faire valoir le point de vue de la France.

Toute production ou coproduction d'enquêtes assurée sur ce financement devrait obligatoirement faire l'objet d'un dépôt gratuit à l'Institut créé. Il n'apparaît pas utile par contre de confondre financement de coproduction et politique d'acquisition des données.

IV.4. Une politique informatique pour les sciences sociales

Il faut prendre en compte au niveau de la politique informatique pour les sciences sociales, tant dans les universités qu'au CNRS, le coût de l'instrumentation nécessaire pour les chercheurs qui sont de plus en plus conduits à traiter des données d'enquêtes de grande taille. Le réseau des Maisons de la recherche est une aide en ce sens. Mais les chercheurs sont localisés dans l'ensemble des laboratoires universitaires ou du CNRS et sur l'ensemble des sites. Le coût d'équipement en micro-ordinateurs, de mise à niveau chaque année, de renouvellement et de licences de logiciels, n'a pas jusqu'à présent été pris en compte.

IV.5. La validation scientifique de la documentation des données

Le nettoyage et la documentation des données constituent des opérations coûteuses en temps de travail mais où les connaissances et le travail des chercheurs sont indispensables. Si l'on veut impliquer les chercheurs dans ces opérations qui conditionnent le partage des données, en particulier pour celles qui sont produites dans un cadre universitaire, il faut que ce travail soit reconnu scientifiquement à côté des publications. Ceci est du ressort des instances d'évaluation des chercheurs.

Conclusions

Un accord large entre chercheurs et producteurs de la statistique publique sur les principes et le mode d'organisation du partage des données peut être trouvé.

De grandes infrastructures pour les sciences sociales ont vu le jour depuis les années 50, qui ont permis de franchir un pas dans la rigueur et la cumulativité. Il s'agit d'instituts, dénommés *Data Archives*, diffusant aux chercheurs les grandes enquêtes universitaires ou issues de la statistique publique, largement inexploitées. Cette évolution s'est accompagnée d'une politique de recherche sur le long terme qui a permis le financement de grandes enquêtes généralistes, réalisées par des universitaires, ainsi que la participation des chercheurs de chaque pays aux grandes enquêtes européennes et internationales.

La France a été jusqu'à présent largement absente de ces deux dispositifs. L'existence d'un institut national de la statistique, l'Insee, à caractère scientifique est un atout, envié à l'étranger, mais aussi un facteur parmi d'autres d'un éloignement progressif des chercheurs et universitaires des données et de leurs conditions de production. L'étroitesse du milieu aujourd'hui intéressé et formé à l'utilisation de ces données, l'absence de financement permettant à la France de participer aux grandes enquêtes européennes et internationales, ou à des co-productions, avec l'Insee notamment dans le cadre de collaborations de recherche, sont contre-productives, de l'avis même des producteurs de la statistique publique. L'accès aux données de la statistique publique demeure cependant encore difficile et pourrait le devenir encore plus, dans le contexte actuel d'inquiétudes grandissantes et légitimes sur la protection de la vie privée. Si le CNRS a pu signer une convention avec l'Insee, la diffusion aux Universités n'est pas réglée. Les contraintes introduites par la Cnil pour les données infracommunales du Recensement de 1999 suscitent une très grande inquiétude chez les géographes. Régler ces problèmes requiert une véritable organisation de la diffusion des données des grandes enquêtes pour les chercheurs. C'est un problème de structure comme de moyens.

Cela devient d'autant plus nécessaire que la révolution informatique, qui permet aujourd'hui un accès de plus en plus rapide à des sources diverses, et l'organisation en réseau des grands *Data Archives*, en particulier européens, vont permettre aux sciences sociales d'effectuer un saut significatif.

Pour que la France puisse être en mesure de s'insérer dans ce dispositif, il faut impulser une politique de recherche ambitieuse et de long terme et surmonter les difficultés qui ont été résolues ailleurs.

Il faut viser trois objectifs de front : faciliter l'accès aux données, accroître cette utilisation en assurant une meilleure formation des jeunes chercheurs à l'utilisation et au traitement de ces données, rapprocher les chercheurs de la production des données en ouvrant des possibilités de financement autonome, ou en collaboration avec les producteurs de données publiques, de grandes enquêtes intéressant la recherche.

L'accès aux données individuelles, qui sont seules adaptées au processus de la recherche, pose des problèmes juridiques de garanties relatives à la protection de la vie privée, rendus plus vifs aujourd'hui par la puissance des outils informatiques. Cette protection est assurée dans le cadre européen qui a en même temps reconnu la finalité de recherche comme statut spécifique. On peut espérer qu'elle sera complètement prise en compte dans la refonte en cours de la loi Informatique et Liberté de 1978 en France. La contrepartie doit naturellement et nécessairement en être une professionnalisation du milieu de la recherche en sciences sociales, assumant ses responsabilités.

La création d'un Institut doté d'un conseil scientifique où sont représentés les producteurs de données aux côtés des universitaires, l'élaboration de codes professionnels permettent de répondre à cette exigence. Ceci peut être fait au moyen d'une Convention cadre associant le ministère de l'Éducation nationale, de la Recherche et de la Technologie, le ministère de l'Économie, des Finances et de l'Industrie, et le ministère de l'Emploi et de la Solidarité. Cette convention est la voie qui permet d'ouvrir la diffusion des données à l'ensemble des chercheurs (CNRS comme universitaires).

Cet Institut doit disposer de moyens et de personnels au niveau des instituts analogues en Europe. C'est une condition indispensable pour garantir la sécurité des données. C'est également le moyen d'impulser une politique d'aide aux producteurs de données pour documenter les enquêtes, condition indispensable de leur diffusion aux chercheurs. Il s'agit là d'un point de blocage central sur lequel les organismes producteurs de la statistique publique seraient prêts à instituer un partenariat.

Il faut en même temps mieux former les jeunes chercheurs à l'utilisation des grandes enquêtes. Cette formation, qui est aussi un instrument de garantie de bonnes pratiques, ne peut être enclenchée que par des mesures diversifiées. L'Institut créé doit, à l'image de ce qui existe à l'étranger, assurer un rôle d'impulsion important. Les écoles doctorales constituent un cadre approprié pour avoir une action incitative en matière d'allocations de recherche. Des actions de formation permanente coordonnées entre le CNRS et les universités doivent pouvoir s'ouvrir aux doctorants et aux post-doc.

Enfin, il faut rapprocher les chercheurs de la production même des données, leur permettre, par l'existence de financements qui pourraient être inscrits au Fonds National de la Science, d'être présents dans la production d'enquêtes, de façon autonome, dans des collaborations avec l'Insee et dans des collaborations internationales. Ceci doit naturellement se faire sous contrôle d'un conseil scientifique propre. Ces mécanismes doivent être indépendants de l'Institut à créer qui pourra apporter un soutien méthodologique, avec l'Insee, à ce type de production. Faire participer davantage les chercheurs à la production des données permettra d'accroître le poids de la recherche française dans le processus d'harmonisation des systèmes d'enquêtes et des nomenclatures, en cours au niveau européen.

C'est donc d'un saut qualitatif qu'il s'agit. Il faut inscrire les jalons construits par le CNRS au niveau d'une politique nationale de la recherche. L'instrumentation pour les sciences sociales doit être prise en compte. Ceci doit se traduire aussi dans une politique informatique qui prenne en compte ces évolutions. L'utilisateur doit avoir accès à des réseaux à gros débits, il doit disposer de micro-ordinateurs puissants. C'est une politique d'ensemble qu'il faut mener en prenant en compte les moyens nécessaires à un Institut diffusant les données et impulsant leur utilisation, ainsi que ceux nécessaires aux utilisateurs.

Un accord large des chercheurs et de l'INSEE et d'autres services ou organismes producteurs de la statistique publique peut être trouvé sur les principes et le mode de fonctionnement de cette diffusion, comme sur les collaborations entre les deux milieux qui peuvent s'y organiser. Ceci devrait faciliter la mise en œuvre d'une politique à long terme qui peut s'appuyer sur les jalons construits au CNRS.

Annexes

Annexe I : Le déroulement de la mission

1. Fonctionnement général

La mission s'est appuyée sur les liens construits entre le CNRS et l'Insee via le Lasmas-IdL. Les organismes publics producteurs de données et partenaires habituels ont été dès le début associés à la réflexion. Mais il fallait d'emblée couvrir l'ensemble du champ de données intéressant les sciences sociales. Dès le départ le CIDSP-BDSP a donc été associé étroitement à cette mission. Il apportait le champ des données d'opinion et des données électorales intéressant les sciences politiques et issues en plus grande proportion de la recherche ou des instituts de sondage.

Un délai très court avait été assigné à cette mission. Plusieurs étapes ont dues être menées concurremment : état des lieux, identification des difficultés et groupes de travail sur les problèmes identifiés ont donc fonctionné parallèlement.

1. L'état des lieux

Des éléments de comparaison à l'étranger où existent depuis longtemps des *Data Archives* étaient importants. On s'est efforcé de regarder précisément les missions et le fonctionnement des centres d'archivage et de diffusion dans quelques pays étrangers, de situer leur place dans le système global de production et de diffusion des données. Des responsables de ces centres ont été consultés et associés aux groupes de travail. Un repérage rapide et non exhaustif de la production de données du côté de la recherche a également été mené. Ont été plus particulièrement examinés les cas de la Grande-Bretagne et de l'Allemagne, ainsi que plus succinctement le Canada et les États-Unis. Nous avons ainsi pu consulter Denise Lievesley, directrice du UK *Data Archive* de 1992 à 1998, Simon Musgrave, l'actuel Directeur, Repke De Vries du *Netherlands Institute for Scientific Information Services*, Paul Bernard professeur à l'Université de Montréal et auteur d'un rapport sur l'Avancement de la recherche utilisant les statistiques sociales au Canada, enfin Margaret Adams responsable du Département Archives électoniques des Archives nationales aux États-Unis. Deux missions ont été faites sur place au Royaume-Uni et en Allemagne qui ont permis d'examiner dans le détail le fonctionnement des institutions d'archivage et de diffusion des données pour les chercheurs. Un questionnaire en anglais a été mis sur le réseau du Cessda, qui a permis d'élargir le champ d'investigation.

Le débat très important aux États-Unis sur les coûts et avantages du partage des données (*sharing data*) a donné lieu depuis une vingtaine d'année à de nombreux articles et ouvrages. La réflexion menée a pris appui sur cette littérature.

On a également cherché à prendre en compte les différences de cadre juridique entre pays. Ce cadre fournit un contexte plus ou moins favorable au partage des données : existence ou pas d'un dépôt légal, dispositions sur l'accès aux données publiques, dispositions sur la protection de la vie privée. Les textes les plus utilisés ont été la Directive européenne, le *Privacy Act* de 1974, le *Freedom of Information Act* aux États-Unis. L'équipe d'Isabelle de Lamberterie, le Cecoji, a rendu plusieurs notes sur les questions juridiques relatives à la propriété intellectuelle sur les bases de données. René Padieu a rédigé des notes précieuses pour cette mission.

Au niveau européen et international, l'attention s'est portée sur le processus actuel de mise en réseaux des centres d'archivage et de diffusion des données, l'Ifdo et le Cessda, les initiatives en cours à la NFS et à l'OCDE, les niveaux européens de production des données (Eurostat), et les processus de production d'enquêtes européennes et internationales.

En France, la consultation a été menée auprès des organismes producteurs de données, l'Insee, mais aussi quelques administrations ou agences gouvernementales (Céreq, DPD, Darés, Drees, ANPE, plusieurs services dépendant du ministère de l'Intérieur). Il n'était pas question de faire une enquête exhaustive à la fois par manque de temps mais aussi parce qu'il s'agissait surtout de repérer quelques éléments de politique générale. Nous avons plus particulièrement mené cette enquête auprès de ceux qui étaient des partenaires habituels des chercheurs en sciences sociales. Ces organismes ont été interrogés sur leur politique d'archivage, de documentation et de diffusion des fichiers d'enquêtes, ainsi que sur leur politique d'association des chercheurs à la production des données. Ils ont également été interrogés sur leur utilisation de données provenant d'autres producteurs. Cette consultation n'est pas complètement achevée. Du côté des organismes privés, on

a utilisé la connaissance déjà accumulée par la BDSP qui archive certaines données d'instituts de sondage.

Le monde de la recherche a été exploré à travers la diffusion très large d'un questionnaire à un grand nombre de laboratoires du CNRS (quel que soit leur statut) et des universités (ces derniers ayant probablement été couverts moins exhaustivement, faute de liste centralisée). Les instituts de recherche (CEE, Ined, Inra, IRD) ont fait l'objet d'un questionnement du même type que les agences gouvernementales. Dans tous les cas, on s'est intéressé à l'utilisation des données, à leur production, aux conditions d'archivage et de documentation, aux conditions du partage des données (gestion du droit d'usage). Le milieu s'est très fortement mobilisé pour répondre à cette enquête de façon détaillée.

Enfin sur certains points il nous a paru nécessaire de compléter le repérage des difficultés par une couverture par domaine et par discipline. Quelques chercheurs ont bien voulu apporter leur collaboration sur ce point.

2. L'identification des grands besoins et le fonctionnement des groupes de travail

Quatre groupes de questions ont été identifiés à partir de cet état des lieux. Elles ont donné lieu à la mise en place de groupes de travail (voir liste ci-dessous) pour préparer des propositions. Ces groupes ont associé des chercheurs de plusieurs disciplines et domaines de recherche, des représentants d'instituts de recherche producteurs et utilisateurs de fichiers d'enquêtes, des représentants de l'Insee, d'administrations ou d'agences gouvernementales, des spécialistes des problèmes d'archivage et de documentation des données, en particulier les Archives contemporaines, des responsables étrangers des Data Archives, des juristes et des spécialistes de la déontologie. Un lien a été assuré avec la Commission de déontologie mise en place par la Société française de statistique qui travaille sur la traduction et l'application de la Directive européenne en France. Le Cnis a également été consulté dans le cadre de cette mission sur la base des propositions faites. La Direction de la Recherche et la Direction du Département SHS qui ont soutenu les activités du Lasmas-IdL et du cidsp-bdsp ont naturellement constamment été tenues au courant du déroulement de la mission.

Une première synthèse des travaux des groupes de travail a été présentée et discutée au cours d'une réunion élargie le 17 mai 1999 sous la présidence de J. Lautman et A. Grelon (voir liste des participants ci-dessous).

Liste des participants aux groupes de travail

Groupe de travail 1 : Champ et questions juridiques (7 et 8 avril 99)

Le groupe a bénéficié de la participation de

Jacques Antoine, professeur honoraire Cnam, L. Arrondel (Delta, CNRS), Mme Benoit (CIDSP-BDSP, Grenoble), Alain Chenu, professeur à l'Université de Versailles-Saint-Quentin et chercheur au Crest-Insee, Mme Laurence Coutrot (CNRS et Céreq), Jacques Galezot (LEIA, Inra), Alain Godinot (Insee), Mme Denise Lievesley, directrice du *Data Archive* de l'Université d'Essex jusqu'en 98, et actuellement chef de la division Statistiques à l'Unesco, Philippe Méhaut (Céreq), René Padieu (Insee), Jean-Pierre Pagès, directeur d'Agoramétrie, Edmond Préteceille (CSU, CNRS), Benoît Riandey (Ined), Mme Thérèse Saint-Julien professeur à Paris I, ex-directrice adjointe du département SHS du CNRS, Mme Isabelle de Lamberterie, directrice de recherche CNRS (Cecoji).

La présidence du groupe a été assurée par Jacques Lautman, professeur à l'Université de Provence, ex-directeur du département SHS du CNRS, qui a rédigé le rapport.

Groupe de travail 2 : Archivage, documentation et outils de diffusion (20 avril 1999)

Participants : Alain Degenne (Lasmas, CNRS), Marie-Claude de la Godelinais (Insee), Pierre Hallier (Céreq), Annick Kieffer (Lasmas, CNRS), Alexandre Kych (Lasmas, CNRS), Jocelyne Léger (Lasmas, CNRS), Marie-Odile Lebeaux (Lasmas, CNRS), Daniel Masson (CIDSP, CNRS), Christine Pétillat (Archives nationales), Bénédicte Sabatier (Delta, CNRS), Annie-Claude Salomon (CIDSP, CNRS), Isabelle Séguy (Ined), Jean-Pierre Teil (Archives nationales).

Rapporteur : Bruno Cautrès.

Groupe de travail 3 : Les échanges producteurs-utilisateurs et la formation à l'utilisation des données (4 mai 1999)

Participants : Bruno Aubusson de Cavarlay (Cesdip et Association Pénombre), Jean Bourdon (Iredu), Mireille Dadoy (CNRS), Nicole Duchet (responsable du bureau national de formation permanente du CNRS), Michel Gollac (CEE), Irène Fournier (Lasmas-IdL), Alexandre Kych (Lasmas-IdL), Marie-Odile Lebeaux (Lasmas-IdL), Dominique Méda (Darés), Lise Mounier (Lasmas-IdL), Denise Pumain (Géographe Paris 1), Bénédicte Sabatier (Delta).

.Rapporteurs : Alain Chenu et Alain Degenne.

Groupe de travail 4 : La production de données par la recherche (3 mai 1999)

Participants : Bruno Cautrès (CIDSP, CNRS), Alain Degenne (Lasmas-IdL), Franz Kraus (MZES, Mannheim), Yannick Lemel (LSQ, Crest-Insee), Béatrice Roy (OIP, FNSP), Roxane Silberman (Lasmas-IdL) , Louis-André Vallet (Lasmas), Pierre Vergès (Lames).

Rapporteur : Michel Forsé.

Réunion de synthèse du 17 mai 1999

Jacques Antoine (professeur honoraire CNAM), Jean-Pierre Butault (Inra), Nicolas Catzaras (Cevipof), Bruno Cautrès (CIDSP-BDSP), Alain Chenu (professeur à l'Université de Saint-Quentin et Crest-Insee), Philippe Chevet (CNRS), Jean-Paul Combessie (Iresco), Michèle Conchon (Archives nationales), Laurence Coutrot (Lasmas-IdL), Nicole Dausque (CNRS), Alain Degenne (Lasmas-IdL), Repke De Vries (*Netherlands Institute for Scientific Information Services*), Michèle Dodeur (Inserm), Jean-Luc Dubois (IRD), Jean-Marie Firdion (Ined), Michel Forsé (Lasmas-IdL), Irène Fournier (Lasmas-IdL), Olivier Galland (OSC), Michel Glaude (Insee), Marie-Odile Gascon (Certu), Francis Godard (Direction des Programmes), André Grelon (EHESS et Lasmas-IdL), Jean-Paul Grémy, François Héran (Ined), Robert Hérin (professeur à l'Université de Caen), Annick Kieffer (Lasmas-IdL), Alexandre Kych (Lasmas-IdL), Isabelle de Lamberterie (Cecoji, CNRS), Gérard Lang (Insee), Jacques Lautman (Université d'Aix-en-Provence), Emmanuel Lazega (Lasmas-IdL), Marie-Odile Lebeaux (Lasmas-IdL), Yannick Lemel (Crest-Insee), Dominique Méda (Darés), Philippe Méhaut (Céreq), Pierre-Michel Menger (EHESS), Simon Musgrave (ESRC *Data Archives)*, René Padieu (Insee), Catherine Paradeise (ENS Cachan), Bruno Péquignot (Direction scientifique SHS, CNRS), Benoît Riandey (Ined), Béatrice Roy (OIP, CNRS), Claude Seibel (Darés), Nadir Sidhoum (ANPE), Roxane Silberman (Lasmas-IdL), Richard Topol (Direction scientifique SHS, CNRS).

2. L'enquête auprès des laboratoires

Irène Fournier, Alexandre Kych, Marie-Odile Lebeaux (Lasmas-IdL)

Dans le cadre de la mission confiée par le ministère de l'Éducation nationale, de la Recherche et de la Technologie, il nous a semblé important de faire un état des lieux sur l'utilisation des données et leur production dans les laboratoires de recherche CNRS et universitaires en sciences sociales. Un questionnaire a été envoyé à plus de 150 unités. Nous avons choisi les formations qui, d'après leur description dans les listings administratifs, relevaient de l'économie, la sociologie, la géographie et l'histoire contemporaine. Nous avons veillé à inclure toutes les unités qui, à notre connaissance, utilisaient ou produisaient des données. Après un envoi par courrier, nous avons fait une relance électronique des non-répondants. Notre échantillon souffre sûrement de manque d'exhaustivité et sur représente les utilisateurs les plus connus. Les laboratoires ayant répondu ont cependant des thèmes de recherche assez variés pour permettre d'esquisser un tableau de l'utilisation et de la production de grandes enquêtes dans le milieu de la recherche.

Taux de réponses au questionnaire

Les questionnaires ont été envoyés pour un tiers dans des unités de géographie, un autre tiers en économie, pour un quart en sociologie et dans quelques labos d'histoire. Le taux de réponse est de 36 %. Ce sont les économistes qui ont le plus fortement répondu (55 % des unités contactées dans ce domaine). Les géographes ont beaucoup moins répondu (25 %) mais nous avions fait un envoi très large, incluant l'environnement, l'urbanisme, l'architecture, et beaucoup d'unités ne se sont sans doute pas senties concernées.

Comme il est habituel dans ce genre d'enquête par questionnaire, les unités répondantes sont très majoritairement intéressées par le sujet puisque 80 % déclarent utiliser des enquêtes produites par d'autres organismes. 38 % des unités ayant répondu disent également produire ou coproduire des enquêtes.

Utilisation d'enquêtes produites par d'autres organismes

Nous allons faire le point sur les enquêtes et les bases de données utilisées. Le comptage est difficile ; en effet comment comparer l'enquête Emploi qui est annuelle et pour laquelle nous disposons des fichiers de 68 à 98, les recensements que l'on peut utiliser sur le plan national ou très localement, les fichiers administratifs, comme celui des permis de construire, les panels, les bases de données du FMI, etc. Le simple dénombrement des données utilisées illustre parfaitement la variété des façons de compter. 161 enquêtes ou bases de données sont citées au moins une fois par au moins l'une des 45 unités utilisatrices. Elles correspondent à 281 citations si l'on considère le nombre des unités utilisatrices et à 634 citations si l'on tient compte en plus de la multiplicité des années pour les données annuelles.

Si on classe rapidement les sources, on peut les répartir ainsi : enquêtes dans 67 % des cas, fichiers administratifs (18 %), panels (5 %) et bases de données (10 %, certainement très sous-évalué). Les enquêtes annuelles ou les extractions de fichiers administratifs répétées annuellement représentent 22 % des cas (et même 55 % si l'on compte les années).

Les vedettes sont le Recensement Général de la Population (sous toutes ses formes : fichiers-détails ou agrégés, fichiers RRA, PALR ou PALT) qui est utilisé 71 fois et les enquêtes Emploi et leurs enquêtes complémentaires, utilisées 113 fois en tenant compte des différentes années. Les autres fichiers cités plusieurs fois sont : les enquêtes Structure des emplois et Structure des salaires, l'Enquête annuelle d'entreprises, le fichier SIRENE, l'Inventaire communal, l'enquête Formation-qualification professionnelle, le recensement général de l'agriculture et les enquêtes Céreq. Mais il faut aussi souligner la diversité des sources utilisées: 85 enquêtes ou bases de données (soit la moitié des 161 fichiers cités au moins une fois) ne sont citées qu'une seule fois par un seul laboratoire et pour une seule année.

L'Insee vient en tête des producteurs de données : 40 % des cas. Les autres ministères, et singulièrement les ministères du Travail et de l'Industrie, représentent encore 21 % des cas. Il ne

faut pas oublier les institutions nationales gestionnaires de la protection sociale qui procurent le tiers des extractions de fichiers administratifs et les institutions internationales et les pays étrangers qui fournissent les deux tiers des bases de données utilisées par nos laboratoires.

Trois fois sur quatre, les données sont fournies directement par le producteur. Le Lasmas a fourni 15 % de l'ensemble des fichiers et bases de données, mais 36 % des données produites par l'Insee. Si la majorité des données est d'accès libre, dans 32 % des cas, les unités ont eu accès aux données par l'intermédiaire d'une convention, d'un contrat ou d'une collaboration établis auprès de l'organisme producteur ou fournisseur des données. Un accord de la Cnil ou du Cnis a été nécessaire dans 15 % des cas. Enfin les laboratoires pensent que, dans 13 % des cas, leurs bonnes relations avec l'organisme producteur ont été indispensables pour accéder aux données qu'ils convoitaient.

Dans 80 % des cas, les sources utilisées sont françaises et trois fois sur quatre elles ont été obtenues gratuitement. Il faut remarquer que les données provenant la Communauté européenne ou de pays européens ont dû être payées deux fois sur trois. Ceci explique peut-être la faible utilisation (8 %) de données européennes.

La moitié des sources sont nommées par des économistes, l'autre moitié à peu près également par les géographes et les sociologues. Un quart des unités ayant utilisé des fichiers ou bases de données en utilise moins de 2 ou 3, la moitié moins de 7 ou 8, les trois quarts moins de 20 à 25 quand 2 labos en utilisent 51 et 62. Il n'y a pas de différences notables dans le nombre de données utilisées selon la discipline des unités, selon leur taille, selon qu'elles déclarent avoir besoin d'autres enquêtes ou non, selon qu'elles sont elles-mêmes productrices d'enquêtes ou non.

Production ou coproduction d'enquêtes

Vingt-et-une unités ont déclaré produire des enquêtes et en décrivent plus d'une centaine, mais seules 5 unités en décrivent plus de 10. Deux unités produisent des enquêtes socio-politiques qui, dans leur grande majorité, sont mises à disposition des chercheurs par l'intermédiaire de la BDSP (CIDSP, Grenoble). Pour utiliser ces enquêtes, une demande d'autorisation auprès des producteurs est la seule contrainte. Les enquêtes décrites par la 3ème unité sont des enquêtes de suivi des étudiants dans une université de la région parisienne. Leur mise à disposition aux chercheurs ne pourrait se faire qu'avec l'autorisation de l'université concernée. Les deux autres unités font depuis de nombreuses années des enquêtes annuelles, l'une dans le domaine de l'éducation et l'autre en géographie sociale.

Pour les 16 autres unités, ce sont des enquêtes souvent ponctuelles (11 unités n'en ont produites que une ou deux) sur des populations souvent très spécifiques (les magistrats de la Cour des Comptes, les comédiens professionnels, les retraités parisiens nés entre 1907 et 1912, par exemple). Dans leur très grande majorité, ces enquêtes peuvent être mises à la disposition des autres chercheurs ; quand ce n'est pas possible, la raison en est souvent la spécificité et la perte d'anonymat de la population visée.

Ce sont souvent des enquêtes lourdes : dans la moitié des cas, elles sont passées en face-à-face, l'autre moitié, par questionnaire. On peut remarquer une enquête faite par minitel. 25 % des enquêtes datent des années 80 et plus de la moitié a été effectuée dans les cinq dernières années.

Dans plus de la moitié des enquêtes, il n'est fait appel à aucun organisme extérieur, que ce soit pour la passation, la saisie ou le dépouillement. Les financements sont très variés mais plus d'une fois sur deux le CNRS est cité parmi les financeurs.

Il est à remarquer que 4 panels ont été mis en œuvre, dont deux sont encore en cours. Le premier, consacré à l'étude de l'entrée en politique des jeunes, a démarré en 1988 et a réalisé en 97 sa 7ème vague. Le second suit une cohorte de retraités parisiens depuis 1974 et se poursuit encore, il s'arrêtera faute de combattants. Un panel est constitué à partir d'un fichier administratif et est enrichi chaque année. Le dernier panel décrit est l'enquête auprès des ménages lorrains, appelé panel lorrain, qui a suivi des ménages pendant 6 ans.

Pour 17 enquêtes, une autorisation a été demandée auprès de la Cnil. Pour les autres enquêtes, l'anonymat les protégeait de toute demande.

Difficultés rencontrées par les chercheurs

Les difficultés pointées par les chercheurs recoupent en grande partie les observations déjà faites au niveau du Lasmas. L'impact négatif des coproductions qui repose le problème de la diffusion, les questions de coût d'accès pour les universitaires, pour les fichiers de l'Insee non couverts par la convention avec le Lasmas, la difficulté d'accès à certains fichiers particuliers (coût d'accès au fichier Sirène et problème de l'accès en ligne pour améliorer les appariements avec d'autres enquêtes, par exemple), la question du recensement, les conditions de cession des données par plusieurs administrations qui en limitent l'usage à des contrats d'études spécifiés, mobilisent l'attention. Est également soulignée l'absence de négociation au niveau global avec la Cnil. Chacun mène seul la négociation. Il en va de même avec le Comité du secret qui gère l'accès aux données des entreprises. Les remarques portent également sur le manque de formation pour certains à l'utilisation des données (ceci touche y compris les économistes sur des formations pointues, alors que pour les sociologues il s'agit plus du coût initialement lourd d'entrée dans la maîtrise des logiciels d'analyse), et dans certains cas d'insuffisance des équipements pour traiter des grosses données, des problèmes de réseaux pour accéder aux centres de calcul.

Conclusion

La situation telle qu'elle ressort de cette enquête est différente en fonction des disciplines et des domaines. Ceci tient à la fois à la formation inégale des chercheurs en matière d'utilisation de fichiers d'enquêtes de grande taille, à des politiques différentes des organismes détenteurs de données mais aussi aux caractéristiques particulières de certaines données (par exemple données d'entreprises) plus ou moins sensibles. Cela tient aussi au rôle et à la position variable des commanditaires de recherches qui peuvent avoir ou non des accès particuliers aux données que les chercheurs n'ont pas pu obtenir pour leur propre compte. Le rôle des relations personnelles dans nombre de cas demeure important.

Les remarques fournies en fin de questionnaire sont un reflet des joies et peines des utilisateurs d'enquêtes. En positif, ce sont les négociations collectives entre la recherche et les organismes producteurs qui facilitent l'accès aux données. Il est demandé d'étendre de plus en plus ces négociations. En négatif, ce sont les difficultés d'accéder à des fichiers administratifs, difficultés liées le plus souvent à la non-anonymisation des données, le manque de valorisation pour le travail de production et de documentation d'enquêtes ; les coûts sont également un obstacle souvent cité.

Liste des laboratoires ayant répondu à l'enquête

EA 438, Laboratoire pour l'étude du développement et de l'aménagement local et régional (Lédalor)

EA 2252**,** Innovation, travail et emploi dans les espaces en mutation (ITEEM), Université de Poitiers

EA 2324, Centre de recherche « populations et sociétés » (Cerpos), Université de Paris X

EA 2329, Centre de sociologie des religions et d'éthique sociale

EP 1789, PRINTEMPS, Université de Versailles-Saint-Quentin

ESA 5038, Laboratoire de la montagne alpine (Lama)

ESA 5039, Institut de recherche économique sur la production et le développement (IREPD)

ESA 5040**,** Groupe de recherche sur la socialisation (GSR)

ESA 5043, Centre de recherche et d'études sociologiques appliquées de la loire (Cresal)

ESA 5066, Laboratoire Interdisciplinaire de recherche sur les ressources Humaines et l'emploi (LIRHE)

ESA 7003, Emploi et Politiques sociales

ESA 7011, Image et ville

ESA 7028, Fondements des organisations et des régulations de l'univers marchand

ESA 7048, Centre d'étude de la vie politique française (Cevipof)

ESA 7082, Laboratoire technique, territoires et sociétés (Latts)

ESA 7088, Centre de Recherche sur la gestion (Cereg)

GDR 903, GDR Réseaux

JE 2131, Centre d'étude et de recherche sur la production, Institut d'urbanisme de Paris

JE 2208, Géomorphologie dynamique et aménagement des littoraux

UMR 109, Département et laboratoire d'économie théorique et appliquées (Delta)
UMR 159, Groupe de sociologie des religions et de la laïcité (GSRL)
UMR 694, Modélisation et simulation pour l'architecture, l'urbanisme et le paysage (MAP)
UMR 5597, Institut de recherche sur l'économie de l'éducation (Iredu)
UMR 5600, Environnement, ville et société
UMR 5601, Laboratoire d'analyse des techniques économiques (Latec)
UMR 5603, Société, environnement, territoire (SET)
UMR 5604, Groupe de recherche en économie mathématique et quantitative (Gremaq)
UMR 6590, Espaces géographiques et sociétés (ESO)
UMR 6824, Groupe d'analyse et de théorie économique (Gate)
UMR 6579, Groupement de recherche en économie quantitative d'Aix-Marseille (Greqam)
UMR 6585, Centre de recherche Rennais en économie et en gestion
UMR 6586, Laboratoire d'économie d'Orléans (IOF)
UMR 6587, Université de Clermont-Ferrand, Centre Gergovia
UMR 7533, Laboratoire dynamiques sociales et recomposition des espaces (Ladyss)
UMR 7545, Modélisation de la dynamique économique et monétaire (Modem)
UMS 828, Observatoire Interrégional du politique (OIP)
UPR 266, Groupe d'étude sur la division sociale et sexuelle du travail (Gedisst)
UPR 267, Culture et sociétés urbaines (CSU)
UPR 320, Lasmas -IdL
UPR 9059, Laboratoire d'économie et de sociologie du travail (Lest)
URA 209, Centre de sociologie des arts (CSA)
URA 210, Centre de sociologie de l'éducation et de la culture (CSEC)
URA 313, Centre de recherches sociologiques sur le droit et les institutions pénales (Cesdip)
URA 362**,** Laboratoire de recherche économiques et sociales (Labores)
URA 919, Mutations Espace et environnement, travail et emploi, industrie et services, stratégies (Metis)
URA 922, Régulation ressources humaines et économie publique, Centre d'études prospectives d'économie Mathématique Appliquée à la planification
URA 926, UMR 8594**,** Macro-économie et analyse des déséquilibres (MAD)
URA 928, Recherche Fondamentale en économie mathématique (Cepremap)
URA 941, Laboratoire d'économie sociale, économie des ressources humaines et gestion non marchande
URA 1243**,** PARIS-EHGO
URA 1249, Laboratoire méditerranéen en sociologie
URA 1738, Histoire sociale
URA 2036, Centre d'enseignement et de recherche en analyse socio-économique (Ceras)
URA 2048, Laboratoire Georges Friedmann
URA 2221, Centre de recherche politique Raymond Aron
Centre de recherche en géographie et aménagement (CRGA), Université Jean Moulin
SEIGAD, Institut de Géographie Alpine

3. L'enquête auprès des instituts de recherche (epst) et des organismes producteurs de données

Les instituts de recherche ainsi que les organismes producteurs de données publiques ont été interrogés sur la base d'un questionnaire qui portait sur :

1) la production et coproduction d'enquêtes ainsi que la détention de fichiers administratifs éventuellement utiles à la recherche,
2) l'archivage et la documentation de l'ensemble de leurs fichiers,
3) leur politique de diffusion des fichiers aux chercheurs, qui tient naturellement compte de la distinction entre fichiers administratifs et enquêtes sur échantillon et de l'encadrement juridique de la Loi de 1951 et de la Loi Informatique et Libertés,
4) les formes de cette mise à disposition (coût, convention d'étude ciblée, restrictions de publication ou pas, groupes d'utilisateurs, problèmes liés aux coproductions, etc.),
5) leur politique d'association des chercheurs extérieurs en amont des enquêtes,
6) l'utilisation par eux-mêmes de donnés produites ailleurs.

Nous n'avons assuré aucune couverture exhaustive en particulier du côté des producteurs de données publiques qui outre l'Insee, comprend ce que l'on peut englober sous le terme d'agences gouvernementales (par exemple le Céreq) et tous les départements statistiques et de recherche des Ministères. Il s'agissait dans le délai très court imparti à la mission, d'effectuer un repérage très général des pratiques, des points de vue et des difficultés.

Les instituts et organismes ont pour la plupart répondu par écrit de façon très détaillée. Ils ont également exprimé à cette occasion leurs demandes à l'égard des objectifs de la mission. On trouvera les principaux résultats de cette enquête dans le corps du rapport (ch II et III). Soulignons simplement que l'enquête a été l'occasion pour ces organismes, dans le cadre d'une réflexion sur leur politique, de s'interroger sur ce qui pourrait être construit en partenariat avec la recherche.

Pour les producteurs de données publiques, deux questions sont apparues essentielles :

1) celle de la documentation, où les moyens et le temps sont partout insuffisants, la priorité étant naturellement accordée aux objectifs de production immédiate d'exploitation et de résultats,
2) celle de la sous-utilisation des données, qui va de pair avec une demande de meilleure visibilité pour les producteurs de données publiques du champ de la recherche et des utilisateurs potentiels.

Sur ces deux points, il y a la possibilité d'établir un partenariat. Des propositions de bourses ont été formulées.

Enfin, il est apparu que des questions de circulation interne des données entre organismes publics étaient soulevées par nos interlocuteurs. Ces questions ne relèvent pas de la mission, dont l'objet est la finalité de recherche pour des organismes de recherche. Elle ressort clairement de l'État et doit de surcroît être replacée dans le cadre du contrôle exercé sur la non-connexion des fichiers détenus par les administrations (qui doit être autorisée par la Cnil). Les organismes ont également fait état de leur intérêt pour l'accès à des données d'autres pays de l'ue.

Instituts de recherche et organismes producteurs de données publiques ayant répondu à l'enquête :

Ined, Inra, CEE, IRD, Insee, Céreq, Darés, Drees, DPD, ANPE, plusieurs services du ministère de l'Intérieur (voir Annexe III).

4. Liste des personnes consultées

Centre National de la Recherche Scientifique, CNRS, Département des Sciences Humaines et Sociales

Marie-Claude Maurel Directrice du Département SHS

Richard Topol, Directeur Scientifique Adjoint

Bruno Péquignot, Chargé de mission

CNRS, Direction de la Formation Permanente

Nicole Duchet, Responsable de la Formation permanente au CNRS

Institut National de la Statistique et des Études Économiques, Insee

Alain Godinot, Chef du département de la Coordination statistique

Michel Glaude, Directeur des statistiques démographiques

Gérard Lang, Chef de la division Environnement juridique de la statistique

Marie-Claude de la Godelinais, Responsable de la Cellule de mise à disposition des enquêtes

Yannick Lemel, Directeur du LSQ

Cnis

Jean-Pierre Puig, Directeur de la coordination statistique et des relations internationales, Secrétaire général

Marie-Hélène Amiel, Secrétaire adjointe

Société Française de Statistique

René Padieu, Inspecteur Général de l'Insee

Ministère de l'Éducation Nationale, de la Recherche et de la Technologie

Antoine Lyon-Caen, Conseiller du Directeur de la recherche

Francis Godard, Directeur adjoint pour l'action concertée incitative Ville, Direction des Programme

Direction de la Programmation et du Développement (DPD-MEN)

Michel Garnier, Directeur

Centre d'Études et de Recherches sur les Qualifications (Céreq)

Hugues Bertrand, Directeur

Philippe Mehaut, Directeur Adjoint

Pierre Hallier

Direction de l'Animation de la Recherche, des Études et des statistiques (Darés), ministère de l'Emploi et de la Solidarité

Claude Seibel, Directeur

Dominique Méda, chef de mission

Institut National d'Étude Démographique (Ined)

François Héran, Directeur

Benoist Riandey

Jean-Marie Firdion

Françoise Moreau

Isabelle Seguy

Centre des Archives contemporaines

Christine Petillat, Conservateur Général

Michèle Conchon

Jean-Pierre Teil

Institut National de Recherches Agronomiques (Inra)

Jean-Pierre Butault, Chef adjoint du Département ESR

Direction de la Recherche des Études de l'Évaluation et des Statistiques (Ministère de l'Emploi et de la Solidarité)

Pierre Strobel, Mission Recherche (DREES)

Catherine Mermilliod, Mission Coordination des Programmes (DREES)

Dominique Euriat, Département des Méthodes et des Systèmes d'Information (DREES)

Agence Nationale Pour l'Emploi (ANPE)

Nadir Sidhoum

CESEM-OPINION

Jacques Antoine

Inserm

Michèle Dodeur

Certu

Marie Odile Gascon

Agoramétrie

Jean-Pierre Pagés, Président

Institut de recherche sur le Développement (IRD)

Philippe Chevet

Unesco

Denise Lievesley, Chef de la division Statistique à l'Unesco Directrice du ESRC-*Data Archive* de 1992 à 98

Netherlands Institute for Scientific Information Services

Monsieur Repke De Vries

The National Archives at College Park

Margaret O'Neill Adams

Data Archive-ESRC

Simon Musgrave, Acting Director de l'ESRC

Université de Montréal

Paul Bernard, professeur, Directeur du département de sociologie

Laboratoire CNRS, Université, EHESS, et ENS

Philippe Amblard, Cecoji
Luc Arrondel, Delta
Bruno Aubusson de Carvalay, Cesdip
Anne-Marie Benoît, CIDSP
Patrice Bourdelais, École des Hautes Études en Sciences Sociales (EHESS)
Jean Bourdon**,** Directeur de l'Institut de Recherche et d'Étude sur l'Éducation (Irédu)
François Bourguignon, Delta
Nicolas Catzaras, Secrétaire général du Centre d'étude de la vie politique française (Cévipof)
Bruno Cautrès, Directeur du Centre d'Informatisation des Données Socio-Politiques (CIDSP-BDSP)
Alain Chenu, Professeur à l'Université de Versailles-Saint Quentin en Yvelines
Jean- Claude Combessie, Directeur de l'Iresco

Frédérique Cormier, Cecoji
Laurence Coutrot, Lasmas - IdL
Mireille Dadoy, Laboratoire Georges Friedmann
Jean-Philippe Damais, Professeur à l'Université de Paris 1 et Paris XIII
Nicole Dausque, Urec
Alain Degenne, Lasmas - IdL
Jean-Luc Dubois, Cecoji
Elisabeth Dupoirier , Directrice de l'Observatoire Interégional du Politique (OIP)
Michel Forsé, Lasmas - IdL
Annie Fouquet, Directrice du Centre d'Études de l'Emploi (CEE)
Irène Fournier, Lasmas - IdL
Olivier Galland, Observatoire Sociologique Changement (OSC)
Michel Gollac, CEE
André Grelon, École des Hautes Études en Sciences Sociales (EHESS), Lasmas-IdL
Jean Paul Grémy, Lasmas - IdL
Yves Guermond, Professeur à l'Université de Rouen
Robert Hérin, Directeur de Maison de la Recherche en Sciences Humaines (MRSH de Caen)
Pierre Janet, Lasmas - IdL
Annick Kieffer, Lasmas - IdL
Alexandre Kych, Lasmas - IdL
Emmanuel Lazega, Lasmas - IdL
Isabelle de Lamberterie, Directrice du Centre d'Etudes sur la Coopération Juridique Internationale (Cecoji)
Jacques Lautman, Professeur à l'Université d'Aix Marseille
Marie-Odile Lebeaux, Lasmas - IdL
Jocelyne Léger, Lasmas - IdL
Daniel Masson, CIDSP
Pierre-Michel Menger , École des Hautes Études en Sciences Sociales (EHESS)
Fabrice Mollo, Cecoji
Lise Mounier, Lasmas - IdL
Catherine Paradeise, Professeur à l'École Normale Supérieure de Cachan
Jean-Jacques Paul, Irédu
Pascal Perrineau, Directeur du Centre d'étude de la vie politique française (Cévipof)
Edmond Préteceille, CSU
Denise Pumain, Professeur à Paris 1, Directrice de l'équipe PARIS-EHGO
Philippe Robert, Directeur du Cesdip
Béatrice Roy, OIP
Bénédicte Sabatier- Labeyrie, Delta
Thérèse Saint-Julien, Professeur à Paris 1
Annie-Claude Salomon, CIDSP
Louis- André Vallet, Lasmas - IdL
Karl Van Meter, Lasmas - IdL
Pierre Verges, Directeur du Lames

Voir également les listes de personnes consultées en Allemagne, en Grande-Bretagne et au Canada par Annick Kieffer en annexe II, ainsi que celles consultées par Jean-Paul Grémy en annexe III.

Annexe II : Les chercheurs et leurs données dans quelques pays d'Europe et d'Amérique du Nord

1. L'archivage et la mise à disposition des données d'enquêtes auprès des chercheurs en Grande-Bretagne

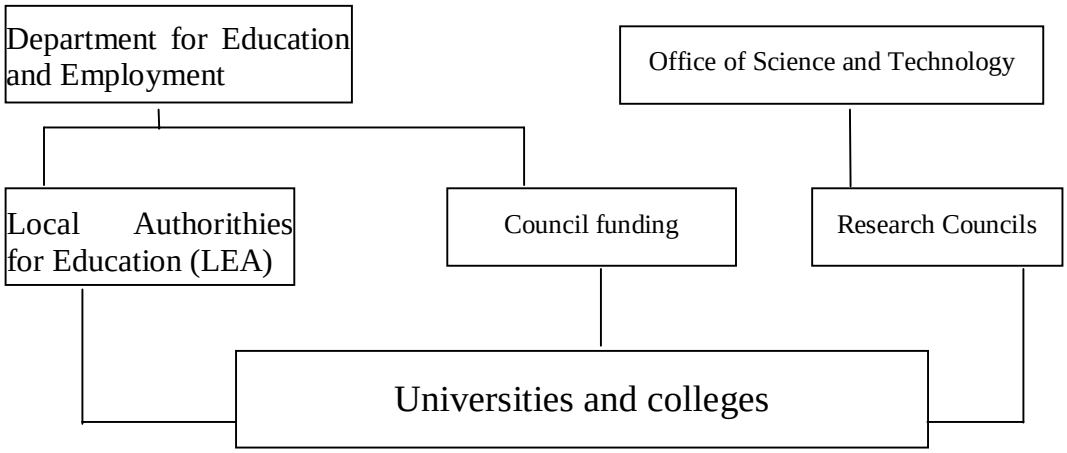
A. Kieffer, J. Léger (Lasmas-IdL)

Centres visités : le *Cathie Marsh Centre for Census and Survey Research (CCSR)*, situé à Manchester et dirigé par Angela Dale ; *L'Institute for Social and Economic Research* situé à l'université d'Essex (Colchester) et dirigé par Jonathan Gershuny ; enfin *l'ESRC Data Archive* situé également à l'Université d'Essex.

Pour mieux comprendre : un détour par l'organisation de la recherche en sciences sociales en Grande-Bretagne.

L'*Office of Science and Technology*, placé auprès du Ministre des sciences de l'énergie et de l'industrie, est l'instance supérieure de la recherche qui définit les grands axes de la recherche et répartit les fonds du budget des sciences entre les 7 conseils de recherche. Il donne les lignes directrices de la politique menée par ces conseils. La politique de l'enseignement supérieur est définie par le *Department for Education and Employment* (DfEE) et financée par le *Higher Education Funding Council* d'Angleterre, d'Écosse du Pays de Galle et d'Irlande du Nord. Il semble que la répartition des prérogatives entre ces deux instances soit assez claire : la politique et les activités de recherche (y compris celles concernant les post-graduate et les doctorants) dépendent de l'OST par l'intermédiaire de ses *Councils* tandis que les coûts indirects (salaires des membres permanents des équipes, les locaux, l'équipement informatique) sont financés par les différents *Higher Education Funding Councils*. Les financements privés viennent essentiellement des organisations caritatives, de l'industrie et du commerce.

Graphique : Principales sources de financement public des institutions de recherche et d'enseignement supérieur en 1996-97



(Source : extrait de HESA finance records 1996-97 English HEIs)

Les centres que nous avons visités sont tous situés dans des Universités. Pour autant, ils ne dépendent d'elles que d'une manière limitée, le véritable moteur de la recherche et des technologies qu'elles mobilisent en Grande-Bretagne sont les 7 *Research Councils* organisés par grands groupes de disciplines. Cette organisation ne facilite par les collaborations entre disciplines ni les programmes interdisciplinaires. La place des sciences sociales est assez marginale dans cet ensemble. Elles dépendent de *l'Economic and Social Research Council* (ESRC). Fondée en 1965, cette agence de moyens a pour rôle de fixer les orientations de recherche grâce à de grands programmes finalisés et de financer les activités de recherche, allant de la formation des personnels aux crédits de fonctionnement, en passant par les équipements et les données d'enquêtes. Son budget annuel est de 65 millions de livres soit environ 650 millions de francs français. Il finance des

centres et des groupements de recherche, des programmes qui impliquent plusieurs centres, des projets qui émanent de chercheurs individuels et de jeunes post-doctorants, ainsi que les ressources ; il délivre également un « label » de la qualité de la recherche. Il est membre de la *Joint Infrastructure Fund* qui finance les équipements et les infrastructures de la recherche universitaire (y compris les bâtiments). Il comporte des comités d'une part (*Research Boards*) et une division de la recherche.

La répartition des financements est opérée par trois comités. Le *Research Priorities Board* (budget d'environ 220 Mo FF) assure le financement de plus de 50 centres et des programmes de recherche. Il contribue à définir les recherches prioritaires (« *what you are told to do* »). Il est composé d'une quinzaine de personnes, des universitaires, des chercheurs du public ou du privé, des représentants de ministères ou de municipalités, du patronat et des trades unions. Les programmes sont composés de 9 thèmes prioritaires qui ont été définis « après une large consultation » des chercheurs, des milieux politiques et des industriels, pour une durée de 10 ans, révisés annuellement. Chaque thème dispose d'un budget indicatif. *Le Research Grants Board* (budget qui représente 1/3 du budget de l'ESRC) finance les projets individuels soumis par les chercheurs ou les doctorants et post-doctorants, et des séminaires. Enfin, le *Research Resources Board* assure le développement des instruments méthodologiques, des équipements technologiques et des ressources (production de données, bibliothèques, archivage et mise à disposition des données etc.). Il définit la politique en la matière. Le *Sample of Anonymous Records* (SAR), le *British Household Panel Study* (BHPS) et le *Data Archive* sont financés par cette instance.

La division de la recherche est organisée en 3 groupes de disciplines : Management, psychologie, linguistique et éducation (MPLE), politique, économie et géographie (PEG), sociologie, histoire, anthropologie et ressources (SHAR), composées chacune d'une équipe de direction (*Research Support Team*) d'une dizaine de membres qui participent à la définition de nouveaux développements et des priorités scientifiques et nomment les experts chargés d'évaluer les projets.

Les centres de recherche sont financés par l'ESRC pour une durée de 10 ans. À la compétition annuelle des centres pour être reconnus par l'ESRC, est substituée depuis peu une compétition organisée autour des 9 thèmes prioritaires. Les centres sont évalués tous les 5 ans avec un bilan effectué à mi-parcours, à 2 ans ½. L'évaluation consiste à regarder si le contrat et les critères (d'une recherche de qualité) ont été respectés. Elle est effectuée par des experts nommés par l'ESRC, qui viennent sur place un ou deux jours. Les rapporteurs font des recommandations le plus souvent. Chaque unité financée par l'ESRC est donc évaluée a priori lorsque l'unité postule, régulièrement au cours de l'exécution du contrat, puis en fin de contrat. Le financement d'une unité par l'ESRC assure le salaire des personnels non permanents et les frais d'équipement pendant la durée du contrat, cependant des financements complémentaires sont le plus souvent nécessaires pour financer les salaires d'un grand nombre de salariés sous contrats précaires. D'autres structures sont également soutenues, telles que des réseaux de recherche, des coopérations internationales, des programmes en coopération avec des « utilisateurs » de la recherche (*Link programmes*).

Le *Joint Information Systems Committee* (JISC) joue un rôle important aux côtés de l'ESRC dans le développement des infrastructures. Il est financé par les ministères de l'enseignement supérieur des différents pays du Royaume-Uni (Écosse, Angleterre, Pays de Galles et Irlande du Nord). Son budget est de 3.87 billions de £ en 1998-99. Il définit les conditions d'attribution, de gestion et de contrôle des crédits attribués aux universités et le nombre d'étudiants correspondant. Il finance des moyens des bibliothèques et des bases de données pour les chercheurs.

L'élaboration progressive d'une politique de données pour la recherche en sciences sociales

Si, comme on le verra, l'origine de la création en Grande-Bretagne d'un *Data Archive* pour la recherche en sciences sociales revient à l'initiative individuelle de politologues soucieux de reproduire dans leur pays l'expérience américaine de l'ICPSR ou du *Roper Center*, très tôt la question d'une infrastructure d'archivage a été liée à celle d'une politique de données pour la recherche. Construire une infrastructure suppose en effet que soient réglées en amont les questions de l'acquisition de fichiers d'enquêtes, du champ et de l'ampleur de l'archivage, en aval celles de leur diffusion, de leur coût pour les chercheurs, de la formation des utilisateurs, du retour vers les

producteurs, donc aussi celle de la valeur ajoutée par l'archivage et la recherche quantitative aux données elles-mêmes.

L'accès aux fichiers d'enquêtes produites par la statistique publique nécessite un niveau de négociation qui dépasse la procédure du simple gré à gré entre la structure d'archivage et le chercheur producteur ou l'institut de sondage privé. Le niveau très élevé des coûts de production des enquêtes publiques ou des Recensements de la population entraîne un déplacement nécessaire vers une politique nationale. L'ESRC a joué ici un rôle majeur. Les chercheurs en sciences sociales qui ont compris l'enjeu des données pour l'avenir et la qualité de la recherche, pour la reconnaissance de son utilité (expertise sociale) par les instances politiques comme par la population, ont progressivement réussi à convaincre l'ESRC de la nécessité de mener une politique cohérente et systématique en matière d'archivage des données pour une mise à disposition auprès de l'ensemble de la recherche publique.

Les freins du côté des différents offices statistiques britanniques ont été faibles, la législation britannique est très libérale en matière de droit d'accès des citoyens aux informations publiques, sous réserve que les droits des individus soient respectés. Les droits de propriété intellectuelle appartiennent clairement aux producteurs, le problème de l'accès est réglé par la licence de distribution, qui peut être concédée à un centre d'archivage. De plus les offices statistiques ont plus vite que dans les autres pays européens été convaincus de l'intérêt de l'analyse secondaire, pour la qualité de la statistique publique, comme du point de vue de la rentabilité des investissements dans les enquêtes. Le *Data Archive* a donc pu signer avec eux des conventions permettant aux chercheurs financés sur fonds publics d'accéder aux fichiers des enquêtes publiques au coût marginal, les fichiers étant déposés par les offices statistiques (maintenant l'ONS) au *Data Archive* dans ce but.

Reste la question des fichiers d'enquêtes produites par les chercheurs. Or on l'a vu, l'ESRC finance une grande partie de la recherche britannique, assume la responsabilité du *Data Archive* d'Essex et enfin mène les négociations avec les offices statistiques, il maîtrise par conséquent toute la chaîne : production, acquisition et accès aux données. En 1997 le Conseil a donc mis au point une politique de données[1] qui édicte une série de règles en la matière qui sont autant de contraintes pour les chercheurs producteurs mais sont aussi des préalables à toute réutilisation par d'autres chercheurs.

* Toute recherche financée par l'ESRC qui comprend une collecte de donnée doit prévoir d'enregistrer et de documenter de manière correcte le fichier d'enquête pour le déposer au *Data Archive*. Passé un certain délai (trois mois après la fin du contrat), la dernière partie du financement est suspendue jusqu'à ce que les matériaux collectés obéissent aux normes de l'archivage du *Data Archive*. Le projet du chercheur doit comprendre une revue des données existantes et justifier l'originalité de la collecte qu'il envisage de faire.

* Mais la décision d'archivage, fondée sur l'intérêt scientifique des données recueillies, revient in fine au comité scientifique du *Data Archive* qui doit en même temps faire une évaluation des coûts de maintenance et d'exploitation de ces données.

* Une licence de distribution peut être signée entre producteur et distributeur, fixant les conditions d'accès, les frais (sachant que l'accès doit être gratuit pour les chercheurs) et spécifiant les autorisations d'accès à des tiers (par exemple les étudiants), le retour des informations vers le producteur (erreurs, création de variables dérivées...). L'ESRC prend les coûts d'acquisition à sa charge. Les conditions d'accès pour les chercheurs étrangers doivent être identiques à celles établies pour les chercheurs britanniques, la différence résidant dans les tarifs (pas de gratuité).

* L'utilisateur de son côté doit citer la source des données qu'il utilise, se conformer aux règles du copyright et de confidentialité des informations obtenues auprès des personnes et aux principes d'éthique de la recherche.

* L'ESRC encourage le *Data Archive* à mener des négociations afin d'élargir l'accès de la communauté sicentifique aux sources gouvernementales et non gouvernementales.

[1]. Les données sont définies ainsi par l'ESRC : « data generated by or of particular interest to the social science community which may be considerated for archiving. This may include computer-readable data, audio and visual recordings, hand-written documents such as diaries and fieldwork notes, photographs and artefacts...»

* Le centre d'archivage et les unités de recherche chargées de la diffusion de certains fichiers (telles que l'ISER ou le CCRS par exemple) doivent mener une politique effective d'information auprès des chercheurs sur les données dont ils ont la responsabilité : l'accès des chercheurs aux fichiers doit être effectif.

Le *Data Archive*

La création du *Data Archive* résulte de l'initiative individuelle, avec le soutien du *Social Science research Council*, d'un professeur de l'Université d'Essex (alors jeune université ouverte à des initiatives originales) le politologue Allen Potter et d'un informaticien de l'université, Eric Roughley, désireux de créer en Grande-Bretagne un *British National Social Science Data Archive* inspiré de l'ICPSR d'Ann Arbor. Fondé en 1967 à l'Université d'Essex, le *Data Archive*, alors dénommé *Data Bank*, le centre d'archivage a pour mission d'assurer la collecte et la diffusion des fichiers informatisés d'enquêtes sur la société britannique en direction des chercheurs, de définir le champ de la collecte, les normes de qualité permettant leur archivage et leur diffusion. Il est pourvu dès sa création d'un conseil scientifique composé de représentants de l'Université d'Essex, de la statistique publique et de ministères, du *Social Science Research Council* (ancêtre de l'ESRC) et d'autres universités, créant ainsi d'emblée les conditions d'une politique nationale d'archivage. Le texte fondateur met en place un consortium d'utilisateurs, le conseil scientifique en fixant les conditions d'entrée.

Les enquêtes archivées sont d'abord des enquêtes socio-politiques, et s'étendent peu à peu (dès le début des années 70) à l'archivage systématique d'enquêtes publiques, la *Family Expenditure Survey* (produite par le *Department for Employment*), puis les enquêtes produites par l'OCPS et le CSO (en particulier la *General Household Survey*). L'implication rapide des organismes de statistique publique est fondée sur la reconnaissance précoce de la valeur ajoutée aux enquêtes par le *Data Archive* (archivage des données, information issue de l'utilisation des fichiers, travail original de diffusion et de documentaion).

Le DA archive en 1998 plus de 700 fichiers d'enquêtes pour l'ensemble de la communauté académique (appartenant à des universités ou centres de recherche publique dont les recherches sont financées sur fonds exclusivement publics) et les étudiants. Il abrite trois unités spécialisées, le *History Data Service* et le Laboratoire virtuel de psychologie expérimentale. Le centre de ressources pour l'accès aux données en Europe (R•Cade), abrité dans un premier temps par le DA, est actuellement pris en charge par la Durham University (il est soumis en effet à une procédure d'appel d'offres au terme de chaque contrat).

Une quarantaine de personnes assurent toutes les activités concernant l'acquisition, l'archivage, la mise à disposition des données et le développement d'outils performants de documentation et d'accès en ligne.

La plus grande partie des enquêtes sont issues de la statistique publique et administrative, l'ONS (*Office for National Statistics*), les ministères (tels que par exemple le *Department for Education and Employment*, *DfEE* principalement avec des enquêtes régulières comme les *General Household Surveys* ou les *Labour Force Surveys* mais le DA archive également des fichiers d'enquêtes produites par des chercheurs ou des instituts privés, à condition qu'elles soient informatisées. Le centre n'est pas propriétaire des fichiers qu'il met à disposition, il n'en est que le diffuseur. D'où l'importance des contrats signés par toutes les parties concernées qui garantissent les obligations et les droits de chacune : le producteur (le propriétaire), le diffuseur et l'utilisateur. Le contrôle de l'usage des données est donc assuré par le DA qui en échange obtient que les utilisateurs disposent du fichier sans avoir à solliciter d'autorisation pour chacune des utilisations. De même le DA présente l'intérêt pour les producteurs de données d'assurer la maintenance des fichiers d'enquêtes et surtout le bon usage des fichiers.

Une des missions du DA est la préservation des données, en particulier des données anciennes et le développement des outils qui assurent l'avenir de l'accès des utilisateurs aux fichiers d'enquêtes. Les données sont reçues ou mises à disposition dans des formats différents et par des moyens ou des supports différents [disquettes, cd-rom, bandes (exabyte), bandes digitales (dat), par réseau (ftp)]. Elles sont stockées au centre qui est responsable de leur maintenance et de leur sécurité selon les normes en vigueur dans le pays. Des copies de sauvegarde sont toutefois effectuées

systématiquement sur cd-rom, exabyte et à l'université de Londres. Les formats de stockage sont indépendants du système et principalement dans des formats fondés sur l'ASCII (format export de SPSS).

Le DA s'assure que la documentation des enquêtes est complète et l'effectue lui-même lorsque le producteur a fournit un fichier mal documenté. Il référence, catalogue et indexe les enquêtes sous une forme standardisée (de la même manière que tout document bibliographique), les utilisateurs ayant l'obligation de citer les sources correctement. Il donne des conseils aux dépositaires sur la manière de documenter les enquêtes et pour cela a été amené à travailler avec eux et avec les utilisateurs pour créer une documentation standardisée. Il a donc créé des groupes de travail pour cela et a participé à des groupes de même nature mis en place dans d'autres pays. L'échange d'expériences au niveau national et international est un atout majeur du DA.

La numérisation de tous les documents nécessaires pour l'analyse secondaire est en cours (base de données relationnelle SIR). Le développement des échanges entre les centres d'archivage de différents pays amène à standardiser la documentation ainsi que les modes d'interrogation des bases. Le DA utilise donc un format utilisé par la plupart des centres d'archivage d'Europe, le « *standard study description* ». Les enquêtes sont cataloguées et indexées selon un système standardisé de description par thème qui recourt à un thesaurus spécialisé. La recherche documentaire peut s'effectuer en ligne grâce au système bibliographique *Biron (Information Retrieval On-line)* sur le site *http:// daww.essex.ac.uk*. Ce système permet d'effectuer une recherche méthodologique, spatiale, par thème, par enquête. L'information obtenue comprend une liste de termes avec tous les thèmes couverts par les données et un catalogue qui fournit le numéro d'archive du fichier, le titre, les conditions d'accès, les producteurs et les sponsors, un résumé des objectifs de l'enquête et les variables. Quelques tableaux sont également fournis. Le thesaurus peut être consulté indépendamment par *Hasset*. Ce point est important, *Hasset* contribue en effet à la socialisation des utilisateurs à la recherche correcte des données dont ils ont besoin.

Le DA délivre gratuitement les enquêtes pour toute recherche financée directement par l'ESRC (seul le prix du matériel est facturé, brochures, support, etc.). Pour les recherches publiques financées par d'autres sources, le chercheur doit payer, la somme étant fonction de la source de financement : modique quand le financement est de source caritative ou associative sans but lucratif (100£), ou public (200£, les chercheurs étrangers sont compris dans cette catégorie), plus élevée lorsqu'il s'agit de recherches financées par des entreprises privées ou entreprises par elles (1000£ outre les taxes qui peuvent être imposées par le dépositaire ou propriétaire du fichier). En outre, pour ces derniers, ne sont délivrées que les enquêtes pour lesquelles les producteurs ont donné leur accord explicite lors du dépôt. Dans ce cas, le DA perçoit les royalties pour le compte du producteur. Les données fournies à des fins d'enseignement sont le plus souvent produites par le DA avec des éditeurs spécialisés ou des organisations d'enseignement ou encore par des utilisateurs qui doivent dans ce cas remplir leurs obligations en redéposant au DA les fichiers qu'ils ont élaborés. L'utilisateur doit remplir un formulaire où il résume le projet de recherche pour lequel il demande tel fichier d'enquête et précise la (ou les) source(s) de financement.

Le centre d'Essex anime des groupes d'utilisateurs autour des enquêtes les plus importantes, afin d'une part de développer les connaissances autour des enquêtes par l'échange d'expériences, mais surtout de disposer d'un lieu d'échange entre utilisateurs et producteurs qui permette aux uns une meilleure utilisation des enquêtes (réflexions communes sur des problèmes rencontrés, échanges de solutions, pertinence des données selon l'objet de la recherche, etc.) et aux seconds de bénéficier des critiques, suggestions, interrogations ou besoins des chercheurs. De plus des séminaires sont mis en place autour d'enquêtes particulières (les plus importants sont autour du *General Household Survey* et de la *Labour Force Survey*), de types de fichiers (par exemple les problèmes de traitement des panels), de méthodes statistiques. Afin de promouvoir l'usage des données d'enquêtes dans les universités et les centres de recherche, le DA a mis en place un réseau de correspondants implantés dans les bibliothèques universitaires qu'il réunit périodiquement. Ces correspondants sont également chargés du retour d'information des utilisateurs vers le DA.

Le DA est membre de l'*International Consortium for Political and Social Research* (ICPSR) de l'Université de Michigan et du *Council for European Social Science Data Archives* (Cessda) ce qui permet d'élargir l'accès aux fichiers étrangers. Il appartient également à l'*International Federation*

of Data Organisations (Ifdo) et une partie de l'équipe appartient à l'*International Association for Social Science Information Services and Technology* (Iassist), ou à l'*International Statistical Institute* (ISI) dont l'adhésion est individuelle mais qui jouent l'un et l'autre un rôle important dans la réflexion et l'échange d'expérience propre à l'analyse secondaire.

Un *Advisory Committee* donne ses avis concernant la politique menée par le centre. Il est composé outre le vice-chancelier d'Essex et du président (ministère de la santé), de deux membres du DA (le directeur et l'administratrice), de deux représentants de l'Université (départements de sociologie et d'économie), de trois représentants d'autres universités, d'un membre du secrétariat de l'ESRC, de deux représentants de dépositaires (dont un de l'ONS) et de trois personnes nommées par l'ESRC (deux universitaires et Ian Maclean, du *Statistics User's Council*).

Budget

Le *Data Archive* a un budget de 1447,3k£ issu de plusieurs sources, la plus importante est l'ESRC (971,3k£, soit environ 63 %), le JISC (304,6k£ soit environ 20 %) et 271.5k£ viennent d'autres sources (Communauté européenne, NFF, vente de services et financements ad hoc, l'université d'Essex). La plus grande partie des dépenses concerne les salaires du personnel et les charges afférentes, viennent ensuite les équipements, les frais de gestion, les frais de gestion de l'université, ceux de recherche et développement ainsi que les équipements. Il assure le financement des différentes activités du centre : acquisition (34,1k£), l'archivage proprement dit (20 % des coûts de production proprement dite) ; la création de métadonnées (dont le catalogage et l'indexation pour l'accès en ligne, 14 % des dépenses), la maintenance des données et leur sauvegarde sous plusieurs formats (12 % des dépenses), l'actualisation de la documentation, l'établissement et l'application des dispositions légales de dépôt et de mise à disposition des données, la création de standards internationaux pour préserver les données, les activités de promotion des enquêtes (ateliers spécifiques, listes mél, publications), l'aide à la localisation des données, la gestion des commandes, la mise à disposition des données (16 % des dépenses) etc. Le *Data Archive* mène des activités de R&D (telle que *NESSTAR*) et comprend un service de données historiques et des collections spéciales dont le budget total s'élève à 3,00k£ environ. Le document fourni avec le budget insiste sur les économies d'échelle en termes tant d'équipement, de R&D, qu'en termes de partage d'expertise, des relations anciennes avec les utilisateurs et avec les producteurs de données, tout ceci mettant le centre dans une très bonne position pour obtenir des financements pour des projets internationaux par exemple.

Le Cathie Marsh Centre for Census and Survey Research (CCRS), Université de Manchester

Le CCRS, dirigé actuellement par le professeur Angela Dale, a été créé en 1992 par l'ESRC comme centre de ressource pour le soutien, la promotion et la mise à disposition des *Samples of Anonymised Records* (SAR) issus du recensement britannique de 1991, permettant pour la première fois de mettre des données individuelles issues du recensement à la disposition des chercheurs, dans le strict respect de la protection des droits individuels (règles de confidentialité et d'anonymisation). Deux échantillons ont été extraits du recensement, le premier au niveau ménage au taux de sondage de 1%, le second au niveau des individus des ménages avec un taux de sondage de 2 %. Le chercheur peut désormais construire ses propres tableaux alors qu'auparavant seules les données agrégées étaient disponibles. Un rapport publié en 1989 par C. Marsh (fondatrice du centre), à l'instigation de l'ESRC et des *Census Offices* de Grande-Bretagne, insistait sur l'importance d'une telle opération et sur les conditions de sa réalisation. La question a été présentée au parlement dès la fin de 1989, et son principe accepté en juillet 1990 par le *Registrars General for England and Wales and for Scotland* (en fait par les services statistiques chargés du Recensement).

Tim Holt (directeur actuel de l'ONS et professeur de statistiques à l'Université de Southampton) et un groupe de statisticiens (dont Denise Lievesley, directrice de l'ESRC *Data Archive* jusqu'à ces dernières années) a rédigé un rapport à l'ONS sur les aspects de confidentialité des SAR. L'autorisation de création des SAR a finalement été prise en mars 1992 ; sa réalisation a été confiée à l'ESRC pour l'Angleterre, l'Écosse et le Pays de Galles, un autre étant réalisé également de manière indépendante, pour l'Irlande du Nord (également sous l'égide de l'ESRC), tout en respectant les conditions permettant une harmonisation au niveau du Royaume-Uni.

L'ONS a le copyright pour le recensement et l'ESRC la licence de distribution. De cette manière l'ONS reconnaît la valeur ajoutée par les chercheurs à ses enquêtes. Il sous-utilise les enquêtes, en les mettant à disposition, il accroît et élargi leur utilisation et leur exploitation et leur donne plus de valeur tout en répartissant les coûts. La licence de diffusion des données a été attribuée à l'Université de Manchester qui a pour cela signé un contrat avec l'ESRC. Tous les établissements et institutions universitaires doivent signer une autorisation (*End User Licence Agreement*) qui leur donne une délégation de responsabilité pour tous les utilisateurs membres de leur établissement, qu'ils soient étudiants ou personnels. Les utilisateurs s'engagent à ne pas tenter d'utiliser les SAR pour identifier un individu ou un ménage et à ne pas céder les données à des utilisateurs qui ne sont pas inscrits. Le CCSR doit en échange s'assurer que les personnes qui demandent les fichiers satisfont bien aux conditions requises et contrôler l'utilisation qui a été faite des fichiers. La sanction est la suppression de toutes les copies du SAR utilisées par l'établissement concerné. Les SAR sont également diffusés par le centre aux établissements du secteur privé qui en font la demande, dans les mêmes conditions (utilisation restreinte au personnel de l'établissement).

L'obtention des fichiers est gratuite pour les chercheurs dont l'équipe est reconnue par l'ESRC et financée par une institution d'enseignement supérieur ou par un *Research Council.* Pour les autres utilisateurs, ou pour les chercheurs financés par d'autres sources, le prix s'élève à 1000£ plus les taxes pour chaque fichier au niveau national (1800£ pour les deux) et 500£ pour un fichier au niveau d'une aire (soit une taille de 120.000 personnes au minimum). La moitié est reversée à l'ESRC qui a acheté les données. Les fichiers sont délivrés dans tous les formats courants (ASCII, SPSS, SAS, SIR, USAR, STATA, etc.), sur différents supports (le plus souvent sur cd-rom, et par FTP). Les utilisateurs peuvent travailler aussi en ligne (via le *Manchester Information Datasets and Associated Services*, ou MIDAS, de l'Université de Manchester, financé par le JISC, l'ESRC et par l'université) ou demander l'extraction de sous-fichiers. Le nombre d'utilisateurs s'élève à 300 environ.

Le Recensement est acheté par l'ESRC à l'ONS à la différence des enquêtes, qui sont données au *Data Archive*. Le prix est de 145 000 £ pour les deux SARs. Le fichier est également archivé au *Data Archive*. La taille est de 40 megabytes pour le fichier ménages et 80 megabytes pour le fichier individus, mais la création de variables peut doubler la taille du fichier.

L'équipe qui assure le SAR est constituée de trois personnes dont Angela Dale, une administratrice et une informaticienne à temps partiel. Elle assure la documentation du fichier, les tests de qualité, les travaux méthodologiques tels que les risques d'identification ou les statistiques locales, les critères d'extraction et prend en charge l'aide aux utilisateurs. Une des activités importantes du centre a été la création de variables dérivées en rapport étroit avec les utilisateurs. Il organise régulièrement des formations courtes pour les utilisateurs (formations aux méthodes statistiques, à l'analyse d'enquêtes). Les compétences acquises par le centre dans l'analyse du recensement, les problèmes de techniques d'échantillonnage, de gestion des données, d'analyse secondaire, d'analyse et de modélisation statistique, etc., permet à l'équipe d'assurer des services de conseils et de consultances auprès des autorités locales ou régionales, des ministères, des organisations caritatives ou des *Research Councils*.

Un groupe d'utilisateurs a été créé. Il est réuni deux fois par an, avec la participation des spécialistes de l'ONS. Le centre a également créé une liste mél d'utilisateurs. L'équipe est membre du *User's Council* de l'ONS pour le recensement et pour la *Longitudinal Study* (l'équivalent de notre Échantillon Démographique Permanent). Elle assure des enseignements ponctuels dans d'autres universités pour présenter les SARs ou leurs travaux fondés sur l'utilisation des SARs.

Le CCSR est toutefois avant tout un centre de recherche pluridisciplinaire, spécialisé dans les recherches sur l'activité des femmes, les rapports entre famille et activité, et les discriminations ethniques face à l'éducation et l'emploi.

L'équipe prépare depuis quelques années les conditions de création de nouveaux SARs à partir du recensement de 2001. L'ESRC a lancé un appel d'offres aux universités pour développer les SAR 2001 et il n'est pas encore certain que le CCSR soit retenu. Le CCSR assure la commercialisation des SARs auprès des entreprises. Ceci est dû au fait que l'ESRC étant l'unique acheteur des recensements de 1991, il en possédait la licence de distribution. Ce ne sera sans doute pas le cas pour les SAR

2001, l'ONS souhaitant conserver une partie de ses droits sur la distribution et la vente des SARs au secteur non universitaire. Ce point est encore en débat.

Budget

Une grande partie du financement est attribué par l'ESRC : 300k£ pour 5 ans soit 60k£ par an pour financer les salaires, l'équipement, les missions et 10k£ pour le fonctionnement, etc. La vente des SARs aux entreprises assure également un petit revenu (mais la plus grande partie est reversée à l'ESRC qui a acheté les fichiers du recensement). Plus de la moitié des dépenses sont liées aux salaires (seul celui d'Angela Dale, professeur sur contrat permanent, n'est pas pris en charge par le centre). L'université assure les frais de gestion, mais ils sont facturés au CCSR.

Un exemple de production d'enquêtes : le British Household Panel Study (BHPS)

un entretien avec Nicholas Buck, directeur adjoint de l'ISER, Université d'Essex

Il y a 11 ans, l'ESRC a inscrit dans ses priorités la création de données socio-économiques longitudinales auprès des ménages. À l'époque, des panels socio-économiques ont été entrepris dans d'autres pays, Allemagne, Luxembourg, Belgique, Lorraine (France) et en partie en Irlande. Les influences les plus marquantes ont été l'ICPSR de l'Université de Michigan et le SOEP allemand. Il importait pour l'ESRC qu'un tel travail soit sous la responsabilité scientifique et technique d'universitaires, et non de la statistique publique, afin de garantir la continuité du projet. L'ESRC a donc organisé une compétition entre les centres universitaires de recherche pour la production d'un panel. L'*ESRC Centre on Micro-social Change* dirigé alors par Tony Coxon et David Rose a obtenu la responsabilité de produire le *British Household Panel Study* dont la première vague a eu lieu en 1988. Ce centre a été intégré récemment dans l'*Institute for Social and Economic Research* (ISER).

On peut distinguer deux phases dans l'histoire de BHPS. Jusqu'en 1993 il s'agissait avant tout d'asseoir le panel sur des bases solides, donc de pouvoir collecter les informations pendant 3 vagues (3 ans). L'accent a donc été mis d'abord sur la méthodologie et la maintenance du panel. Un point a été fait après l'exploitation de ces trois premières vagues : il s'agissait surtout de justifier les dépenses que le panel nécessite. À partir de 1994 non seulement le panel est solidement établi, mais surtout l'équipe dispose des compétences et de l'expérience qui lui permettent d'assurer la continuité, la priorité va alors se déplacer vers les recherches. De 52 en 1993, le nombre d'utilisateurs est passé à 400 en 1997.

Lorsque le panel européen des ménages a été mis en place par Eurostat, l'ONS a décidé de produire son propre panel. Mais la production de deux panels est très coûteuse pour un seul pays. Le coût du BHPS est d'environ 850k£. Il y a eu des arbitrages parfois difficile et finalement le BHPS est devenu la contribution britannique au panel européen. La même situation s'est produite en Allemagne et en Belgique entre autres. Un échantillon de 1000 nouveaux ménages à revenus faibles a été incorporé à partir de la vague 7 du BHPS à cette fin.

Le problème d'attrition s'est posé avec acuité en 1996. De plus de 10000 individus lors de la vague 1 (10300 individus vivant dans 5500 ménages en 1988), le nombre est passé à 6000. Il s'est stabilisé ensuite grâce à une politique de fidélisation des personnes enquêtées, comme cela se pratique dans les autres pays, par un meilleur suivi des ménages lors de leur déménagement, l'inclusion des nouveaux membres vivant dans un ménage formé par un individu-panel, une politique de formation des enquêteurs, etc.

L'équipe de l'ISER élabore le questionnaire et exploite l'enquête, mais c'est le NOP (un institut de sondage) qui assure la passation du questionnaire, produit le fichier de collecte et le codage des réponses. À partir du fichier de collecte, l'équipe restructure le fichier pour les utilisateurs en un fichier rectangulaire. L'équipe assure l'apurement du fichier et la maintenance du panel, traite des questions des non-réponses, des problèmes du suivi des enquêtés, de la qualité des données, de la documentation, de l'organisation des recherches. Elle élabore les conventions relatives aux valeurs manquantes, construit les variables longitudinales avec un même préfixe, le suffixe étant la date de l'enquête, assure la codification des variables plus complexes telles que la position sur une échelle de prestige des professions ou la position sociale (nomenclature SOC). Les scientifiques ont bien ici la maîtrise des opérations les plus importantes : l'élaboration et la construction du questionnaire, la structure du fichier, la création des variables et leur codification.

D'où l'importance centrale de deux autres activités de l'équipe : les relations avec les utilisateurs et la formation. Les relations avec les utilisateurs sont centrales, que ce soit pour intégrer de nouvelles questions de recherche dans le questionnaire, améliorer la qualité de l'enquête, la description des faits et des pratiques, les nomenclatures, ou pour promouvoir les recherches fondées sur le panel. C'est pour cette raison que l'équipe est essentiellement constituée de chercheurs (au nombre de 7 professeurs et 15 chercheurs sous contrat) qui prennent en charge toute la chaîne de production, de traitement, d'analyse et de recherches tant méthodologiques que disciplinaires (ceci occupant environ la moitié de leur temps de travail). Les utilisateurs sont consultés pour qu'ils adressent leurs recommandations pour une série de vagues (la dernière consultation a porté sur la préparation des vagues 9 à 13). Les formations sont de nature diverses. Un master a été créé à l'université d'Essex, mais les étudiants étant peu nombreux à s'y inscrire, la question de son maintien est posé. Des actions de formation permanente aux méthodes d'analyse et de traitement des données longitudinales ont été mises au point avec le SSRU (*Social Statistics Research Unit*) de la City University. En 1993 l'ESRC avait élaboré un projet visant à financer des formations associées pour développer le traitement des données, maintenant les formations portent davantage sur les problèmes d'analyse des données longitudinales.

L'équipe vient d'être déclarée *National Longitudinal Resource Centre*. Elle aura en charge désormais la *Cohort Survey*, ce qui renforce et élargit ses compétences sur les données longitudinales.

Budget

Jusqu'en septembre 1999, l'ESRC a attribué un budget de près de 2000k£ par an (salaires compris, en dehors de ceux des professeurs permanents, la part des salaires assurée par l'ESRC est passée de 77 % en 1994/95 à 53 %). L'université d'Essex accorde environ 200k£ pour les salaires et les financements récurrents, et 75k£ pour les coûts indirects. Des contrats de recherche viennent compléter ces sources de financement, pour la recherche. L'ISER reçoit également d'autres financements, tel que le *British Telecom* pour la publication des BHPS *Newletter*. Les dépenses liées à l'enquête se répartissent approximativement comme suit : 1500k£ sont consacrés à tout le travail lié à la production, la documentation, l'archivage, le traitement et la mise à disposition des données, auxquels il faut ajouter 400k£ pour les salaires exclusivement liés à ces fonctions ; le recours à MAI/NOP (l'institut de sondage) coûte 900k£, les frais divers, d'équipement, etc., 200k£.

L'équipe, outre les chercheurs, comprend 7 personnes pour la gestion du panel (maintenance, nettoyage, etc.), 5 personnes pour les activités informatiques, 2 documentalistes, et une personne assure à temps partiel les relations avec les utilisateurs.

Les relations entre l'ONS et la recherche

Il faudrait développer plus longuement les caractéristiques du système statistique britannique, regroupé depuis peu au sein de l'ONS. Il faut souligner la collaboration ancienne et régulière entre les universitaires et les statisticiens tant pour l'expertise des enquêtes, pour l'analyse secondaire que pour mener des réflexions plus larges sur la conception de nomenclatures, des questions méthodologiques (les problèmes de mémoire par exemple, ou les conséquences d'une codification automatique des professions) ou encore la pertinence de tel mode de description des faits économiques ou sociaux. Récemment la réflexion sur la rénovation du système des classifications sociales a été confié à des universitaires (sous la direction de David Rose et Karen O'Reilly de l'ISER). Mais ces rapports sont considérés comme des contributions de la recherche au développement des outils de description économiques et sociaux. Il ne s'agit pas de véritable partenariat, et, à notre connaissance, il existe peu d'exemples en Grande-Bretagne de groupes composés de chercheurs et de statisticiens, sous la responsabilité de l'ONS, pour l'analyse primaire d'une enquête. Le partage des tâches semble plus marqué même si les relations semblent mieux établies et plus anciennes.

Liste des personnes consultées :

Cathie Marsh Centre for Census and Survey Research (CCSR), Université de Manchester :

Pr. Angela Dale, directrice

ESRC-Data Archive, Université d'Essex :

Simon Musgrave, Acting Director de l'ESRC-Data Archive

Rowan Currie, Senior administrator

Melanie Wright, acting director, end user services

Sheila Anderson, acting director, depositor services

Simon Jones, director ressource management

Hilary Beedham, head of processing, depositor services

Pr. Denise Lievesley, directrice de l'ESRC-Data Archive (1992-1996)

Institute for Social and Economic Research, (ISER), Université d'essex

Nicholas Buck, Associate Director

2. La production d'enquêtes et la mise à disposition des données auprès des chercheurs en Allemagne

Annick Kieffer (Lasmas-IdL)

Centres visités : le Zuma et le MZES (Université de Mannheim), le Deutsche Institut für Wirtschaftsforschung (DIW) et le Max-Planck Institut für Bildungsforschung (Berlin)

La situation allemande est complexe, en raison d'une législation particulièrement restrictive, y compris pour les chercheurs (avec, notons-le un infléchissement récent) en matière de protection des individus. Les fichiers d'enquête sont archivés au *Zentral-Archiv*, de l'université de Bonn. Le centre est d'abord financé directement par l'État fédéral (mais une fondation, la *Volkswagenstiftung*, a également contribué pendant un temps au financement du ZA), afin de garantir la continuité. Un groupe de travail constitué au sein de la *Deutsche Forschungsgemeinschaft* (DFG) au début des années 80, s'est penché sur la manière dont une infrastructure pour les sciences sociales pourrait être organisée. Ce travail a abouti au regroupement du ZA, du Zuma chargé du développement méthodologique et statistique sur les enquêtes, et de l'*Information Zentrum* au sein du GESIS, dont le financement est assuré par le budget fédéral. Les chercheurs ne peuvent cependant que rarement travailler directement sur tous les fichiers d'enquête dans leur Université ou centre de recherche, en particulier lorsque le taux de sondage est très élevé, ne fournissant pas alors les protections satisfaisantes au regard de la législation. Deux solutions s'offraient à eux jusqu'à ces dernières années : demander des tableaux ad hoc au Zuma ou s'y rendre. Par contre l'obtention des fichiers des différents panels est aisée à un coût très faible directement auprès des producteurs ou bien au ZA. Les chercheurs étrangers qui souhaitent travailler sur le Panel Socio-Économique Allemand peuvent aisément obtenir le fichier spécialement anonymisé à leur intention (qui contient moins de données « sensibles » que le fichier fourni aux nationaux), grâce à l'Université de Syracuse, USA.

Pour comprendre la situation allemande à l'égard de la production et de l'archivage des données pour la recherche en sciences sociales, il apparaît nécessaire de prendre en compte des considérations d'ordre institutionnel, pami lesquelles nous avons retenu :

- la structure fédérale de l'État qui implique une définition des prérogatives et des contributions respectives de l'État fédéral (le *Bund)* et des États membres (les *Länder*) à l'égard de la recherche ;

- la forte autonomie des structures de recherche à l'égard des pouvoirs politiques ;

- la législation qui définit de manière stricte les usages des fichiers de données individuelles produites par l'État fédéral et par les *Länder* afin d'assurer la protection effective des individus ;

- la reconnaissance par les instances politiques (ministères concernés) et scientifiques de la nécessité d'une infrastructure de long terme pour l'archivage et l'aide à la production et aux traitements de données d'enquêtes.

État fédéral et autonomie des structures de recherche

La politique de recherche et les crédits qui lui sont affectés sont déterminés de manière conjointe par le ministère fédéral et les ministères compétents des *Länder* au sein d'une instance commune : la *Bund-Länder Kommission* (BLK, qui définit également les orientations en matière d'éducation et de formation professionnelle). Les modalités et le champs d'application (institutions et organisations concernées) de l'effort commun de recherche ainsi que le rôle et le mode de fonctionnement de la commission sont fixés par un accord-cadre. Ainsi la répartition du financement entre le *Bund* et les *Länder* est établie de la manière suivante : 50/50 pour le financement de la *Deutsche Forschungsgemeinschaft* (DFG) et la *Max-Planck Gesellschaft* (MPG), 90/10 pour la *Fraunhofer-Gesellschaft*, 72/25 pour les programmes particuliers de la DFG et pour les programmes de la « *Blaue Liste* » (intitulée depuis *Wissenschaftsgemeinschaft Gottfried Wilhelm Leibnitz*, WGL), 90/10 pour les grands équipements. Le *Land* dans lequel se trouve l'unité de recherche apporte une contribution plus importante du fait qu'il bénéficie de cette localisation.

Pour prendre ses décision, la commission s'appuie sur les recommandations du conseil de la science, (*Wissenschaftsrat*), instance indépendante formée de représentants des différentes disciplines scientifiques. Elle est composée d'une commission scientifique formée de 32 scientifiques ou de personnalités reconnues et d'un conseil d'administration dans lequel siègent 22 représentants des gouvernement de l'État fédéral et des *Länder*. Les évaluations des institutions et des unités de recherche sont effectuées par le *Wissenschaftsrat.* Une évaluation négative entraîne l'arrêt ou la réduction du financement par la BLK.

Le financement de la recherche intervient essentiellement par le biais des fondations indépendantes, financées, on l'a vu, par l'État fédéral et les *Länder*. La *Deutsche Forschungsgemeinschaft* (DFG) est la principale agence de moyens, financée presque exclusivement par l'argent public (97 %). Elle conseille le Parlement et les autorités régionales sur les questions scientifiques, organise les relations entre recherche et monde économique d'une part, et les relations scientifiques internationales de l'autre. Elle joue un rôle central dans le financement des jeunes chercheurs, en particulier des séjours à l'étranger des post-doctorants. Il s'agit en fait d'une association libre d'institutions de recherche, d'universités, l'adhésion étant examinée par l'assemblée des adhérents et acceptée à la majorité des voix. La DFG finance des programmes de recherche (sur appel d'offres ou à l'initiative de chercheurs, dans le cadre de ses *Sonderforschungsbereiche*). La DFG est composée de commissions de spécialités. La « *Blaue Liste* » ou WGL comprend un peu plus de 82 établissements de recherche ou de services à la recherche non universitaires (soit environ 11.000 personnes), qui exigent soit des financements de long terme, ou développent des thématiques qui ne sont pas prises en charge dans les universités, ou enfin mènent des recherche pluridisciplinaires. Leur point commun est leur intérêt national. La WGL comprend 5 commissions : sciences humaines et recherche sur l'éducation ; sciences sociales, économiques et de l'espace ; sciences de la vie ; mathématiques, sciences de la nature et de l'ingénieur ; sciences de l'environnement. Le *Max-Planck Gesellschaft* (MPG), avec ses 17 instituts, est l'institution de recherche fondamentale la plus importante après les Universités.

L'accès aux données de la statistique publique : une législation contraignante

En Allemagne, la statistique administrative est placée sous la responsabilité des Ministères de l'intérieur du *Bund* et des *Länder*. Les liens avec la recherche sont en général très faibles, les plus distendus étant avec les services statistiques des ministères. La diffusion des fichiers de données individuelles à l'extérieur des services statistiques officiels était interdite et depuis 1987 elle est ouverte de manière très limitée aux chercheurs publics par l'article 16 de la loi sur les statistiques fédérales. De plus, la réutilisation par d'autres, y compris à des fins de recherche et strictement non commerciales, est soumise à des règles contraignantes de confidentialité, qui ont interdit de fait pendant longtemps l'accès des chercheurs aux fichiers des grandes enquêtes auprès des individus, des ménages ou des entreprises, produites par le *Statistisches Bundesamt*. La législation limite strictement les fichiers de ces enquêtes et exige pour toute enquête le risque zéro d'identification d'un individu ou d'une entreprise à partir des informations recueillies.

Cette difficulté pour les chercheurs d'accéder à des informations considérées comme des outils importants de connaissance et comme des instruments empiriques de mesure, a eu deux conséquences importantes pour la recherche empirique en sciences sociales en Allemagne :

- elle a permis une mobilisation importante des chercheurs pour produire eux-mêmes des enquêtes et par conséquent d'obtenir les infrastructures nécessaires. C'est ce qui a présidé en particulier à la création du Zuma ;

- elle a nourri un vif débat au sein de la communauté des chercheurs en sciences sociales et avec les statisticiens sur la question du statut de la recherche publique et sur les règles d'anonymisation.

À cela, il faut ajouter la faible confiance de nombreux citoyens allemands en l'État, dont témoigne le faible taux de réponses au Recensement, le problème prenant une ampleur telle que les Recensements ont été à plusieurs reprises repoussés. L'accès au *Mikrozensus*, enquête proche de notre enquête Emploi, mais dont le taux de sondage est plus élevé (1/100 contre 1/300 pour l'enquête Emploi française) s'est posé de façon cruciale aux chercheurs allemands. Les débats ont d'abord été centrés sur les conditions d'une anonymisation effective (mais non absolue, au risque

zéro) du *Mikrozensus*, condition préalable pour l'accès des chercheurs aux fichiers. Un rapport a été demandé à un groupe de travail placé sous la responsabilité de Walter Müller, professeur à l'université de Mannheim, sur cette question : après avoir mené une expérience de piratage (appariement d'un fichier comprenant des informations issues de sources extérieures et du fichier anonymisé du *Mikrozensus* pour des individus un certain nombre de caractéristiques identiques) et conclu à l'impossibilité (mais non absolue) de réidentification d'un individu, il a élaboré un certain nombre de conditions assurant la sécurité des personnes.

Ce premier travail effectué, un second obstacle restait à lever : faire reconnaître l'intérêt de la réutilisation par les chercheurs des fichiers d'enquêtes produites par la statistique publique. Un accord a été signé entre le *Statistisches Bundesamt* et le Ministère fédéral de la recherche qui permet l'accès des chercheurs à un certain nombre de fichiers, dont au *Mikrozensus* anonymisé. Le coût est fixé à 130 Marks (par chercheur) lorsque le travail du chercheur entre dans le projet pilote du ministère, entre 200 à 4000 DM selon la taille du fichier (nombre de variables) pour les autres.

Lors de l'examen des instituts de recherche en économie relevant de la « *Blaue Liste* » en janvier 1998, le *Wissenschaftsrat* soulève le problème de la recherche empirique en Allemagne, et demande qu'une réflexion soit menée sur ses structures, son organisation, son efficacité. Dans ce cadre, trois économistes de renom, Richard Hauser, Gert Wagner et Klaus Zimmermann, ont rédigé un mémorandum publié en juin de la même année qui a le mérite de montrer le rôle central de la recherche empirique dans l'établissement de connaissances scientifiques : l'examen minutieux et critique des hypothèses exige un accès libre aux données et un processus d'échange entre les chercheurs qui analysent les données et le producteurs de données. Ce mémorandum brosse avec beaucoup de lucidité et de manière très complète les conséquences de la situation allemande à l'égard de la production statistique et de la diffusion des données à des fins de recherche : coupure entre chercheurs et statisticiens, faible influence des chercheurs sur la programmation statistique en dehors de quelques lieux bien identifiables, insuffisante formation des étudiants au maniement des grandes enquêtes. Les auteurs déplorent l'absence de politique de données auprès des chercheurs : non réglementation de l'accès des chercheurs aux données administratives, réglementation de l'accès aux statistiques d'enquêtes selon des règles d'anonymisation qui limitent fortement l'intérêt de ces fichiers pour la recherche ; production de données agrégées à des fins ou dans des délais qui en limitent là aussi l'intérêt pour des scientifiques ; enfin les chercheurs eux-mêmes ne remettent pas systématiquement leurs enquêtes au *Zentralarchiv* de Cologne, peu de contrats de recherche prévoyant explicitement l'archivage à des fins de réutilisation par d'autres chercheurs (la DFG faisant figure de modèle à cet égard).

Les auteurs résument les conditions de développement de la recherche empirique en sciences sociales par la réutilisation de fichiers d'enquêtes :

- la mise au point de procédures d'anonymisation effective ou absolue, au niveau du fichier lui-même comme au niveau de la diffusion d'algorithmes informatiques automatiques au poste de travail du chercheur ; et la prise en charge des coûts d'anonymisation ;

- la continuité, la rapidité et la standardisation de la diffusion des fichiers ;

- la qualité de la documentation des enquêtes au niveau du *Statistisches Bundesamt*, le retour des informations vers le producteur et plus généralement une meilleure coopération entre chercheurs et statisticiens que ce soit sur les questions de nettoyage des donnée ou en amont sur les aspects plus théoriques ; l'élargissement de ces procédures à d'autres producteurs de données individuelles.

Les auteurs produisent enfin un certain nombre de thèses concernant :

- les conditions de financement et d'organisation de la mise à disposition de fichiers anonymisés. Partant de la considération qu'il s'agit de bien public, ils plaident pour un accès à un coût faible pour la recherche publique et pour l'autonomisation de la statistique publique (séparation du Ministère de l'intérieur) qui doit se doter de conseils scientifiques ;

- une meilleure participation des chercheurs à l'établissement des programmes d'enquêtes de la statistique publique par la création de conseil consultatif ;

- l'amélioration des procédures d'anonymisation, l'ouverture de l'accès des chercheurs aux données « sensibles » au sein des établissements de production statistique et l'élaboration d'un code d'éthique des chercheurs (avec retrait de la licence au contrevenant) ;

- l'accès en ligne aux données agrégées, ce qui implique qu'une institution s'occupe de l'information mise sur le site et de la documentation complète, en complémentarité avec le *Statistisches Bundesamt*, et que l'accès aux données soit standardisé et d'un coût faible ;

- le développement de programmes en direction de jeunes chercheurs fondés sur l'utilisation de données d'enquêtes et plus généralement le renforcement de la recherche empirique et l'amélioration de la formation à l'analyse quantitative et ceci dès l'université.

Une infrastructure intégrée d'informations pour les sciences sociales : Le GESIS

Le GESIS (***Ge**sellschaft **S**ozialwissenschaftlicher **I**nfra**s**truktureinrichtungen*) offre aux chercheurs en sciences sociales un ensemble de services, allant de l'acquisition et de la mise à disposition de données quantitatives, à une banque de données bibliographiques, en passant par le conseil et le développement méthodologique et la production d'une enquête annuelle sur le développement social. Il est constitué de trois établissements : un centre de documentation en sciences sociales, *l'Informationszentrum Sozialwissenschaften* (IZ) localisé à Bonn, un centre d'archivage de données d'enquêtes, le *Zentralarchiv für Empirische Sozialforschung* (ZA), qui dépend de l'Université de Cologne, et un centre de compétences, le *Zentrum für Umfragen, Methoden und Analysen* (Zuma) qui produit l'enquête Allbus, contribution allemande au programme ISSP (*International Social Survey Program*). Le GESIS a créé un centre à Berlin, spécialisé dans les relations avec les pays de l'Est sur ces trois domaines.

Créé en 1986 à la suite des recommandations d'un groupe de travail de la DFG (mis en place en 1983) sur l'organisation de l'infrastructure de production et de diffusion d'informations pour la recherche en sciences sociales, il est né de la volonté d'intégrer dans une même structure un ensemble de services et de conseils pour la recherche en sciences sociales. Institution appartenant à la « *Blaue Liste* », le GESIS est financé à hauteur de 80 % par le *Bund* et de 20 % par les *Länder*. Il est évalué, ainsi que ses instituts et l'enquête *Allbus*, tous les 8 ans par le *Wissenschaftsrat*.

Le Zentralarchiv (ZA)

Créé en 1960, il est le centre d'archivage le plus ancien d'Europe. Sa configuration doit beaucoup à la personnalité de Günther Schmölders et surtout d'Erwin Scheuch qui avait travaillé quelque temps au *Roper Center* aux États-Unis. Les coûts élevés de production des enquêtes (estimés actuellement entre 250.000 et 500.000 DM) ont été un argument important pour faire admettre l'importance d'une infrastructure permettant leur réutilisation par d'autres chercheurs. À ses débuts, le ZA était financé par l'Université de Cologne. Scheuch réussit à convaincre le ministre allemand de l'époque de l'intérêt de créer une structure dont le financement et les postes sont garantis dans la durée. En 1971 le Ministère fédéral de la recherche et de la technologie décida de financer directement le ZA.

Ses instances de direction sont composées d'un conseil d'administration, d'un comité de direction et d'un conseil scientifique.

Le ZA archive plus de 4.000 fichiers d'enquêtes et données textuelles, la plupart dans les domaines socio-politiques. Parmi ses enquêtes les plus importantes figurent les enquêtes Eurobaromètres, des enquêtes électorales, les résultats des élections législatives allemandes depuis 1953, les enquêtes Allbus (depuis 1980), les données de l'ISSP (issues de 22 pays, depuis 1985) les enquêtes auprès d'un échantillon d'actifs (*Beschäftigtenstichprobe*) produites par l'*Institut für Arbeitsmarkt und Berufsforschung* (IAB) de Nuremberg et un certain nombre de fichiers d'enquêtes de l'ex-RDA en particulier auprès des cadres de l'État. Le ZA assure la vérification, le nettoyage, la correction des erreurs de codage, la documentation et la mise à disposition des fichiers qui lui sont fournis. Il fournit une aide aux chercheurs pour documenter leurs enquêtes et donne des conseils en analyse secondaire. Il organise régulièrement des formations (environ deux fois par an) au traitement des données et aux méthodes statistiques et des séminaires ou des colloques sur des problèmes tels que l'accès des chercheurs aux donnée publiques par exemple.

Les données sont fournies aux chercheurs sous format SAS ou SPSS via FTP, ou sous d'autres supports tels que disquettes, cd-rom, dat ou cartouches. Outre le fichier d'enquête, sont fournis le dictionnaire des codes, le questionnaire, les informations méthodologiques (plan de sondage, conditions techniques de production de l'enquête, etc.). Les documents sont répartis en trois catégories : 0 accès libre, A, accès limité à la recherche publique ou financée sur fonds publics, B, accès libre limité à la recherche publique ou financée sur fonds publics, mais dont la publication est soumise à autorisation préalable du producteur, C, accès soumis à l'agrément du producteur. Les prix dépendent de la catégorie d'utilisateurs et du support, une réduction de 10 et 50 % est consentie aux universitaires, et les produits sont gratuits pour les fournisseurs de données d'enquêtes : 50 DM pour les dictionnaires des codes et les copies sur support magnétique, 300 à 400 DM de droit l'usage du fichier.

Le ZA assure le catalogage comprenant des informations détaillées sur les données archivées. L'information est disponible en ligne soit à partir des enquêtes, soit à partir des variables (projet en cours d'élaboration à partir des enquêtes ISSP). Il mène des activités de développement en matière de standard d'archivage, de documentation d'enquête et de mode de recherche sur internet, en collaboration avec les centres d'archivage d'autres pays. Il produit également des séries chronologiques qui permettent de suivre des évolutions temporelles ou de faire des comparaisons dans le temps.

Le ZA est membre du Cessda et de l'Ifdo, et son directeur, le Pr. Mochmann est également un des responsables pour l'Europe du Iassist.

Le Zuma

La production de données représentatives ou l'accès aux données existantes est une question centrale en sociologie. Les enquêtes sont en effet les outils d'observation des sociologues. Les sociologues allemands se sont emparés de ces questions à la fin des années 60 et au début des années 70 d'une manière tout à fait originale. En effet, développer une base empirique en sociologie suppose d'accéder aux données existantes quand elles sont de bonnes qualité, donc de créer une institution qui centralise l'acquisition et la mise à disposition de fichiers, cela suppose aussi de produire des enquêtes. Mais les compétences en méthodes et en statistiques nécessaires à la production et au traitement d'enquêtes sont peu développées au sein de la discipline. D'autre part, la coupure forte entre les statisticiens des services officiels statistiques et le monde scientifique en général (y compris avec la statistique mathématique universitaire) introduit un frein supplémentaire au développement d'une sociologie quantitative. Partant de ce constat, les sociologues allemands ont œuvré à la création d'un centre de compétences autonome fournissant aux chercheurs des services dans ces domaines. Ils ne l'ont pas conçu comme un service séparé de la recherche ; bien au contraire, conscients que la qualité et l'efficacité dépendaient de sa relation à la recherche, ils se sont attachés à rassembler des personnes spécialisées dans les méthodes mais menant en même temps des projets de recherche. Enfin ils n'ont pas conçu ce centre de manière exclusive, mais ont œuvré parallèlement à la production d'enquêtes par des centres de recherche, pouvant faire appel à l'expérience et aux conseils du centre de compétences.

Le Zuma a été fondée en 1974 par la *Deutsche Forschungsgemeinschaft* dans le but de doter les sciences sociales d'Allemagne d'une infrastructure de production de données, d'aide à la production d'enquêtes et d'intermédiaire pour l'accès aux données produites par la statistique publique. Depuis son intégration dans une nouvelle configuration, le Gesis, il est devenu un institut de la « *Blaue Liste* » ou WGL. Il est financé conjointement par le *Bund* et les *Länder*.

Dirigé par un professeur de sociologie de l'université de Mannheim (actuellement le Pr. Peter Mohler), le Zuma est néanmoins un institut indépendant de cette université, au contraire du ZA qui est rattaché à l'université de Cologne. Ce choix témoigne de la volonté constante en Allemagne, que l'on retrouve également dans la conception du *Data Archive* d'Essex en Grande-Bretagne, d'articuler fortement les infrastructures pour la recherche à la recherche proprement dite. On estime que les activités de service représentent environ 75 % des activités du centre, celles de recherche 25 % (les chercheurs du Zuma sont dans leur majorité des sociologues). Le centre comprend une centaine de personnes (102 en 1999). La plus grande partie des travaux de recherche du Zuma sont financés par les *Sonderprojekte* de la DFG, par la communauté européenne, les ministères, etc. et

participent du financement complémentaire (*Drittmittel)* qui intervient sur la hauteur du financement de base. Si la plupart de ces travaux de recherche sont centrés sur des questions méthodologiques, beaucoup portent également sur des thématiques sociologiques (inégalités face à l'éducation, choix du conjoint etc.).

Le Zuma offre un ensemble original et cohérent de services à la recherche. Un premier domaine d'activité est lié directement aux services aux utilisateurs. Le département de statistique du centre de compétences développe des travaux et des outils sur les méthodes statistiques (théorie et méthodes de sondage, problèmes et traitement des non-réponses et des données manquantes...) et les procédures d'échantillonnage (récemment sur les méthodes d'échantillonnage d'enquêtes par téléphone par exemple), tandis qu'un second département conseille et aide les chercheurs dans la conception de leur enquête (conception du questionnaire, méthode d'interrogation, méthodes de prétest) et effectue des travaux originaux sur les méthodes d'interview, la conception d'enquêtes téléphoniques, postales ou directes, les problèmes d'interrogation auprès d'une catégorie particulière de population, etc. Le département d'analyse textuelle et de codification effectue le codage des enquêtes produites par les chercheurs ou les conseille dans ce domaine. Il mène en même temps des travaux visant une plus grande standardisation des procédures de classification grâce à l'expérience acquise dans le département de production de l'enquête *Allbus* (*Allgemeine Bevölkerungsumfrage der Sozialwissenschaften*). Cela a permis l'élaboration de « standards démographiques », permettant une homogénéisation des procédures de classement (et donc aussi la constitution d'un corps commun de questions socio-démographiques dans les enquêtes), et une diffusion des classifications standards professionnelles et des branches d'activité économique (en collaboration avec le *Statistisches Bundesamt* et deux instituts de recherche en économie).

Un second domaine d'activités regroupe des compétences plus directement en rapport avec la production d'enquêtes et les activités spécifiques à l'analyse secondaire. Elles visent toutes à faciliter l'accès des chercheurs aux enquêtes ainsi que leur traitement par la constitution d'un savoir commun par grands domaines thématiques. Outre la production d'*Allbus*, le Zuma comprend ainsi une équipe spécialisée dans les données de consommation des ménages (enquêtes de consommation des instituts privés pour la plupart, mais aussi sur l'enquête anonymisée « revenus et budget des ménages » *du Statistisches Bundesamt*), dans le but d'améliorer leur traitement et leur analyse. Le département données individuelles (*Mikrodaten*), placé sous la responsabilité de Bernhard Schimpl-Neimanns, a été créé en 1987 au moment de l'ouverture limitée de la législation. L'équipe sert d'intermédiaire pour faciliter l'accès aux données publiques, centraliser les besoins des chercheurs, en même temps qu'elle développe des compétences en analyse secondaire des grandes enquêtes publiques telles que le Mikrozensus par exemple. Cette équipe a joué un rôle central dans les négociations avec le StBa, dans la réflexion sur les questions d'anonymisation de données sensibles et sur les risque d'identification (réflexion menée avec le StBa, Walter Müller de l'Université de Mannheim, et avec l'IAB pour les enquêtes auprès des employés des entreprises). Le ministère a ainsi pu acheter, pour les mettre à la disposition des chercheurs qui viennent travailler au Zuma, quelques fichiers anonymisés du Mikrozensus (1962-1969, 1973, 1976, 1980, 1982, 1985, 1987, 1989, 1991, 1993, 1995). L'équipe s'occupe de la mise à disposition de la documentation en particulier sur le web, elle a produit des clés de passage lorsque les classifications ont été modifiées, elle organise des ateliers afin d'améliorer les connaissances des utilisateurs sur ces enquêtes. La politique développée ici est fondée sur la durée, y compris pour résoudre les problèmes de coûts d'accès aux données de la statistique publique. Enfin, le département indicateurs, animé par Heinz-Herbert Noll, est spécialisé dans la conception des indicateurs de mode de vie ou du changement social en Allemagne, dans un optique comparatiste. Il a mis au point une banque de données d'indicateurs sociaux et a la responsabilité de la contribution du Zuma au *Datenreport* du *Statistisches Bundesamt*.

Les grandes enquêtes de sociologie en Allemagne

Trois séries d'enquêtes seront présentées ici, en raison de leur importance et de leur continuité : l'enquête *Allbus* du Zuma, le panel socio-économique allemand (SOEP) produit par le *Deutsches Institut für Wirtschaftsforschung* (DIW) et les enquêtes parcours de vie *(Lebensverlaüfe und gesellschatlicher Wandel*) produites par l'*Institut Max-Planck für Bildungsforschung* (MPIB)

Allbus

La première enquête *Allbus* a vu le jour en 1978. Financée ponctuellement, tous les deux ans à partir de 1980, elle dépendait fortement des intérêts des chercheurs qui en assumaient la responsabilité. La DFG, qui en assurait le financement, en a confié la réalisation au Zuma au moment de sa création, lui assurant une continuité dans son financement comme dans sa conception. *Allbus* est la base de la contribution germanique à l'enquête internationale annuelle ISSP depuis 1986. Il s'agit d'une enquête sociale effectuée auprès de 3.000 individus environ élargie à l'est de l'Allemagne à partir de 1992 (un peu plus de 2.000 répondants à l'ouest de 1.000 à l'est), permettant de suivre les évolutions des attitudes, représentations, comportements des allemands ainsi que les principales évolutions de la structure sociale de ce pays. Son ancienneté alliée à sa grande régularité en fait un outil important d'observation des évolutions de la société allemande. D'où l'attention portée par ses responsables à la comparabilité des enquêtes entre elles (standardisation, homogénéisation des procédures, des méthodes, des interrogations, des codages et des systèmes de classification), afin de disposer de séries temporelles de très bonne qualité. Elle est constituée d'une enquête centrale qui comprend une série de questions stables pendant quelques années (caractéristiques socio-démographiques, revenu, religion, intentions de vote, prestige de la profession, classe selon le schéma de Goldthorpe, etc.) et de modules approfondis (par exemple la religion en 1982 et 92, comportements à l'égard des groupes ethniques en 1984 et 94) et des modules plus courts (contacts avec les étrangers en 80-84-88-90-94 et 96, fierté nationale en 88-91-92-96). *Allbus* est doté d'un conseil scientifique propre et est étroitement lié à l'association allemande des sociologues qui présentent leurs besoins.

Le département du Zuma chargé d'*Allbus* comporte 5 personnes, dont une est totalement affectée à la partie ISSP. Le coût de production d'une enquête est estimé à 500.000 à 700.000 DM (hors salaire du personnel du Zuma). L'enquête et son management sont effectués par un institut de sondage, la durée de l'interview est d'une heure en moyenne. L'équipe assure le travail en amont (formulation des questions, problèmes méthodologiques), et de la validité des résultats (mais pas directement la comparaison avec les statistiques officielles, il y a là évidemment un problème du contrôle de la qualité des enquêtes).

78 % des utilisateurs d'*Allbus* sont des sociologues et 61 % des universitaires. Les liens avec les utilisateurs ne sont pas directs, puisque les fichiers sont archivés au ZA. Le retour vers les producteurs est classique : insuffisant car les chercheurs ne fournissent pas leurs publications, malgré l'engagement qu'ils ont signé, et le ZA ne fournit au Zuma que la liste des institutions qui ont demandé le fichier et non celle nominative des utilisateurs. Le cd-rom de l'enquête coûte 50 DM pour chaque chercheur.

L'enquête *Allbus* est archivée au ZA et mise à la disposition des chercheurs extérieurs au Zuma dès que le fichier a été nettoyé et documenté. Pour autant l'équipe du Zuma, dirigée par Achim Koch, ne fait pas que produire l'enquête, elle en effectue également l'analyse régulière.

Le panel socio-économique allemand (SOEP)

Le panel socio-économique allemand est l'un des plus anciens panels auprès des ménages en Europe. Le problème rencontré par les chercheurs pour accéder aux données publiques a amené la DFG à lancer un vaste programme d'enquêtes dans le cadre de sa programmation à moyen terme 1975-1989 : enquêtes transversales menées par Richard Hauser entre 1977 et 1979, enquêtes sur le bien-être de Heinz-Herbert Noll, études de cohortes de Karl-Ulrich Mayer, une enquête auprès des salariés, une enquête auprès des étrangers et en 1983 le projet d'un panel socio-économique sur le modèle du PSID américain.

Le projet débuta en 1986, mais il fallu 5 années jusqu'à sa réalisation effective que la DFG a confié au DIW (Berlin). Comme pour *Allbus*, il s'agit d'emblée de fournir des données à l'ensemble des chercheurs en sciences sociales. Si le DIW en est le producteur, il n'en est pas le propriétaire, son seul rôle est de fournir aux chercheurs des données de très bonne qualité. C'est pourquoi les responsables du projet n'ont pas souhaité de financements extérieurs, par exemple des ministères intéressés par tel ou tel aspect. En n'acceptant que les financements ordinaires de la recherche en Allemagne, l'équipe du SOEP a choisi l'indépendance. Le panel est donc financé, dans le cadre des programmes de la DFG, à 50 % par le *Bund* et à 50 % par les *Länder*, essentiellement par le *Land* de

Berlin (tandis que le DIW est financé dans le cadre de la *Blaue Liste)*. Le SOEP pourrait être institutionnalisé (comme le GESIS par exemple), ce qui assurerait un financement sur le long terme, mais le nombre d'instituts financés est en nombre constant, il y a donc un problème de file d'attente. Le SOEP est évalué de manière séparée par la DFG tous les trois ans et avec l'ensemble du DIW tous les 10 ans.

Les coûts de production du SOEP sont estimés à 4 Mio DM par an, salaires et équipements compris, dont 3.1 Mio pour la collecte par l'institut de sondage (Infratest, Münich). L'équipe, dirigée par Gert Wagner, comprend 9 personnes, des sociologues et des économistes. L'institut de sondage effectue tout le travail de production de l'échantillon, de collecte, de codage, de contrôle et de nettoyage du fichier transversal (à l'inverse des choix effectués pour *Allbus*, qui dispose de compétences sur place ou de l'équipe du BHPS en Grande-Bretagne). L'équipe du SOEP élabore les règles de production de l'enquête pour Infratest, collabore étroitement au travail de l'institut. En revanche elle prend totalement en charge la partie la plus délicate d'un panel, la construction des variables, des indicateurs et des fichiers longitudinaux, elle met également au point le dictionnaire des codes. La documentation ainsi que les tris à plat sont disponibles sur le web. Si le SOEP est bien archivé au ZA, ce sont les producteurs qui prennent totalement en charge la diffusion des fichiers auprès des utilisateurs en raison de leur caractère particulier : les compétences sont au DIW, c'est là que les utilisateurs obtiendront l'aide dont ils ont besoin.

L'équipe du SOEP accorde un soin attentif aux utilisateurs (leur capital en fait). Pour obtenir le fichier, chaque chercheur doit signer un contrat où il décrit sa recherche et s'engage à ne pas diffuser le fichier à d'autres. Les producteurs du panel insistent sur l'importance des liens avec les utilisateurs et mènent une politique très cohérente dans ce domaine, sollicitant critiques, remarques, problèmes rencontrés, propositions. Une association des amis du panel a également été créée, qui décerne un prix pour les meilleures publications issues du panel.

300 chercheurs utilisent le SOEP, tant en Allemagne qu'à l'étranger. En effet, l'Université de Syracuse, intéressée par une comparaison, participe à la création d'un fichier anonymisé selon la législation allemande et prend en charge sa diffusion à l'étranger. Elle assure la traduction en anglais de toute la documentation. Les frais pour les chercheurs sont faibles : 50 DM pour un Cd-Rom par vague, 100 $ pour le fichier américain. Syracuse a également construit, avec l'équipe allemande, un fichier unique harmonisé SOEP-PSID. Un projet en cours permettra de fournir un fichier unique harmonisé sur 4 pays par l'intégration des données du panel britannique et du panel produit par Statistique Canada. En outre le StBa ayant décidé de se retirer du projet du panel européen des ménages, c'est désormais le SOEP qui fournit les données à Eurostat, ce qui a exigé un travail de mise en conformité de certaines de ses questions et la création d'indicateurs selon les normes communautaires de la statistique publique. Enfin le SOEP constitue la contribution allemande au projet PACO d'Eurostat, base commune des panels européens. Ajoutons que le SOEP fournit à l'OCDE des indicateurs dès lors que le StBa est défaillant. Il produit en outre du matériel pédagogique dans le cadre des livres scolaires.

L'équipe du SOEP édite une lettre d'information (également disponible sur le web) qui fournit les informations sur les évolutions du questionnaire, donne la liste des publications des chercheurs utilisateurs, et même une liste à jour des utilisateurs, ceci afin d'assurer la visibilité de l'utilisation des données. C'est également un bon moyen pour avoir un fichier d'adresses. Notons que cette visibilité porte ses fruits : rares sont les producteurs qui peuvent fournir un tel détail d'informations sur les utilisateurs et sur leurs publications. Une liste d'utilisateurs est également créée sur internet qui permet d'assurer rapidement les échanges entre utilisateurs et producteurs. Le SOEP organise des ateliers de formation à l'utilisation du panel et des données de panel et, tous les deux ans, un colloque scientifique rassemble les utilisateurs des deux rives de l'Atlantique. Une version d'apprentissage des données du panel a été conçue à partir d'un échantillon de 50 % de l'échantillon initial.

Les enquêtes parcours de vie du Max-Planck Insititut für Bildungsforschung

Les enquêtes produites par l'équipe de Karl-Ulrich Mayer sont au cœur de sa théorie des parcours de vie (*Lebensverlaüfe*) : pour comprendre la perception des événements par les individus et l'évolution de leurs comportements, il faut tenir compte de l'histoire de leur génération, donc à la

fois du contexte dans lequel ils ont vécu tel événement et de l'âge auquel ils l'ont vécu. Il faut donc articuler, âge, génération et période. C'est pourquoi il a conçu une série d'enquête rétrospectives fondées sur des analyses de cohortes. Trois enquêtes ont été effectuées depuis 1981-83 auprès de générations nées avant et après la seconde guerre mondiale, puis auprès d'une génération née juste après la première guerre mondiale, puis auprès de cohortes plus récentes. Une enquête est actuellement en cours auprès de la cohorte née en 1971. Depuis 1991 une série d'enquêtes auprès des mêmes cohortes a été effectuée auprès des personnes de l'ex-RDA (restées à l'est).

L'enquête sur la cohorte née en 1971 (cofinancée par l'IAB, ce qui permet de rapporter les informations issues des entreprises) coûte 600.000 à 700.000 DM, hors salaire des chercheurs du MPIB. La collecte est confiée à un institut de sondage (Infratest) mais l'équipe du MPIB effectue toutes les autres étapes de la chaîne de production. Pour cela elle a fait appel à la collaboration du Zuma pour la codification en particulier et pour le nettoyage du fichier. Mais les délais ont amené l'équipe du MPIB à mener elle-même cette opération. Enfin pour permettre une comparaison avec les classifications mises au point par la statistique publique, l'équipe effectue la codification des professions avec l'IAB. Le problème est particulièrement crucial pour l'ex-RDA, les énoncés par les répondants de leurs métiers, de leur formation ou de leur position professionnelle, sont donnés dans la nomenclature et dans les intitulés de ce pays. Il s'agit de garder ces énoncés mais aussi de concevoir des clés de passage permettant des comparaisons intra-allemandes.

L'enquête est financée par le Max-Planck, ceci permet d'assurer la continuité du financement. Six personnes sont sur l'ensemble du projet, la plupart étant des doctorants et post-doctorants. Ceci est voulu par Karl-Ulrich Mayer qui voit là le meilleur moyen de disséminer les compétences en sociologie quantitative. Mais l'équipe souffre du manque de personnel permanent (au moins une personne) qui assurerait le transfert des compétences et de la mémoire au sein de l'équipe. Ceci ne lui permet pas de prendre en charge les relations avec les utilisateurs. L'enquête est archivée et documentée au ZA, qui assure également les relations avec les utilisateurs.

Liste des personnes consultées :

Mannheimer Zentrum für europäische Sozialforschung

Franz Kraus, Eurodata

Zuma, Mannheim

Pr. Peter Mohler, directeur

Achim Koch, Abteilung Allbus

Bernhard Schimpl-Neimanns, Abteilung Mikrodaten (responsable)

Heike Wirth, Abteilung Mikrodaten

Simone Schmidt, Abteilung Mikrodaten

Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

Jürgen Schupp, chercheur, panel socio-économique

Max-Planck Institut für Bildungsforschung, Berlin

Pr. Karl-Ulrich Mayer, directeur

Heike Solga, chercheuse, Lebensverlaüfe in der ehemaliger DDR

Inneke Maas, chercheuse, berliner Altersstudie

3. La mise à disposition des données auprès des chercheurs en sciences sociales au Canada

A. Kieffer

Les universités occupent, au Canada comme dans beaucoup de pays, une place centrale dans le dispositif de recherche publique. La programmation et les financements relèvent des agences. Deux d'entre elles jouent un rôle prédominant au niveau fédéral pour les sciences de l'homme et de la société, le Conseil de Recherche en Sciences Humaines (CRSH, organisme autonome qui rend compte qui rend compte de ses dépenses au Parlement, par l'intermédiaire du ministre de l'Industrie, mais jouit d'une grande autonomie quant à l'établissement de ses choix et de sa politique scientifique) et, récemment la Fondation Canadienne pour l'Innovation (FCI), créée en 1997 pour financer les infrastructures (qui exigent des financements de long terme).

La situation du Canada à l'égard de la production des données est, malgré la proximité des États-Unis, assez proche de la situation française par le poids de Statistique Canada, l'un des instituts nationaux de statistique les plus performants, tant par la richesse des données collectées que par sa capacité d'innovation en méthodes de sondage et instruments de collecte. Certains fichiers de structure particulièrement complexes exigent une solide compétence en traitements statistiques (analyse multiniveau par exemple) que peu de chercheurs maîtrisent. La contrepartie de cette situation est, comme en France, le faible niveau de production d'enquêtes académiques.

Or l'accès aux données de Statistique Canada était difficile tant du fait des coûts élevés des fichiers, que de la législation : Statistique Canada est en effet garant et responsable de la protection de la confidentialité des données qu'il collecte. Une autre caractéristique du Canada du point de vue de l'archivage des données est l'absence de centre national d'archivage de données informatisées, comme cela existe aux États-Unis, mais aussi depuis quelques années, en France.

Les universités se sont donc tout naturellement tournées vers leur voisin américain pour obtenir des données d'enquêtes. Dès 1980, des universités comme celle d'Alberta par exemple, a créé, sur le modèle américain, des data services de données au sein des bibliothèques. Simples adhérentes au départ de l'ICPSR d'Ann Arbor, Michigan, leur rôle va progressivement s'étoffer : diffuser auprès des chercheurs les données obtenues auprès de l'ICPSR, du centre Roper, mais aussi certaines données de Statistique Canada, des départements statistiques provinciaux, ou du secteur privé, fournir une aide à la localisation et à l'obtention des données, assister les utilisateurs (chercheurs et étudiants) dans le traitement des données (formatage, recodage, création de variables originales etc.), enfin veiller à ce que la documentation des enquêtes soit correcte, et produire des métadonnées. Les enquêtes produites par les chercheurs universitaires y sont également archivées. De tels centres n'existent toutefois pas dans toutes les universités du pays, certaines sont en effet trop petites ou trop pauvres pour cela. Aussi vont-elles s'organiser en consortium : ainsi l'université d'Alberta assure ce service pour 14 autres universités de l'ouest canadien. Un accord de partenariat a été signé avec des laboratoires ou instituts de recherche, permettant des échanges de services (environ 20% de l'activité du centre). L'aide aux étudiants est une activité importante de ces centres (le data service de l'université d'Alberta estime que cette aide représente 1/3 de l'activité qu'elle consacre aux utilisateurs de leur université).

Les universités canadiennes accordent les moyens nécessaires aux universités pour piloter ces activités. Ainsi, l'équipe de l'université d'Alberta est composée de deux personnes à temps plein et de trois à temps partiel. Les équipements sont également à la hauteur de l'ambition : deux stations Unix, de nombreux postes de travail équipés, un terminal X, deux Mac Intosh dont un équipé d'un graveur de cd-Rom, et 3 PC. Le budget du service est de 60.000 $. Le service assure enfin des sessions de formation aux données qu'il archive et organise une école d'été en analyse quantitative.

De même les grandes universités du Québec assurent un service identique pour l'ensemble des universités québécoises. Elles ont mis au point un interface internet, Sherlock, qui fournit en ligne

une description complète des enquêtes, des informations sur la documentation (métadonnées), et pour les personnes autorisées, une extraction du fichier ou de variables.

Du point de vue juridique, ces centres disposent d'une licence de diffusion auprès des distributeurs de données.

À l'instar de leurs collègues américains, les archivistes universitaires spécialisés dans les données d'enquêtes, ont formé des associations dynamiques ; ils jouent un rôle important aux côtés des universitaires et d'autres catégories d'utilisateurs au sein de l'Association Canadienne des utilisateurs de données publiques (Capdu). Cette association a rédigé un rapport en septembre 1998, qui lance un cri d'alerte sur les conséquences de l'absence d'une politique nationale d'archivage de données publiques, prône la création d'un centre national d'archivage et en établit les principes.

Mais l'initiative et le dynamisme des archivistes d'enquête ne suffit pas. Le CRSH de son côté s'est ému de la faiblesse des chercheurs canadiens en analyse quantitative (mais ce constat est le même dans de nombreux pays, y compris aux USA, voir le colloque organisé en novembre 1998 par le *National Research Council* américain qui s'inquiète de la diminution du nombre de chercheurs compétents en méthodes quantitatives. Un groupe de travail formé de statisticiens et de chercheurs a donc été créé en 1998, afin d'étudier les questions de l'utilisation des données, tout particulièrement celles de Statistique Canada, et les obstacles à leur utilisation. Le rapport, remis en janvier 1999, fait un certain nombre de recommandations visant à améliorer la formation des étudiants et des chercheurs, à promouvoir des recherches fondées sur l'utilisation de données d'enquêtes, en particulier de statistique Canada et à resserrer les liens entre la recherche en sciences sociales et les décideurs politiques.

Mais la décision la plus significative dans ce domaine a été la mise en place de l'Initiative de Démocratisation des Données (IDD). Cette initiative, adoptée en 1996, associe la Fondation des Sciences Sociales et des Humanités (HSSFC), l'Association des bibliothèques de recherche (ABRC) l'Association des bibliothèques des petites universités (Casul), l'Association des utilisateurs des données publiques (Capdu), Statistique Canada et les ministères fédéraux, sur une logique de partage des compétences, des services et des financements. Elle permet aux universités adhérentes d'acquérir des données selon des tarifs fixés annuellement : l'abonnement coûte 3.000 $ aux universités membres de la Casul, 12.000 $ à celles qui sont membres de l'ABRC) afin de les mettre à la disposition de leurs enseignants et étudiants, à des fins de recherche non commerciales. Chaque université doit assurer l'aide aux utilisateurs en personnel et en logiciels, Statistique canada apporte le soutien technique et le protocole de transfert de fichiers, via internet, les ministères fédéraux assurent une partie des financements de l'opération. Les universités adhérentes (63 en 1999) jouent ainsi le rôle de guichets. La propriété intellectuelle des données reste au gouvernement fédéral canadien, les universités signant une licence d'utilisation.

L'IDD permet également d'améliorer la formation des étudiants en techniques quantitatives. Une attention toute particulière a été accordé à ce point. Statistique Canada fournit des petits fichiers sur cd-Rom aux élèves des écoles (lycées) et aux étudiants des premiers cycles universitaires afin de les initier aux traitements des données et d'attirer des jeunes vers les recherches quantitatives. De même des bourses sont accordées aux étudiants avancés pour promouvoir des recherches qui s'appuient sur le traitement des données. Mais surtout, l'IDD permet de poser les premiers jalons pour une politique nationale d'archivage de données d'enquêtes, de coordination des centres existants, de promotion des recherches quantitatives et de formation des chercheurs en ce domaine. Elle permet en outre de resserrer les liens entre chercheurs et statisticiens canadiens.

Liste des personnes consultées :

Chuck Humphrey, Université d'Alberta

Gaëtan Drolet, Université de Québec

Paul Bernard, Université de Montréal

4. Les États-Unis. Quelques remarques sur l'évolution actuelle d'une institution pionnière : l'ICPSR

Annick Kieffer, Roxane Silberman (Lasmas-IdL)

Les États-Unis ont vu se constituer les deux premières banques importantes de grandes enquêtes pour les chercheurs en sciences sociales, le Roper Centre et l'ICPSR à l'Université de Michigan. Ces deux centres avaient nettement au départ une ambition mondiale, justifiée à la fois par le projet de travaux comparatifs de leurs fondateurs et le fait qu'il n'existait pas par ailleurs de centre de même type. S'ils abritent toujours des enquêtes provenant de plusieurs pays, dont certaines auraient d'ailleurs été perdues si elles n'y avaient pas été archivées à l'époque, il est clair que la dimension mondiale est aujourd'hui moins forte. Plusieurs *Data Archives* ont été créés en Europe, assurant une couverture des champs nationaux et rassemblant des compétences sur les enquêtes. Aux États-Unis même, les banques de données sur les grandes enquêtes se sont elles-mêmes multipliées. De très nombreuses universités possèdent aujourd'hui un centre. C'est le résultat d'une part de la politique de diffusion très libérale de l'ICPSR, qui a cherché dès le départ à créer des relais pour inciter à utiliser les enquêtes et pour trouver des partenaires. Mais c'est aussi l'effet d'une multiplication de centres thématiques au plus près des chercheurs qui ont cherché à construire des instruments plus performants dans leur domaine propre, comme celui de l'existence de grosses enquêtes produites par quelques universités qui en assurent la diffusion. On assiste donc à une mise en réseau de ces centres qui a conduit l'ICPSR à une réflexion approfondie et exemplaire sur l'évolution nécessaire dans ce contexte.

L'ICPSR (*Inter-University Consortium for Political and Social Research*) s'est d'abord constitué à l'Université du Michigan (Ann Arbor) comme une fédération d'universités cofinançant par leurs cotisations annuelles le plus grand centre mondial d'archivage de données en sciences sociales. Il regroupe aujourd'hui 325 institutions partenaires, principalement américaines mais pas seulement. Les pays européens ont chacun un représentant national (*National Member*), la BDSP du CIDSP pour la France. Les données archivées à l'ICPSR portent sur plus d'une centaine de pays et constituent un élément très important du « patrimoine » de la recherche empirique en sciences sociales (ainsi c'est à l'ICPSR qu'a été archivée et informatisée la Statistique Générale de la France). L'adhésion à l'ICPSR fonctionne sous une double dimension. Les partenaires mettent à disposition les données qu'ils archivent ou produisent, et accèdent dans le cadre de ce partenariat à l'ensemble des autres données gratuitement, une fois réglée une cotisation qui varie selon les pays et la taille de l'institution. Les utilisateurs non partenaires peuvent également utiliser les données moyennant paiement variable selon l'utilisateur. Partenaires et autres utilisateurs ont également accès à des modules de formation à l'utilisation des enquêtes. L'ICPSR dispose grâce à ces cotisations d'un fonds propre de fonctionnement qui a cependant été abondé très vite par d'autres sources, notamment par la *National Science Foundation*. Elle dispose aussi d'un financement stable qui permet une politique de conservation et de diffusion sur le long terme. La gratuité des données publiques aux États-Unis facilite l'acquisition des données par l'ICPSR.

L'ICPSR fonctionne sous l'égide de l'Université du Michigan et en étroite collaboration avec l'*Institute for Social Research* de cette même université. Cet institut est à la fois un centre de production de données et de formation aux méthodes de l'enquête en sciences sociales.

L'ICPSR abrite chaque année une école d'été de grande envergure dans le domaine des méthodes quantitatives en sciences sociales. Il s'agit sans aucun doute de l'école de formation de plus haut niveau en ce domaine au plan international. Une trentaine de modules de deux à trois semaines de cours chacun sont proposés. Les points forts de cette école d'été sont le renouvellement des cours proposés et les applications rendues possibles par l'accès aux données de l'ICPSR. L'adhésion de la BDSP du CIDSP à l'ICPSR lui permet d'envoyer chaque année des chercheurs à cette école d'été, avec une bourse de $300. L'ICPSR dispose d'un conseil scientifique composé de personnalités scientifiques importantes dans le domaine de l'analyse des données en sciences sociales, qui se réunit deux fois par an. Chaque banque de données européenne membre du Cessda assure à tour de rôle la représentation européenne à ce conseil scientifique.

L'ICPSR travaille en collaboration étroite avec les banques de données européennes, notamment vis-à-vis des renouvellements suscitées par l'Internet en matière de stockage et de diffusion des données.

La réflexion actuelle porte sur l'évolution nécessaire de l'ICPSR dans un contexte de multiplication des centres thématiques qui met au premier plan la question du référencement des centres les uns sur les autres, de la navigation de l'utilisateur sur un réseau de centres pour trouver les données les plus adéquates, de la coopération entre les centres en matière d'outils de diffusion. Parmi les critiques faites à la structure de l'ICPSR, il faut souligner celle qui porte sur le lien entre mise à disposition des données et partenariat. Lorsque l'un des partenaires est amené à se retirer, la question du devenir des données qu'il avait mises à disposition du centre dans le cadre de ce partenariat est posée. La modulation du tarif des cotisations pour les partenaires, qui relèvent d'universités et de pays aux moyens très différents, a également été source de problèmes constants.

Les infrastructures de soutien aux sciences sociales sont évaluées tous les deux ans par la Direction des sciences économiques et sociales et des sciences du comportement de la *National Science Foundation* et par la Commission des sciences du comportement, des sciences sociales et de l'éducation (CBASSE) du *National Research Council*, les deux principales agences de moyen américaines. L'examen des infrastructures de recherche, mené en 1997, a donné lieu à une réflexion approfondie sur la politique et les moyens à leur accorder, en particulier pour les statistiques. Un séminaire a été organisé sur ce thème, qui a réunit chercheurs et statisticiens. Sur cette base la CBASSE a émis un certain nombre de constats et exprimé des recommandations visant à modifier les procédures de financements afin de remédier aux insuffisances.

La Commission a estimé que le niveau de soutien aux infrastructures doit être augmenté et surtout qu'il faut stimuler les relations entre la recherche, la production et l'utilisation de sources statistiques, relations qui sont insuffisantes dans un certain nombre de disciplines. Pour cela, recommande-t-elle, il faut améliorer la cohérence des sources (production de séries par exemple), l'archivage des données, rendre leur accès non seulement plus aisé (via internet en particulier) mais surtout avec une documentation de qualité, mettre au point des logiciels permettant de traiter les données par un système de filtres qui réponde aux impératifs de confidentialité, et enfin améliorer la formation des chercheurs en méthodologie quantitative. De son côté, la Direction des sciences économiques et sociales et des sciences du comportement de la *National Science Foundation* a lancé une procédure d'appel d'offres en 1999 visant à accroître l'infrastructure de recherche pour les sciences sociales : collecte de données nouvelles pour la communauté, systèmes d'archivage sur le web, de partage de données au moyen du web en particulier, les projets devant intégrer un programme de formation des chercheurs. Les projets retenus (4 à 8) seront financés pendant 10 ans à la hauteur de ½ à 1 Mio$ par an.

5. Représentants du Council of European Social Sciences Data Archives (Cessda)

Pays	Nom	Ville
Allemagne	ZA - Zentralarchiv für empirische Sozialforschung	Cologne
Autriche	WISDOM - Wiener Institut für Sozialwissenschaftliche Dokumentation und Methodik	Vienne
Belgique	BASS - Belgian Archives for the Social Sciences	Louvain-la-Neuve
Danemark	DDA - Dansk Data Archiv	Odense
Espagne	CIS - Centro de Investigaciones Sociologicas	Madrid
Estonia	ESSDA - Estonian Social Science Data Archive	Tartu
France	BDSP - Banque de Données Socio-Politiques Centre d'Informatisation des Données Socio-Politiques - CIDSP-IEP	Grenoble
Grande-Bretagne	UK-DA - The Data Archive	Colchester
Hongrie	TARKI - Társadalomkutatási Informatikai Egyesülés	Budapest
Italie	ADPSS - Archivio Dati e Programmi per le Scienze Sociali Instituto Superiore di Sociologia	Milan
Norvège	NSD - Norsk samfunnsvitenskapelig datatjeneste	Bergen
Pays-Bas	STAR - Steinmetz-archief	Amsterdam
Pays-Bas	CESSDA secretariat - c/o Steinmetz-archief NIWI - Nederlands Instituut voor Wetenschappelijke Informatiediensten	Amsterdam
Suède	SSD - Svensk Samhällsvetenskaplig Datatjänst	Göteborg
Suisse	SIDOS - Swiss Information and Data Archive Service for the Social Sciences	Neuchâtel

Annexe III : Quelques exemples de données particulières

1. Les données statistiques relatives à la sécurité intérieure

(Police, Gendarmerie, Douanes)

Jean-Paul Grémy

Résumé :

1) Les données relatives à la sécurité intérieure présentent un intérêt scientifique certain ;

2) elles ont été jusqu'à présent peu exploitées par les chercheurs ;

3) elles sont dispersées dans des organismes assez divers ;

4) leur disponibilité pour la recherche varie fortement d'un détenteur à l'autre ;

5) les difficultés d'accès à ces données tiennent aux différences de culture professionnelle entre leurs détenteurs et les chercheurs ;

6) il est possible d'améliorer la coopération entre ces deux univers.

1. Ce sont des données importantes pour la connaissance de la société

1.1. Les données sur la sécurité intérieure sont nombreuses et variées

Selon les finalités des organismes qui les détiennent, il peut s'agir de données administratives (individuelles ou agrégées par unités géographiques ou fonctionnelles), ou de données individuelles d'enquête.

Les données administratives portent sur les personnels (recrutement, évolution de carrière, implantation géographique, emploi, etc.), les crimes et délits (lieu et condition de commission, élucidation, caractéristiques des auteurs, etc.), les suites pénales (condamnations, population pénitentiaire, etc.), les victimes.

Les données d'enquêtes d'opinion (le plus souvent par sondage) concernent principalement le moral et les aspirations des personnels, le comportement et les attentes des victimes, l'image des forces de l'ordre dans l'ensemble de la population, le sentiment d'insécurité et les attentes des Français. Certaines de ces données proviennent d'enquêtes internationales, et permettent par conséquent des comparaisons avec divers pays étrangers.

1.2. Elles concernent un problème politique majeur

Le « rapport Peyrefitte » (1977) marque le début de la sensibilisation des responsables politiques, puis de l'opinion, à l'accroissement de l'insécurité dans notre société. Depuis, l'insécurité et la délinquance sont devenus des thèmes de débat majeurs, suscitant des controverses tant sur leurs causes que sur les remèdes à y apporter.

À ce titre, les données statistiques sur l'évolution des diverses formes de délinquance ont un intérêt évident : leur mise en relation d'une part avec l'évolution des moyens alloués aux forces de l'ordre et aux institutions pénales, et d'autre part avec l'évolution des variables socio-économiques et démographiques, est le moyen le plus sûr d'identifier les causes de l'insécurité.

1.3. Ce sont des données « sensibles »

L'importance des enjeux politiques des controverses autour de l'insécurité, leur médiatisation croissante, la charge affective qui leur est associée dans l'opinion, ont contribué à obscurcir et passionner le débat. De plus, le nombre insuffisant de spécialistes de ces problèmes, tant chez les hommes politiques que chez les journalistes et les chercheurs, a accru les risques d'interprétation erronée et suscité des tentations de présentation malveillante.

Tous ces facteurs ont entraîné chez les détenteurs de certaines de ces données une réticence à les ouvrir à la recherche, d'autant que l'utilisation qui a été faite de ces données « sensibles » a eu dans certains cas des répercussions négatives sur les services qui avaient accepté d'en donner l'accès aux chercheurs.

2. Ces données ont jusqu'à présent été peu exploitées

Nous verrons plus loin qu'il existe des fichiers accessibles sans restriction, et d'autres qui peuvent être exploités dans des conditions déterminées. Malgré ces facilités, le volume des travaux de recherche réalisés en France sur ce type de données reste, toutes proportions gardées, bien inférieur au volume des publications anglo-saxonnes sur le sujet.

2.1. Les administrations disposent de peu de moyens pour exploiter ces données en interne

Les ressources des institutions de la sécurité intérieure en matière de recherches socio-économiques sont limitées. À partir de 1972, la Direction Centrale de la Police Judiciaire, qui élabore au sein de sa Division des Études et de la Prospective les statistiques de la délinquance enregistrée par les services de police et de gendarmerie, a fait appel à des statisticiens de l'Insee pour perfectionner ses procédures de collecte et d'analyse. Le Ministère de la Justice dispose d'une unité de recherche associée au CNRS, le Centre d'Études Sociologiques sur le Droit et les Institutions Pénales (Cesdip), qui constitue le centre de recherches le plus important dans le champ pénal. Enfin, l'Institut des Hautes Études sur la Sécurité Intérieure (IHESI), créé en 1989, compte en son sein plusieurs chercheurs (sociologues, politologues et, depuis 1997, statisticiens) et quelques policiers chargés d'études ; les recherches qu'il a produites sont en majorité qualitatives, et ont pour une part importante d'entre elles été commandées à des chercheurs extérieurs.

Il existe quelques autres organismes susceptibles de réaliser ou de commanditer des recherches de type socio-économique. C'est par exemple le cas du Centre de Recherches de la Gendarmerie Nationale, ou du Centre d'Études et de Prévision du Ministère de l'Intérieur. Ce sont en général de petites unités, de création récente, n'ayant encore que peu de production scientifique à leur actif.

2.2. La recherche publique et les universités se sont jusqu'à présent peu intéressées à ces données

Nous l'avons signalé, comparé à celui de la production scientifique anglo-saxonne, le volume des recherches publiées en France sur la sécurité intérieure semble modeste ; en outre, seule une très petite part de celles-ci utilise des données statistiques.

En dehors du Cesdip, qui bénéficie de l'association au CNRS, on ne dénombre que quelques laboratoires de recherches rattachés au CNRS ou aux universités travaillant sur ces thèmes ; les plus importants sont situés à Paris, Grenoble, et Toulouse. D'autre part, quelques chercheurs ou enseignants-chercheurs ne dépendant pas de ces laboratoires spécialisés poursuivent des travaux ou encadrent des mémoires ou des thèses sur ces mêmes thèmes. Enfin, on recense une dizaine environ de formations de troisième cycle susceptibles d'aborder ces sujets ; mais ils sont plutôt orientés vers le DESS que vers le DEA et le doctorat.

Les lacunes de la recherche fondamentale dans ce domaine expliquent que, lors de la création de l'IHESI, les activités de recherche y aient été conçues sur le modèle du CNRS, plutôt que sur celui de la recherche finalisée (ce qui est par exemple le cas au *Home Office*).

3. Ces données sont dispersées dans des organismes divers

Elles se trouvent dispersées dans des sites divers, dépendant de ministères ou de directions différentes. Quatre ministères au moins disposent de données statistiques intéressant la sécurité

intérieure : le ministère de l'Intérieur (police, sécurité civile), le ministère de la Justice (tribunaux, système pénitentiaire), le ministère des Armées (gendarmerie), et le ministère de l'Économie et des finances (douanes, fiscalité).

De plus, il n'est pas impossible que d'autres ministères détiennent des données utilisables pour une analyse fine des problèmes d'insécurité : Éducation nationale (violences scolaires), Affaires étrangères, travail et emploi, Affaires sociales, etc. Par ailleurs, les instituts de sondage réalisent régulièrement des enquêtes d'opinion portant, en totalité ou en partie, sur les problèmes de sécurité ; certaines de ces enquêtes ayant été financées sur fonds publics, il devrait être possible d'y avoir accès.

3.1. Le ministère de l'Intérieur

La Direction Centrale de la Police Judiciaire (DCPJ) est la principale source de données statistiques en matière de sécurité intérieure. Elle élabore et gère un nombre important de fichiers :

- les statistiques de la délinquance enregistrée par les services de police et de gendarmerie (« état 4001 ») ;

- les statistiques d'activité des services de police judiciaire (AGORA) ;

- le système de traitement des infractions constatées (STIC) ;

- les fichiers des sept offices centraux : 1) pour la répression du trafic illicite des stupéfiants (OCRTIS), 2) pour la répression du grand banditisme (OCRB), 3) pour la répression du trafic des armes, des munitions, des produits explosifs et des matières nucléaires, biologiques et chimiques (OCRTAEMS), 4) pour la répression de la traite des êtres humains (OCRETH), 5) pour la répression de la grande délinquance financière (OCRGDF), 6) de lutte contre le trafic des biens culturels (OCBC), 7) pour la répression du faux-monnayage (OCRFM).

La Direction de l'Administration de la Police Nationale (DAPN) gère le fichier des personnels.

La Direction Centrale des Renseignements Généraux (DCRG) a élaboré quelques statistiques sur des problèmes spécifiques, comme les violences dans les banlieues.

La Direction de la Sécurité Civile dispose de statistiques sur la sécurité des Bâtiments.

La préfecture de Police de Paris élabore beaucoup de données destinées à la gestion de la police parisienne : tableaux de bord, rapports d'activité internes, statistiques d'accidents de la circulation, appels police secours.

Enfin, l'IHESI a mis en forme (SAS) et documenté les statistiques annuelles de la délinquance élaborées par la DCPJ, et rassemblé diverses séries socio-économiques de l'Insee afin de les mettre en corrélation avec la délinquance. Il dispose également des fichiers d'enquêtes variées : enquêtes internationales de victimation, analyse d'une cohorte de policiers, etc.

3.2. La Direction Générale de la Gendarmerie Nationale (DGCN)

Le Service Technique de Recherches Judiciaires et de Documentation (STRJD) gère la base JUDEX, créée en 1983, qui rassemble en temps réel les principales informations contenues dans les bases départementales sur les affaires judiciaires traitées (description de chaque affaire, signalement des personnes impliquées, image des objets utilisés ou volés). Selon les divisions, on trouve également divers fichiers sur les personnes recherchées, les véhicules volés, la délinquance itinérante, etc. C'est à partir de cette base que le Bureau de la Police Judiciaire prépare les données qui seront transmises à la DCPJ pour l'élaboration des statistiques de la délinquance enregistrée.

Le fichier des unités (ORG) indique les effectifs de chaque unité, ainsi que les caractéristiques des communes qui en dépendent. Les statistiques de service fournissent un relevé très détaillé de l'activité de chacune des brigades territoriales (nature et durée des mission, contrôles, arrestations, etc.).

Le Service des Ressources Humaines (SDRH) gère le fichier des personnels (fichier PERS), qui contient non seulement les fiches des personnels en activité, mais aussi celles des retraités depuis 1975 ; il dispose également d'un fichier sur les candidats au recrutement dans la Gendarmerie.

Le Sirpa-Gendarmerie réalise peu d'enquêtes spécifiques ; mais la Direction de l'Information et de la Communication Opérationnelle de Défense (DIRCOD) du Sirpa central effectue des enquêtes périodiques sur l'image des forces armées, incluant celle de la Gendarmerie.

3.3. Les Douanes

Le Bureau de la politique des contrôles et de la lutte contre la fraude gère deux fichiers : le fichier national informatisé de documentation (FNID), donnant des informations détaillées sur les infractions constatées, les suspicions de fraudes, les demandes préalables d'enquête ; et le système informatisé de gestion du renseignement sur le trafic des stupéfiants par voie maritime (MARINFO), fichier international (Europe des quinze) analogue dans sa structure au précédent.

D'autres services disposent également de fichiers statistiques susceptibles d'intéresser les chercheurs. La DNSCE gère le fichier SOFI sur le fret international, qui permet d'informer les entreprises sur les chiffres du commerce extérieur ; le Bureau A1 réalise des enquêtes auprès des personnels ; le Bureau d'Information et de Communication réalise des enquêtes auprès des entreprises sur l'image de la douane.

3.4. Les instituts de sondage

Une rapide recension dans *Le Sondoscope* montre qu'il ne se passe pas de mois sans qu'un ou plusieurs des sondages publiés n'abordent les problèmes de sécurité. En outre, les instituts de sondage ont réalisé des enquêtes concernant en totalité ou en partie ces mêmes problèmes, pour le compte de laboratoires ou d'associations variées (Cesdip, IHESI, Observatoire Interrégional du Politique, Agorametrie, etc.). En principe, ces instituts détiennent une copie du fichier des réponses à chacune de ces enquêtes, qui restent la propriété des commanditaires.

Une circulaire, dite « circulaire Balladur », impose aux pouvoirs publics de signaler tout projet d'enquête par sondage au Service d'Information du Gouvernement (SIG, ex-SID), afin de ménager les deniers publics en évitant les doublons ; en principe donc, le SIG doit disposer d'une liste de ces enquêtes, permettant aux chercheurs de repérer celles qui, faute d'avoir fait l'objet d'échos dans les médias, n'auraient pas été recensées dans *Le Sondoscope*.

3.5. Le projet d'un observatoire de la sécurité

L'idée d'un observatoire de la sécurité, qui rassemblerait et exploiterait l'ensemble des données (qualitatives et quantitatives) relatives à la sécurité intérieure, a été plusieurs fois envisagée. Divers projets en ce sens ont été présentés aux pouvoirs publics. Aucune décision n'a encore été prise, malgré la création récente d'un conseil interministériel de la sécurité, qui pourrait s'appuyer sur les analyses conduites dans cet observatoire pour asseoir ses décisions.

4. Les possibilités d'accès aux données varient fortement

L'obtention par les chercheurs d'une autorisation d'accès aux données statistiques sur la sécurité et la délinquance n'est pas systématique ; en effet, dans l'état actuel des choses, cet accès peut dépendre de la conjoncture politique, des expériences passées dans les relations avec les chercheurs, et des positions personnelles du directeur ou du chef de service (en sachant que, dans certains postes, celui-ci change souvent). C'est pourquoi les indications ci-après n'ont pas de portée générale et ne s'appliquent qu'à la situation présente.

4.1. Le ministère de l'Intérieur

Le Directeur Central de la Police Judiciaire s'est montré tout à fait favorable à une collaboration avec les chercheurs. Les statistiques annuelles de la délinquance enregistrée sont naturellement disponibles sans restriction. En ce qui concerne les autres fichiers dont dispose la DCPJ, les conditions de leur accès peuvent être étudiées au cas par cas ; elles incluent évidemment l'anonymisation des fichiers nominatifs, mais surtout la DCPJ compte se prémunir contre les erreurs d'interprétation en ne livrant les données qu'assorties d'un « commentaire d'utilisation », précisant la nature et la portée de ces données. La DCPJ propose en outre de favoriser l'accès des chercheurs aux fichiers d'instances internationales dont elle est le correspondant en France : Interpol (178 pays), Europol, et Schengen (fichier SIRENE).

La Direction de la Sécurité Civile est également prête à accueillir favorablement les demandes des chercheurs intéressés par ses données.

L'IHESI est par vocation un lieu où les demandes des chercheurs devraient être bien accueillies.

La préfecture de Police examinera au cas par cas les demandes qui lui seront adressées ; elle ne formule pas d'objection de principe à la collaboration avec les chercheurs.

Par contre, les autres directions contactées se sont montrées réticentes, et un aval du Cabinet s'avérerait nécessaire avant toute ouverture. Il est donc peu probable que les chercheurs y soient bien accueillis actuellement.

4.2. La Direction Générale de la Gendarmerie Nationale (DGGN)

La DGGN pratique une politique de transparence ; en témoigne l'accueil chaleureux réservé à cette mission. Les expériences passées de collaboration de la DGGN avec des chercheurs ou des thésards se sont révélées positives pour les deux parties. Sous réserve de l'anonymisation des informations individuelles, l'accès à ses fichiers est possible.

Les chercheurs intéressés peuvent adresser leur demande directement au Chef du Service qui détient les données auxquelles ils souhaitent avoir accès ; toutefois, en écrivant au Directeur Général de la Gendarmerie Nationale, les chercheurs ont la garantie que leur demande sera orientée vers le service compétent.

4.3. Les Douanes

Le Bureau de la politique des contrôles et de la lutte contre la fraude s'est montré très coopératif ; en outre, les Douanes et Droits Indirects disposent déjà d'un service de diffusion extérieure qui fournit aux entreprises des données économiques.

4.4. Les instituts de sondage

Les enquêtes réalisées par les instituts de sondage sont la propriété des commanditaires ; c'est donc à ceux-ci qu'il faut s'adresser pour obtenir l'autorisation d'accéder aux fichiers statistiques résultants. Toutefois, il arrive fréquemment que les commanditaires ne disposent pas en interne de ces données, qu'il convient donc de récupérer auprès des instituts après obtention de l'autorisation d'accès.

5. Les obstacles liés aux différences de culture professionnelle

Le contentieux entre les autorités responsables de la sécurité en France et la communauté des chercheurs scientifiques est alimenté par quelques « affaires » qui ont laissé des traces douloureuses des deux côtés. Lorsque l'on analyse le point de vue de chacune des deux parties, on constate que ces conflits trouvent principalement leur source dans les différences de culture entre les fonctionnaires d'autorité d'une part, et d'autre part les chercheurs « fondamentaux » et les universitaires.

S'y ajoutent les différences de culture professionnelle entre d'un côté les enseignants-chercheurs des universités et les chercheurs des laboratoires de type CNRS, qui se consacrent prioritairement à la recherche dite « fondamentale », et de l'autre les chargés d'études et les consultants indépendants, qui font de la recherche dite « appliquée » ou « finalisée ». Ces différences, peu perceptibles chez les chercheurs anglo-saxons, sont fortement marquées chez les chercheurs français.

L'essentiel de ces différences, tel qu'il apparaît à travers les témoignages recueillis au cours de cette mission, peut être résumé sous la forme des trois alternatives suivantes : 1) aider à la prise de décision, ou contribuer à l'avancement de la connaissance ; 2) servir l'État, ou servir la communauté scientifique ; 3) s'adresser aux seuls responsables, ou à l'ensemble des citoyens.

5.1. Aider à la décision, ou contribuer à l'avancement de la connaissance scientifique

Bien que l'histoire des sciences humaines offre des exemples de découvertes importantes faites en tentant d'apporter une réponse opérationnelle à un problème concret, dans la pratique la culture professionnelle des chercheurs « fondamentaux » français s'oppose sur ce point à celle des chercheurs « appliqués », et a fortiori à celle des décideurs privés ou publics. Il arrive que ces derniers se plaignent que les fonds versés par eux pour une recherche finalisée aient surtout servi à financer des recherches fondamentales sur les thèmes de prédilection des chercheurs (parfois de l'aveu même de ceux-ci), au détriment des préconisations attendues par les bailleurs de fonds.

Ainsi, il n'est pas rare au bout du compte qu'une recherche visant en principe à la proposition de solutions concrètes aboutisse en fait à la production d'un rapport dont les neuf dixièmes sont une synthèse de résultats déjà largement connus des spécialistes (selon le modèle académique du bilan de connaissances), tandis que les préconisations attendues se réduisent à un vingtième du texte environ (compte tenu de la place dévolue à la conclusion d'ensemble).

La modicité des ressources allouées aux laboratoires de sciences humaines et sociales, comparée par exemple au coût des enquêtes extensives, permet de comprendre pourquoi certains chercheurs estiment être dans la nécessité de recourir à des financements extérieurs pour poursuivre leurs travaux fondamentaux. Cependant, lorsque le rapport fourni au commanditaire ne répond pas aux termes du contrat qu'il a passé avec le chercheur, il n'y a pas lieu de s'étonner de son mécontentement. Cette forme d'insatisfaction nourrit naturellement les préventions à l'égard des chercheurs.

Une telle mésaventure ne se produit généralement pas lorsqu'il est fait appel à des professionnels de la recherche appliquée ; mais (s'agissant d'associations à buts lucratifs, d'instituts privés, ou de consultants indépendants) leurs tarifs plus élevés que ceux des chercheurs du secteur

public peuvent se révéler dissuasifs (puisqu'ils incluent évidemment les salaires et les frais généraux).

5.2. Servir l'État, ou servir la communauté scientifique

Ces deux termes ne sont pas nécessairement antinomiques. Toutefois, en cas de conflit entre la réserve qu'impose le service de l'État et le souci de diffuser la connaissance scientifique, le fonctionnaire d'autorité penche naturellement vers la réserve, tandis que le chercheur peut juger légitime une diffusion large et rapide de ses travaux.

Le problème de la propriété intellectuelle des textes produits illustre bien l'opposition de ces deux cultures. Un universitaire ou un chercheur du CNRS considère que les écrits qu'il produit lui appartiennent en propre, et que toute reproduction ne mentionnant pas la source (et donc l'auteur) est un plagiat ; elle est par conséquent stigmatisée comme telle. Par contre, il est tout à fait habituel qu'un fonctionnaire d'autorité signe personnellement un texte rédigé par l'un de ses subordonnés ; ce faisant, il avalise les idées exprimées et en endosse la responsabilité.

Ces divergences éthiques dépendent évidemment des « groupes de référence » respectifs de ces deux types d'acteurs. Elles sont actuellement accentuées par deux facteurs. D'une part, l'importance plus grande que l'opinion attache aux informations relatives à la sécurité intérieure confère à celles-ci un poids politique plus grand, et accroît en proportion les risques de polémique à leur sujet. D'autre part, dans le but de « valoriser la recherche », on incite les chercheurs à donner à leur travaux une plus grande visibilité, en portant les résultats de leurs recherches à la connaissance du grand public par l'intermédiaire des médias.

Un exemple récent de ce conflit de cultures est fourni par le rapport rédigé par Alain Bauer, à la demande du Ministre de l'Intérieur, sur l'emploi du temps des policiers. La diffusion dans les médias, sans l'accord du Ministre, d'une partie du contenu de ce rapport a donné lieu à de violentes polémiques. À la suite de cette affaire, la DAPN, qui avait fourni les données utilisées par Alain Bauer, n'a pas souhaité présenter, dans le cadre de cette mission, les fichiers dont elle dispose sur les personnels de la police nationale.

5.3. S'adresser aux seuls responsables, ou à l'ensemble des citoyens

Dans la recherche dite « fondamentale » ou « désintéressée », c'est-à-dire financée sur les fonds publics dans le cadre des laboratoires de recherche ou des universités, on incite les chercheurs à la transparence. Les résultats présentés sont explicitement soumis à la critique des collègues, qui, pour cela, doivent disposer du maximum d'informations sur les conditions de la recherche, les matériaux utilisés, etc. (problème de la « reproductibilité » des recherches). En outre, sans nécessairement tomber dans le travers anglo-saxon du « publish or perish », diverses considérations de carrière poussent les chercheurs à diffuser largement le fruit de leurs travaux. Enfin, de nombreux chercheurs estiment que les résultats de recherches financées sur des fonds publics doivent être accessibles à l'ensemble des citoyens. C'est d'ailleurs la position du CNRS, qui souhaite que toute convention de recherche préserve la liberté de publication des chercheurs.

La culture des chercheurs « fondamentaux » s'oppose d'ailleurs sur ce point aussi à celle des praticiens de la recherche dite « finalisée » ou « appliquée ». En particulier, pour ces derniers, les résultats des recherches qu'ils ont réalisées appartiennent aux organismes, publics ou privés, qui les ont financées ; la propriété exclusive des résultats est la contrepartie du financement perçu. Diffuser ceux-ci sans l'autorisation explicite de ces organismes constituerait une faute professionnelle grave (violation du secret commercial).

Les fonctionnaires d'autorité, auxquels l'obligation de réserve s'impose statutairement, raisonnent naturellement de la même manière. Il n'imaginent pas qu'un rapport de recherche, même établi à la demande d'un service public, fasse l'objet d'une diffusion externe, a fortiori d'une publication, sans l'accord de ce service. Il est d'ailleurs arrivé que des chercheurs « fondamentaux » n'aient apparemment pas prévu les conséquences politiques d'une diffusion dans les médias de résultats destinés à demeurer (au moins pour un certain temps) confidentiels ; ce faisant, ils ont porté préjudice aux autres chercheurs travaillant dans le même champ, qui ont vu se fermer l'accès aux données dont ils avaient besoin.

6. Éléments de solution

L'ensemble des considérations qui précèdent conduit naturellement à préconiser la clarification des relations entre les autorités détentrices des fichiers statistiques et les chercheurs désireux d'en tirer parti.

Cette clarification pourrait comporter trois volets :

1) informer chacune des parties sur les valeurs et les références culturelles de l'autre partie ;

2) établir une charte détaillée servant de cadre à toute convention de recherche, définissant clairement les termes de la coopération entre le chercheur et le détenteur des données ;

3) prévoir une commission d'arbitrage et des sanctions en cas de manquement aux termes de la charte.

Liste des personnes contactées

Ambach Yann, Direction Générale des Douanes et Droits Indirects, Bureau de la politique des contrôles et de la lutte contre la fraude.

Berthe Alain, Sous-Directeur chargé des liaisons extérieures, Direction Centrale de la Police Judiciaire.

Boucard Christian, Inspecteur Principal, Direction Générale des Douanes et Droits Indirects, Bureau de la politique des contrôles et de la lutte contre la fraude.

Caillou Michel, Commissaire Divisionnaire, Direction Générale de la Police Nationale.

Casanova Gilles, Conseiller Technique auprès du Ministre de l'Intérieur (*).

Champon Michel, Sous-Directeur à la Direction de la Sécurité Civile.

Dechamp Claude, Colonel, Chargé de mission du Général Chef du service des opérations et de l'emploi, Direction Générale de la Gendarmerie Nationale.

Declerck Bernard, Adjudant, Service Technique de Recherches Judiciaires et de Documentation, Direction Générale de la Gendarmerie Nationale.

Descombes Gilbert, Direction de la Sécurité Civile.

Dillies Laetitia, statisticienne, Institut des Hautes Études de la Sécurité Intérieure.

Gravet Bernard, Directeur Central de la Police Judiciaire.

Herr Christian, Lieutenant-Colonel, Bureau de la Police Judiciaire, Direction Générale de la Gendarmerie Nationale.

Kerskens Jean-Claude, Major, Service Technique de Recherches Judiciaires et de Documentation, Direction Générale de la Gendarmerie Nationale.

Labrousse Francis, Commissaire Divisionnaire, Conseiller Technique auprès du Préfet de Police de Paris.

Leffondré Daniel, Commandant, Service Technique de Recherches Judiciaires et de Documentation, Direction Générale de la Gendarmerie Nationale.

Legentil Alain, Colonel, Sirpa-Gendarmerie, Direction Générale de la Gendarmerie Nationale.

Lemercier Daniel, Colonel, Service des Ressources Humaines, Direction Générale de la Gendarmerie Nationale.

Melchior Philippe, Inspecteur Général de l'Administration, Directeur de l'Institut des Hautes Études de la Sécurité Intérieure.

Monjardet Dominique, Directeur de Recherche au CNRS, Direction Centrale de la Sécurité Publique (*).

Moreddu François, Direction de la Sécurité Civile.

Penet Alain, Adjudant, Service Technique de Recherches Judiciaires et de Documentation, Direction Générale de la Gendarmerie Nationale.

Pottier Marie-Lys, statisticienne, Centre d'Études Sociologiques sur le Droit et les Institutions Pénales.

Reiller Jacques, Préfet, Directeur du Centre d'Études et de Prévision (Ministère de l'Intérieur).

Watin-Augouard Marc, Colonel, Chef du SIRPA-Gendarmerie, Direction Générale de la Gendarmerie Nationale.

(*) Entretien téléphonique seulement.

2. Les géographes et leur utilisation des recensements

Alexandre Kych (Lasmas-IdL)

La Cnil a posé pour le RP99 des limites très fortes à l'utilisation des fichiers du recensement pour les données infra-communales. La limite à 50 000 habitants rend particulièrement difficile sinon impossible le travail des géographes pour toute une série de recherches.

En première approximation, on peut distinguer quatre cas d'utilisation des recensements par les géographes, qui sont résumés dans le tableau ci-dessous :

Champ	*Zonages*	*Données agrégées spatialement*	*Données individuelles*
local	Îlots, groupe d'îlots, communes, groupes de communes	Cas 1	Cas 2
national	Emboîtement administratif (arrondissement, département, région) Emboîtement urbain (unité urbaine, ZPIU) Autres zonages (zone d'emploi, région agricole) Autres typologies (taille d'unité urbaine, autres)	Cas 3	Cas 4

Le cas 1 (données agrégées au niveau local) est de très loin le plus pratiqué. Les données sont obtenues de différentes manières. Ainsi pour l'exploitation du RP90, le géographe répond à la demande d'une collectivité ou d'une institution locale qui lui fournit les fonds pour acheter les données à la DR de l'Insee ou les lui cède après les avoir acquises pour son propre usage (commune, communauté urbaine, agence d'urbanisme...) ; il peut aussi recourir à un accord particulier (convention particulière entre l'Insee et une université, par exemple) ; il peut encore bénéficier de bonnes relations personnelles avec la DR de l'Insee et payer par des publications ; enfin, autre exemple, un accord avec la Cnil a permis un travail très fin spatialement dans l'agglomération de Rouen.

Le cas 3 (données agrégées au niveau national) est sensiblement moins pratiqué. Cela nécessitait jusqu'à il y a peu de temps de grandes ressources informatiques. Cette approche de l'espace attire aussi moins les géographes. On peut citer le GIP RECLUS et le groupe PARIS.

Le cas 2 (données individuelles au niveau local) se rencontre très rarement. Pendant longtemps il a nécessité des ressources et des compétences informatiques plus rares. Plusieurs équipes de la région parisienne y ont cependant recours.

Le cas 4 (données individuelles au niveau national) est encore plus rarement pratiqué en cumulant les raisons évoquées dans les deux cas précédents.

Les deux clivages qui sont à la base du tableau à quatre cases ci-dessus renvoient schématiquement à des façons différentes de pratiquer la géographie. La distinction ancienne et institutionnelle entre géographie régionale et géographie générale fonde l'opposition local/national. Il y a un autre partage entre ceux qui considèrent que ce sont les lieux en tant que tels qui sont les objets de la géographie et ceux qui s'attachent plutôt aux individus dans des lieux. Pour les premiers, l'utilisation de fichiers agrégés leur semble une évidence et n'est que le préalable à des typologies, des regroupements, des cartographies et des modélisations. Les seconds ne sauraient se passer des fichiers individuels en particulier pour redéfinir les catégorisations des individus, des couples, des familles et des ménages, et fournir par exemple la répartition spatiale des différentes catégories.

En dehors des façons de faire la géographie, jusqu'au RP82, le choix entre les données agrégées et les données individuelles n'est qu'une affaire de ressources et de compétences informatiques d'une part et de ressources financières d'autre part.

À partir du RP90, si les raisons de choix entre les données agrégées et les données individuelles changent peu, la suppression – pour des raisons liées à la protection de la confidentialité des informations recueillies auprès des personnes – de la possibilité d'accéder à l'îlot et aux autres niveaux géographiques infra-communaux (que les données soient agrégées ou individuelles) frappe de plein fouet les géographes qui travaillent sur des champs locaux.

La principale raison évoquée pour travailler au niveau de l'îlot est que l'îlot est l'unité spatiale homogène la plus élémentaire : la brique ou l'atome du spatial. Sans accès à la connaissance de l'îlot, toute étude de ségrégation spatiale est impossible et toute comparaison d'un RP à l'autre l'est également. Le premier argument suscite les réserves de certains : un îlot est-il plus homogène qu'un quartier ou une commune ? L'homogénéité ne serait-elle pas plus grande par rue ou par étage ? Le second argument est plus convaincant : les îlots et les communes ont des limites assez stables d'un recensement au suivant, mais c'est beaucoup moins vrai des agrégats intermédiaires, surtout si ceux-ci sont amenés à assurer des effectifs de population minimums et que les variations de populations sont très contrastées.

Le découpage par îlots souffre d'une seconde faiblesse. Certains îlots peuvent avoir des effectifs très ténus et toute tabulation, sur le quart des personnes recensées de surcroît, devient suspecte. Alors que penser de ceux qui veulent des tabulations, au niveau de l'îlot, fines et correspondant à plusieurs centaines de cases ?

On peut évoquer alors un troisième aspect du recours à l'îlot : la force de l'habitude, retrouver d'un recensement à l'autre les mêmes tableaux pré-définis (cf. les antiques pré-imprimés) sans se demander si quelques tabulations créées ad hoc dans de nouveaux niveaux spatiaux ne feraient pas mieux l'affaire.

En revanche, il y a un domaine où l'accès à l'îlot semble incontournable : c'est tout ce qui concerne les systèmes d'informations géographiques (SIG), en particulier quand ils sont construits pour la gestion des équipements locaux. Avec quoi mettre en relation les réseaux de voirie et d'adductions ou le parcellaire cadastral, sinon avec les îlots, qui sont déjà bien grossiers ?

Quoi qu'il en soit, en particulier dans le cadre de recherches très attachées à un cadre local, l'îlot apparaît indispensable, mais – et la pratique le prouve depuis des années et donc bien avant les nouvelles dispositions de l'Insee – il suffit de travailler en collaboration avec les collectivités locales pour y accéder. Il ne faut cependant pas oublier que la collaboration avec des collectivités locales a ses limites : quand les collectivités locales n'ont pas de demande ou quand les zones étudiées ne correspondent pas à une collectivité. (Comment étudier un espace de plusieurs dizaines de communes quand il n'y correspond aucun regroupement institutionnel de communes, syndicat, communauté ou autre ?)

Le passage au RP99 va apporter un nouveau bouleversement. C'est l'accès aux données individuelles qui est fortement remis en cause. Il y a désormais deux fichiers individuels disponibles : celui des logements et celui des personnes. Le fichier des logements est soumis aux mêmes critères dans le choix des zonages spatiaux fins que les fichiers agrégés, ceci a été évoqué plus haut. L'accès au fichier des personnes n'est possible que si l'on ne peut identifier aucun territoire d'un seul tenant de moins de 50 000 h. C'est rendre impossible le travail de tous ceux qui cherchent à créer leur propres catégories spatiales d'observation : littoral, unités urbaines dans leur définition de 1962, communes touristiques ou toutes autres typologies. Car dans tous ces cas, on part de l'identification d'une commune pour en enrichir la description par la recherche de caractères complémentaires dans des bases de données externes au RP. Sans l'identification de la commune, on ne peut plus rien faire. De surcroît la nouvelle règle du jeu porte en elle une menace supplémentaire : comment, à l'instar des exploitations de tous les RP précédents, accéder à un zonage du type « tranches de taille d'unités urbaines en 8 postes » par départements ? Pour près de la moitié des départements il faudra y renoncer. Et surtout, si un chercheur veut pouvoir tabuler par département et par taille, il devra avoir deux extractions du RP, l'une avec les départements et l'autre avec les tailles, la présence simultanée des deux caractères pouvant amener à passer sous la

barrière des 50 000 h. Multiplier les extractions ou demander à l'Insee à chaque fois une tabulation sont deux solutions aussi inapplicables l'une que l'autre. Enfin, si l'on veut travailler finement à l'intérieur des ménages par un travail particulier de recodage, il faut quasiment renoncer à tous descripteurs spatiaux.

Cette nouvelle disposition empêche une grande partie des recherches qui utilisaient le fichier individuel du RP. Elle comporte encore bien des obscurités : comment utiliser à la fois des zones spatiales d'un seul tenant et des types de communes (on vient d'en voir un exemple), mais que faire aussi avec les communes de résidence antérieure et les communes de lieu de travail, en particulier quand elles coïncident avec les communes de résidence au recensement, leur connaissance sera-t-elle aussi interdite ?

Si l'on ne revient pas aux dispositions en usage pour le RP90, on peut imaginer plusieurs procédures dont aucune n'est vraiment satisfaisante.

1° Multiplier les demandes de tabulation à l'Insee. À quels coûts et dans quels délais ? Est-ce possible, dans la mesure où la recherche implique souvent de passer par plusieurs tabulations avant d'arriver au bon tableau ? L'Insee assure que les chercheurs pourront disposer de jeux d'essai pour élaborer leurs recodages et construire leurs tabulations.

2° Faire une demande d'autorisation auprès de la Cnil pour chaque extraction à partir du fichier individuel des personnes. La Cnil est-elle disposée à une telle procédure et quels en seront les délais ?

3° La création par l'Insee d'un fichier individuel anonymisé à un faible taux de sondage (1/100 ou 2/100) comme cela se fait en Angleterre, voire comme il l'a fait lui-même pour des recensements plus anciens (sans les anonymiser cependant). Un tel fichier, malgré son intérêt, ne résoudra pas le problème de ceux qui travaillent dans des cadres géographique fins ou inhabituels, et empêchera peut-être de créer des recodages fondés sur la confrontation des caractères des différents individus d'un même ménage. Sans compter qu'en construisant certaines sous-populations, on risque d'atteindre assez vite des effectifs trop petits.

4° La création par l'Insee d'un système automatique de tabulations à la demande, qui permettrait des recodages, dans lequel les descripteurs spatiaux ne seraient que des variables comme les autres et qui veillerait à ce qu'un effectif minimum soit présent dans chaque case de la tabulation pour en assurer la confidentialité. Des dispositions de ce type existent dans d'autres pays (cf. le Royaume-Uni).

Le cœur du problème n'est-il pas dans le fait que la Cnil assimile le respect de la confidentialité à la connaissance de la localisation. *La loi dit qu'aucune décision ne doit pouvoir être prise sur une personne à partir d'informations collectées dans le cadre d'un traitement automatisé d'information visant à définir son profil. La Cnil considère qu'une personne appartenant à une petite zone géographique peut être caractérisée grâce au profil de la zone qu'elle habite, tiré des résultats du recensement*[2]. La loi date de 1978, mais la Cnil n'a pas toujours eu la même position, comme en témoigne le RP82. L'Insee non plus n'a pas toujours eu la même position et il montre une prudence peut-être excessive avec le seuil de 50 000 h.

[2]. Cnis, La diffusion du recensement de population de 1999, Réunion du 29 mai 1997. p.12.

En conclusion, il faut souligner l'aspect paradoxal de l'usage du RP mis en place pour 99. Certains l'ont déjà dit, le RP est la grande enquête que l'Insee consacre à l'espace. Dans la mesure où la confidentialité et la protection des personnes sont fondées essentiellement sur la connaissance de la localisation des personnes, l'accès aux caractères spatiaux des individus du RP sera extrêmement laborieux (voire impossible pour beaucoup de chercheurs travaillant sur des espaces très fins ou non standards).

La création d'une zone du secret, comme cela commence à exister à l'étranger (à Statistique Canada par exemple), apparaît comme une solution appropriée.

3. L'accès aux données sur les entreprises du point de vue des sociologues

Emmanuel Lazega, Lise Mounier (Lasmas-IdL)

Pour les économistes, la France dispose de bases d'enquêtes sur les entreprises parmi les meilleures du monde, en termes de couverture (peu d'omissions, peu de répétitions), de contenu et de codification (informations enregistrées), bien que les relations entre statistiques et comptabilité d'entreprise y soient aussi notoirement faibles que dans le reste du monde. L'organisation de la recherche et la production de connaissances sur les entreprises en France est dominée par l'Insee dont sont issus (et qui rassemble) les plus gros fichiers statistiques. D'autres ministères produisent leurs propres enquêtes, comme le ministère de l'Industrie qui dispose de son propre service statistique (le Sessi), ou même d'autres institutions (par exemple, la banque de France a son propre panel d'entreprises). Dans le respect des règles de confidentialité et du cadre réglementaire qui s'impose à tous, ces données ne sont pas facilement mises à disposition ; et lorsqu'elles le sont, c'est sous forme anonymisée, sous convention de recherche – le nombre de conventions étant limité par la volonté de suivre de près les travaux des chercheurs extérieurs. Mais les fichiers principaux, dont *l'Enquête Annuelle d'Entreprises (*EAE*)* et *SIRENE* (le recensement des entreprises), la *Base d'Analyse Longitudinale (*BAL*)*, *Liaisons financières (*LIFI*)*, *SUSE* (données fiscales et données de l'EAE) sont à l'Insee (ou rassemblées par l'Insee). De nombreuses informations sur les organisations participant au système productif français sont dispersées dans des institutions ad hoc, comme par exemple l'Institut national de la propriété intellectuelle (qui vend très cher l'information sur les brevets déposés par les entreprises et organisations de recherche publique). Les universités, contrairement à ce qui se passe aux États-Unis, n'ont qu'un rôle négligeable dans la production de données statistiques sur les entreprises. À ces données publiques, il faut ajouter les fichiers de données privés comme celles de Kompass[3] (dont sont facilement extraites, par exemple, les informations sur les relations interlock entre conseils d'administration).

L'Insee dispose de données administratives, en complément des données d'enquête, qu'il n'a pas le droit de mettre à disposition des chercheurs universitaires. En général, l'accès aux données d'enquête sur les entreprises est difficile – en tout cas pas immédiat – parce que l'anonymat est souvent difficile à conserver[4] ; les fichiers peuvent être mis en concordance et procurer à certains utilisateurs commerciaux des avantages concurrentiels sur d'autres. Pour ces raisons, les fichiers Insee et Sessi ne sont le plus souvent accessibles que par le dépôt d'un projet de recherche et d'une demande d'accès aux données individuelles auprès de la Cnil (à moins de collaborer personnellement avec un chercheur membre de ces institutions). Lorsque le chercheur et universitaire obtient du Comité du secret statistique l'autorisation d'accéder aux données individuelles qui l'intéressent, il paie un droit de mise à disposition. Il ne peut ni les communiquer à des tiers, ni en faire état dans ses relations avec les entreprises ; il s'engage aussi à respecter les règles du secret statistique dans les publications qui seront faites de ses travaux[5] ; enfin il s'engage à détruire les données individuelles une fois l'étude achevée. Un représentant du patronat et un représentant de la Direction des Statistiques d'Entreprises siègent au Comité du secret statistique chargé d'instruire les demandes d'utilisation de bases de données par des universitaires et autres. Il faut noter aussi que la difficulté d'accès n'est pas seulement liée à l'anonymat. L'Insee a maintenant des chercheurs de très haut niveau qui travaillent sur des enquêtes originales et qui ont tendance à ne céder le droit d'accès qu'une fois les fichiers exploités. On pense par exemple aux données sur la concentration verticale et les enquêtes sur les têtes de réseaux d'entreprises dans l'habillement et dans le bricolage menées par la Division Commerce de l'Insee. Notre expérience du Sessi est qu'il

[3]. Des équipes peu nombreuses ont construit et mettent à jour leurs propres fichiers (par exemple celle de François Morin à Toulouse, qui redescend jusqu'aux rapports d'activités des grands groupes français ou européens).

[4]. Il n'est pas difficile de deviner quelle est l'entreprise qui fabrique des pneus à Clermont-Ferrand.

[5]. Par exemple, aucun résultat relatif à un groupe de moins de trois entreprises ou à un groupe d'entreprises dont une seule représente plus de 85 % du résultat total ne peut être diffusé.

délivre gratuitement aux universitaires des extraits de fichiers (moyennant un rapport et la rédaction d'un quatre pages à la fin du délai accordé pour la recherche).

En tant qu'universitaires, nous n'avons pas l'expérience d'une quelconque participation ou consultation en matière de conception d'enquêtes nouvelles. Or ce qui intéresse vraiment la sociologie économique ou la sociologie quantitative des organisations dans l'analyse secondaire des bases de données sur les entreprises en France, en particulier sur les relations inter-entreprises, c'est l'identification de diverses formes de discipline sociale sous-jacentes aux échanges de ressources ayant cours dans l'appareil de production. Les sociologues partent du principe que les entreprises opèrent dans des régimes d'interdépendances de ressources (pour la production) et de marchés contraints (pour les échanges). Avec de bonnes données, pour l'instant très difficiles à exiger des entreprises (ou que l'Insee n'a pas le droit de rétrocéder, comme dans le cas de données dites administratives, qui ne sont pas des données d'enquête), la contribution de cette discipline sociale à la productivité et à la performance de ces entreprises (ou grands groupes d'entreprises) devrait être mesurable. Des enquêtes à grande échelle allant dans ce sens devraient être possibles à mener. En France, pour l'instant, l'Insee ou le Sessi (ou un centre de recherche équivalent) sont apparemment seuls à avoir l'envergure nécessaire pour entreprendre dans ce sens. Ils restent cependant tributaires de la coopération des entreprises pour lesquelles la charge statistique est toujours lourde. Il n'en reste pas moins qu'une telle connaissance est de nature à éclairer, non seulement les chercheurs, mais aussi les institutions sociales et les décideurs publics et privés.

Le manque d'informations quantitatives intéressant le sociologue de l'économie n'est cependant pas seulement dû aux difficultés de rapprochement avec les grandes institutions d'économistes. Il n'est pas nécessaire de faire un sondage spécifique pour se rendre compte que les enquêtes sur les entreprises sont peu utilisées par les sociologues universitaires ou CNRS. Deux raisons au moins contribuent à ce fait. Premièrement, il n'existe pas en France de tradition quantitative forte en sociologie économique procédant par une approche organisationnelle ; or les fichiers entreprises sont souvent complexes. Deuxièmement, ces enquêtes sont pensées et conduites par des économistes soumis à des contraintes administratives et politiques (alléger la charge statistique pesant sur les entreprises en réduisant la taille des questionnaires envoyés par l'Insee). Par exemple, même lorsque les thématiques se rapprochent des préoccupations des sociologues, comme c'est nettement le cas pour l'enquête *Liaisons industrielles* (LI) du SESSI, les données recueillies adoptent comme unité d'analyse les relations contractuelles de l'entreprise focale avec ses sous-traitants « en général » ; ce niveau d'agrégation interdit au sociologue de descendre au niveau d'analyse qui l'intéresse véritablement, c'est-à-dire les relations (contractuelles, sociales) spécifiques entre deux unités bien spécifiées. Autrement dit, l'unité d'analyse reste l'entreprise focale, mais L.I. ne connaît pas les entités avec lesquelles la première est en relation ; ses relations avec les autres entreprises ne sont pas décrites de manière désagrégeable. L'analyse multiniveaux au sens de la statistique des réseaux sociaux n'est donc pas possible sur ce fichier. Il n'est pas possible d'interroger les données pour établir l'existence ou l'absence d'arrangements contractuels d'un type spécifique s'accompagnant d'une discipline sociale, elle-même spécifique, qui offre des garanties pour la prise de risque ou la mise en œuvre des règles contractuelles. Ces fichiers ne recueillent donc pas toujours des informations immédiatement et directement utilisables par des sociologues.

Il n'en reste pas moins que les travaux de recherches menés à l'Insee sur la base de ces données se rapprochent des préoccupations des sociologues de l'économie qui utilisent une entrée intra- et inter-organisationnelle. On pense ici à nouveau essentiellement aux travaux sur les groupes et les relations intra-groupes parus sur la base de l'enquête LIFI et l'accent qu'elle met sur le contrôle capitalistique par une holding tête de groupe. Ou encore aux travaux sur les têtes de réseaux dans l'habillement ou l'enquête sur la commercialisation d'articles de bricolage par la Division Commerce de l'Insee. Il va de soi que d'autres recherches sont menées à l'Insee, notre objectif n'est pas d'en faire un recensement.

Données européennes

À notre connaissance, l'Insee ne communique pas à Eurostat ses bases de données entreprises. Outre les problèmes de confidentialité, la statistique d'entreprise posant des problèmes conceptuels et méthodologiques spécifiques de définition d'unité, la manière dont les pays européens gèrent les répertoires d'entreprises, les sources de statistique sur la démographie des entreprises (créations et cessation), leur maintenance, les usages auxquels ils sont associés varient considérablement. À l'échelle européenne, il y a peu d'homogénéité dans les bases statistiques (bien que sous des appellations différentes peuvent se dissimuler des entités semblables) et dans les politiques de recueil, bien que l'harmonisation comptable soit à l'ordre du jour. Historiquement, chaque pays a son organisation industrielle, sa réglementation sociale et fiscale, qui ont fondé les fichiers administratifs dont ils sont issus. Ce manque d'harmonisation se traduit par une certaine pauvreté des données d'Eurostat et des programmes de comparaisons internationales. Pour l'instant, à notre connaissance, les données Eurostat sont des données économiques agrégées qui ne peuvent pas servir à l'usage intensif auquel les destineraient les sociologues de l'économie.

Annexe IV : Production de données pour la recherche, quelques exemples

1. Note sur la programmation éventuelle d'une nouvelle enquête Formation - Qualification Professionnelle

Annick Kieffer, Louis-André Vallet (Lasmas-IdL)

L'Insee, à l'occasion de l'établissement de son programme d'enquêtes à moyen terme, s'interroge sur l'opportunité de produire une nouvelle enquête Formation - Qualification Professionnelle. Afin de nourrir le débat sur cette question et en vue d'éclairer leur choix, les responsables ont souhaité consulter les utilisateurs intensifs de cette enquête que sont les chercheurs en sciences sociales. Telle est la demande qui a présidé à l'établissement de cette note.

Le Lasmas-Institut du Longitudinal a informé les laboratoires CNRS, utilisateurs de ces enquêtes par son intermédiaire, des différentes options actuellement envisagées par l'Insee et leur a demandé d'exprimer leur point de vue. Il a également sollicité le MZES (Université de Mannheim) dont de nombreux travaux comparatifs s'appuient, s'agissant de la France, sur les enquêtes FQP que ce centre de recherches a lui-même acquises. *Il se dégage de l'examen de ces réponses une demande explicite en faveur du maintien et de la continuation de la série d'enquêtes*. Nous présentons ci-dessous les principaux arguments avancés.

*** Une série d'enquêtes originale et enviée au plan international**

La première enquête FQP a été réalisée en 1964[6]. Elle a été suivie de quatre autres conduites selon une périodicité assez régulière (1970, 1977, 1985 et 1993) et dans un souci de comparabilité, tant du point de vue du protocole d'enquête que de celui du contenu du questionnaire et des indicateurs élaborés. Dès l'enquête sur l'emploi de juin 1953, Jacques Desabie avait introduit trois questions supplémentaires permettant l'examen de la mobilité professionnelle et sociale. La série des enquêtes FQP a prolongé et amplifié cette tentative pionnière. Elle constitue ainsi une source française majeure pour l'étude des phénomènes de mobilité sous leurs différentes facettes (inter-générationnelle, intra-générationnelle et géographique) comme de leurs rapports avec la formation, initiale ou continue, des individus, le statut d'occupation et la position professionnelle qu'ils ont atteints et le salaire qu'ils obtiennent. L'enquête fournit, sur chacun de ces points, une description qui, au fil du temps, est devenue très détaillée. Elle a donc vivement intéressé les spécialistes de l'éducation comme ceux de la mobilité, du travail et des qualifications, sans oublier les chercheurs à l'interface de ces deux domaines (relations entre formation initiale, formation continue et emploi, rentabilisation des diplômes sur le marché du travail). *Elle a de ce fait été très fortement utilisée et continue à l'être, tant par les sociologues que par les économistes, les démographes et les géographes*.

La taille de l'échantillon (hormis l'enquête de 1993 qui peut poser quelques problèmes à cet égard), le mode d'interrogation (c'est l'individu qui répond pour lui-même) et la continuité de la série ont contribué au succès de l'enquête, puisqu'elle a permis par exemple l'étude fiable de sous-populations particulières et qu'elle autorise aussi l'élargissement de l'analyse par la construction de pseudo-panels. La complexité du plan de sondage des différentes enquêtes, si elle a rendu malaisé le calcul exact d'intervalles de confiance, n'a cependant jamais conduit à douter de la validité des résultats produits à partir des données collectées.

Il faut insister sur le fait que, dans le champ qui est le sien, *cette série d'enquêtes au questionnaire fourni et renseigné pour un vaste échantillon, constitue sur le plan international un outil rare et envié*. Ni l'Allemagne, ni la Grande-Bretagne, ni les États-Unis pour ne citer que quelques pays ne disposent d'une source équivalente si l'on prend en compte à la fois l'ancienneté de la série, sa régularité temporelle, la qualité de l'information collectée et le nombre d'individus interrogés. Seule peut-être la Hongrie disposerait de données d'une qualité équivalente.

[6]. Il semblerait que le fichier de cette enquête ne soit aujourd'hui plus lisible à l'Insee. Elle avait été acquise dès 1968 par le Centre d'Études Sociologiques et trois tables SAS sont donc disponibles au Lasmas-IdL entre lesquelles un problème réel d'appariement existait néanmoins. Un travail a consisté à réunir la documentation relative à cette enquête – avec l'aide de M.-A. Estrade – et à résoudre ce problème d'appariement. Un fichier utilisable et bien documenté est disponible pour les utilisateurs de la série d'enquêtes (Degenne, Lebeaux et Vallet, 1998, document de travail, Lasmas).

*** Une collaboration féconde entre statisticiens et chercheurs en sciences sociales**

L'enquête FQP a été la première enquête de l'Insee achetée par le CNRS. Enquête à visée sociologique, mais également importante – on l'a déjà souligné – pour les géographes et les démographes, *elle a été l'occasion d'échanges, de réflexions et de collaborations entre chercheurs et statisticiens de l'institut national*. Cela a par exemple été le cas dans les domaines de la mobilité sociale, de l'évolution des inégalités géographiques et sociales d'éducation, de la transformation des catégories socioprofessionnelles et des classes sociales et, plus généralement, du problème de la catégorisation des groupes sociaux dans une société en forte évolution. *L'enquête FQP a ainsi joué – et joue encore – un rôle important dans la diffusion progressive d'une culture commune aux deux communautés professionnelles*. Cet aspect est lui aussi original car on en trouve peu d'exemples à l'étranger.

Ces rapports plus étroits se sont concrétisés de diverses manières. Une journée d'études « Sociologie et statistique » a notamment été organisée conjointement par la Société française de sociologie et l'INSEE en octobre 1982 (cf. *Économie et Statistique*, 1984, n° 168). D'un point de vue plus opérationnel, le CNRS a souhaité organiser un retour vers l'Insee des travaux, usages et problèmes rencontrés par les chercheurs dans leur utilisation des fichiers statistiques. Cet échange a particulièrement bien fonctionné dans le cas des enquêtes FQP : des éléments d'évaluation, des propositions d'enrichissement ou d'amélioration ont à plusieurs reprises été transmis à l'INSEE et, au fil des enquêtes, la participation des chercheurs à la révision du questionnaire est devenue plus régulière.

*** Un outil unique de connaissance des mouvements de long terme de la société française**

Dans le champ qui est le sien – celui des structures éducatives, des structures d'emploi et des structures sociales –, on peut raisonnablement soutenir que la série des enquêtes FQP est la seule qui permette d'étudier l'évolution séculaire de la société française. Les individus les plus âgés interrogés en 1964 ou 1970 ont vu le jour au tournant du siècle. Leurs parents sont nés approximativement dans le dernier quart du XIX^{e} siècle. À supposer qu'une nouvelle enquête FQP soit réalisée en 2003, on disposerait ainsi d'un outil stable d'observation sur près de 40 ans qui serait susceptible de fournir des informations, par appel à la mémoire, sur toute la durée du XX^{e} siècle.

Un élément essentiel de la valeur des enquêtes FQP réside en effet dans leur ancienneté et leur continuité. *La disponibilité croissante de méthodes statistiques de plus en plus puissantes rend aujourd'hui possible la mise en évidence et l'analyse précise d'évolutions de long terme (évolutions structurelles et évolutions dissociables des seules évolutions de la structure sociale) pour des phénomènes dotés par nature d'une très forte inertie et pour lesquels les évolutions ne sont donc que difficilement décelables sur des périodes brèves*. Tel est par exemple le cas de la relation entre origine et position sociales ou encore du lien entre milieu d'origine et niveau d'éducation atteint.

Décider l'interruption du dispositif d'observation marquerait l'arrêt de l'enrichissement d'un capital qui a désormais une valeur historique et rendrait aussi plus malaisée l'utilisation future du capital déjà accumulé. Des questions telles que celles de l'évolution de la fluidité sociale, de la démocratisation de l'enseignement, des transformations de la valeur économique des diplômes (ou des formations) ou de l'accès à certaines professions ou certains statuts (celui d'indépendant par exemple) sont posées de façon récurrente dans le débat social à propos de la société française. Leur examen scientifique doit s'appuyer sur des comparaisons entre les générations et doit souvent articuler les prises en compte de l'âge, de la période et de la cohorte. *Pour toutes ces questions et avec la série des enquêtes FQP, la statistique publique a fourni un outil d'observation stable, fiable, régulier et de long terme, procurant ainsi, au cours du temps, des informations comparables et susceptibles d'être reliées entre elles. Il est permis de penser que le maintien de cette fonction est essentiel*.

Considérons à titre d'exemple le cas de l'éducation. Les enquêtes FQP sont les seules à fournir depuis 1964 une description détaillée du parcours scolaire des individus – tel qu'ils le reconstituent eux-mêmes – et elles procurent aussi une information sur le niveau d'éducation du père (depuis 1970) ou des deux parents (depuis 1977). À l'opposé, les statistiques annuelles du ministère de l'Éducation nationale ne renseignent que sur des flux et des stocks. Les panels d'élèves (1973,

1980, 1989, 1995) procurent des informations sur les parcours effectués dans l'enseignement secondaire, mais les données sur l'origine sociale – basées sur l'information fournie par l'établissement scolaire et/ou sur un questionnaire auto-administré – sont comparativement de moindre qualité et rien n'est à ce jour recueilli qui concernerait le parcours professionnel ultérieur. Les enquêtes Emploi ont certes intégré, depuis une date récente, des informations assez détaillées sur les cursus scolaires, mais les séries précédentes n'informent que sur les diplômes. Il n'est donc guère envisageable de s'interroger, à partir de cette source, sur les rapports entre formation et certification dans une perspective dynamique. Les informations relatives à l'origine sociale y sont en outre limitées : elles ne concernent que la situation professionnelle du père et ignorent ainsi son niveau d'éducation de même que toutes les caractéristiques maternelles.

La présence, dans un questionnaire statistique, d'un recueil d'informations détaillées relatives aux mères et aux conjointes est au demeurant rare et cela constitue encore une spécificité des enquêtes FQP. Or, les évolutions sociales en cours conduisent naturellement les sociologues à se préoccuper de la construction d'indicateurs assez synthétiques qui permettent de caractériser, de façon plus complète que par la seule personne de référence, la position sociale des familles. Dans le champ de la mobilité sociale, la prise en considération des femmes a été aussi un aspect exploré assez récemment sur le plan international. Dans tous ces travaux et s'agissant de la France, les enquêtes FQP ont été – et seront encore – une source majeure d'informations. On peut citer aussi les recherches sur le choix du conjoint et sur le rôle de la taille de la fratrie dans le niveau d'éducation atteint et la carrière accomplie. L'étude de la mobilité professionnelle sur une période de cinq ans ou depuis l'entrée dans la vie active constitue encore une spécificité des enquêtes FQP.

Ce n'est pas seulement dans l'étude d'évolutions macroscopiques de la société globale, mais aussi dans l'analyse des traits et transformations de certaines professions et groupes professionnels que les enquêtes FQP ont rendu des services importants. On peut penser ici, entre autres recherches, aux travaux sur les artisans (B. Zarca), les employés (A. Chenu) ou les métiers du BTP (M. Dadoy) qui ont su tirer profit de la richesse des informations relatives à la formation initiale et continue jointe à celles concernant le statut, la fonction, la profession et la qualification.

Les enquêtes FQP ont enfin beaucoup apporté au domaine des comparaisons internationales. Cela a été le cas pour l'étude de l'évolution comparée de la sélectivité interne et externe des systèmes éducatifs, pour l'examen des conséquences de l'expansion du système éducatif sur les positions atteintes dans le marché de l'emploi et, plus généralement, pour la recherche sur les processus de stratification sociale. Les enquêtes FQP ont ici permis des comparaisons dans le temps et l'espace et l'on pourrait citer les noms d'universitaires étrangers de réputation internationale – allemands, américains, anglais ou suédois – qui ont utilisé, et parfois même acquis, telle ou telle de ces enquêtes.

*** Inconvénients et risques liés à une interruption de la série**

Un certain nombre d'informations fournies par les enquêtes FQP – mais non la totalité d'entre elles – sont aujourd'hui disponibles dans d'autres sources statistiques, en particulier dans les enquêtes complémentaires aux enquêtes sur l'emploi (insertion professionnelle, suivi des carrières, mobilité résidentielle, conditions de travail notamment). *Cela pourrait constituer un argument important en faveur de la suppression pure et simple de la série d'enquêtes, mais il convient alors d'insister sur les risques de perte d'information qui seraient alors encourus.*

Les sources statistiques que l'on pourrait imaginer substituer à l'enquête FQP ne paraissent en effet guère susceptibles de recueillir, dans un protocole unique, l'ensemble des informations disponibles dans la version actuelle. Il y a ainsi un risque de perte de cohérence. Ces sources statistiques ne sont en outre pas guidées au premier chef par le souci de l'étude de mouvements de long terme. Centrées sur l'étude conjoncturelle de problèmes définis, elles ne sont pas conçues pour l'analyse dans la durée d'évolutions sociales par nature complexes. Un éclatement des différentes parties de l'enquête FQP en enquêtes autonomes, complémentaires ou non à l'enquête Emploi, aboutirait inéluctablement à l'éclatement des champs à un moment où la progression de la modélisation statistique permet au contraire de mieux traiter les interactions entre phénomènes. Il pourrait aussi constituer un obstacle à la comparabilité temporelle et donc à la cumulativité des résultats.

*** Périodicité, taille et coût de l'enquête**

L'intervalle temporel entre deux enquêtes FQP a varié dans le passé de 6 à 8 ans. L'argumentation développée plus haut conduit à souligner la valeur de cette enquête pour l'étude de phénomènes naturellement inscrits dans la durée et pour l'analyse d'évolutions de long terme. Dans cette perspective, un certain accroissement de l'intervalle temporel associé au maintien d'un cadre d'enquête stable ne serait en rien préjudiciable. *Du point de vue des chercheurs en sciences sociales, la réalisation de l'enquête FQP pourrait ainsi, sans inconvénient majeur, adopter une périodicité décennale*.

La réduction drastique de l'échantillon interrogé en 1993 s'est accompagnée d'un coût pour la recherche. Dans certains cas, il n'a plus été possible de travailler sur des populations aussi fines que celles définies auparavant ; l'analyse des évolutions est parfois devenue moins précise – on a dû se contenter de cohortes décennales plutôt que quinquennales – ; le suivi de tendances déjà constatées antérieurement n'a pu être conduit de façon aussi rigoureuse qu'auparavant. Un certain nombre de résultats produits paraissent ainsi, comparativement au passé, moins solides. Une conséquence en a été l'auto-limitation des chercheurs : habitués à un outil de grande qualité, leur déception a pu être forte et les conduire, dans certains cas et au prix d'une limitation de leurs investigations, à rechercher ailleurs une information qu'ils envisageaient de trouver dans l'enquête FQP.

Dans le cas où l'enquête serait maintenue, il conviendrait donc de réexaminer cette question pour définir précisément taille de l'échantillon et plan de sondage permettant d'optimiser les analyses statistiques dans les différents champs couverts. *Il conviendrait, en tout état de cause, de maintenir un taux de sondage assez élevé.* Une attention particulière pourrait être apportée aux sous-populations peu couvertes par d'autres enquêtes (certaines professions, les étrangers ou immigrés, etc.). Il conviendrait également de considérer avec attention l'ampleur du questionnaire et d'examiner attentivement les moyens de l'alléger sans qu'il perde de sa richesse sur ses thèmes originaux. *S'agissant du contenu lui-même, deux aspects novateurs pourraient être introduits qui recouvrent des interrogations aujourd'hui très présentes dans la littérature sociologique internationale*. Le recueil d'informations limitées (niveau d'éducation atteint et position professionnelle) pour le frère (ou sœur) le plus proche en âge de l'individu interrogé permettrait de savoir à quel degré les phénomènes de mobilité sociale opèrent à l'échelle individuelle ou familiale. À l'heure de la montée du divorce comme du développement des familles monoparentales et recomposées, quelques informations sur les circonstances dans lesquelles les individus interrogés ont vécu leur jeunesse pourraient aussi permettre d'examiner les effets éventuels de l'environnement familial sur le devenir scolaire et professionnel.

L'enquête FQP est lourde et de ce fait coûteuse et les chercheurs en sciences sociales en ont conscience. Utilisateurs intensifs de cet outil d'observation, ils ont aussi pour responsabilité de

favoriser, dans la mesure de leurs moyens, sa réalisation et donc de se poser la question de son financement. *Compte tenu de l'importance des informations recueillies pour l'étude de long terme de la société française, on peut penser qu'il serait souhaitable que plusieurs ministères ou organismes contribuent à sa réalisation et marquent ainsi l'intérêt qu'ils accordent à cette collecte d'informations* : le ministère de l'Éducation nationale, de la Recherche et de la Technologie – que ce soit par sa Direction de la Programmation et du Développement ou par sa Direction de la Recherche –, le Ministère de l'Emploi et de la solidarité, le Centre National de la Recherche Scientifique. À son niveau de responsabilité et dans le domaine qui est le sien, le Lasmas-Institut du Longitudinal entame actuellement certaines démarches en ce sens.

*** Conclusion**

Invités à réfléchir sur l'opportunité de la programmation d'une nouvelle enquête Formation - Qualification Professionnelle, les chercheurs en sciences sociales se sont prononcés sans ambiguïté pour le maintien de cette série d'enquêtes et ont souligné qu'elle pouvait, sans inconvénient majeur, adopter une périodicité décennale. Si toutefois il apparaissait que des circonstances graves doivent conduire à la suppression pure et simple de cette série d'enquêtes, ils se prononcent en faveur de sa réalisation sous la forme substitutive d'une unique enquête complémentaire à l'enquête sur l'emploi qui permettrait de maintenir la cohérence du questionnaire acquise au fil du temps. Pour le cas où cette solution devrait être adoptée, les chercheurs en sciences sociales entendent néanmoins insister sur les fortes incertitudes qui lui sont liées, dans un contexte où, la réalisation des enquêtes Emploi évoluant vers une interrogation en continu, la forme des enquêtes complémentaires n'est pas à ce jour précisément définie.

2. Les échantillons longitudinaux d'individus : des expériences étrangères et une perspective française

Louis-André Vallet (Lasmas-IdL, CNRS et Laboratoire de Sociologie Quantitative, CREST – INSEE)

Lorsque la recherche en sciences sociales souligne l'apport des dispositifs longitudinaux d'observation à l'étude et à la compréhension des processus sociaux, ce sont le plus souvent les panels de ménages qui sont retenus à des fins d'illustration. Sont alors fréquemment évoqués les exemples du panel lorrain, prédécesseur pour la France du Panel Communautaire des Ménages, ou bien le *German Socio-Economic Panel* en Allemagne, le *British Household Panel Study* au Royaume-Uni, voire le *Panel Study of Income Dynamics* aux États-Unis. On voudrait dans cette note argumenter que les suivis longitudinaux d'individus sont aussi d'un apport irremplaçable pour les disciplines des sciences sociales. Constitués sur la base d'un échantillon appartenant à une même génération ou cohorte, ces panels adoptent un dispositif d'observation qui débute dans l'enfance ou l'adolescence, puis retrouvent les mêmes individus à leur entrée sur le marché du travail ou dans les années qui suivent immédiatement celle-ci ; dans le meilleur des cas, l'observation des individus peut être prolongée jusqu'à un stade très avancé de leur carrière professionnelle. La richesse des informations collectées dans la jeunesse auprès de l'individu lui-même, de sa famille ou de l'établissement scolaire qu'il a fréquenté autorise alors des investigations approfondies sur le degré auquel les expériences vécues dans la jeunesse ou les aptitudes mesurées dans l'enfance marquent le devenir des individus à l'âge adulte comme leurs itinéraires professionnels et sociaux.

Trois expériences étrangères seront présentées dans cette note. On réservera la plus grande place à la *Wisconsin Longitudinal Study* en raison de l'ampleur de son dispositif d'observation comme de l'influence qu'elle a eue sur la sociologie empirique américaine : c'est en effet à partir des données de cette enquête par panel qu'a pu être étudié concrètement le modèle socio-psychologique de l'acquisition du statut, développé par les sociologues de la stratification de l'Université du Wisconsin. La *National Child Development Study* (Grande-Bretagne) et la *Cohorte du Nord-Brabant* (Pays-Bas) feront ensuite l'objet d'une présentation plus succincte. On conclura enfin en argumentant que les panels nationaux d'élèves suivis à partir de 1989 ou 1995 dans l'enseignement du second degré (DEP et DPD, ministère de l'Éducation nationale) peuvent constituer une base naturelle pour la mise en œuvre d'un dispositif analogue d'observation en France.

La Wisconsin Longitudinal Study ou échantillon longitudinal du Wisconsin

La *Wisconsin Longitudinal Study* consiste en l'observation sur le long terme d'un échantillon aléatoire de 10 317 hommes et femmes qui, en 1957, ont obtenu leur diplôme de fin d'études secondaires dans les établissements de l'état américain du Wisconsin[7]. Ces individus ont fait l'objet d'une enquête qui portait notamment sur leurs aspirations et leurs projets d'avenir durant leur dernière année de lycée (avril 1957) – ils avaient alors environ 18 ans – et ils ont été soumis à une nouvelle interrogation en 1975 (par téléphone), puis en 1992-1993 (par téléphone et questionnaire auto-administré), période à laquelle ils avaient dépassé la cinquantaine. En outre, leurs parents ont été soumis à une enquête postale en octobre 1957, puis en 1964 et des données issues des registres scolaires (notes et classements scolaires), des services de psychologie scolaire (mesures de QI et d'aptitudes intellectuelles) comme des registres d'imposition de l'état du Wisconsin (revenus des parents durant les années 1957 à 1960) ont pu aussi être ajoutées, pour chaque répondant, aux informations directement issues des enquêtes. Lors des enquêtes de suivi de 1975 et 1992-1993 et dans le cas où ils avaient un ou plusieurs enfants, les membres de l'échantillon ont aussi fourni des informations sur le devenir de l'un de ces enfants, choisi au

[7]. Pour cette présentation, on s'est inspiré des deux articles suivants et surtout de la documentation très riche fournie sur le site Internet de la Wisconsin Longitudinal Study : http://dpls.dacc.wisc.edu/WLS/wlsarch.htm.
Sewell W. H., Hauser R. M., Wolf W. C., 1980. – Sex, schooling, and occupational status, American Journal of Sociology, 86(3), pp. 551-583.
Warren J. R., Hauser R. M., 1997. – Social stratification across three generations: new evidence from the Wisconsin Longitudinal Study, American Sociological Review, 62(4), pp. 561-572.

hasard. Enfin, des enquêtes complémentaires ont été conduites en 1977 et de façon plus complète en 1993-1994 auprès d'un échantillon aléatoire obtenu en retenant un frère ou une sœur des membres de l'échantillon original ; cet échantillon était constitué de 2 000 frères ou sœurs en 1977 et 4 800 en 1993-1994. La dernière vague d'enquête, auprès des répondants et d'un membre de leur fratrie, incluait un interview téléphonique d'une heure suivi d'un questionnaire auto-administré d'une vingtaine de pages.

Pour l'échantillon longitudinal du Wisconsin, ce sont donc trente-cinq années qui se sont écoulées entre la première collecte d'information et la plus récente. On pourrait craindre qu'une telle amplitude temporelle s'accompagne d'une très forte attrition qui rende délicate l'exploitation scientifique des données recueillies. Ce n'est pas le cas. Des 10 317 membres de l'échantillon original, ce sont 9 139 (88,6 %) qui ont pu être ré-interrogés en 1975 et 8 493 (82,3 %) qui l'ont été en 1992-1993. À cette dernière date, parmi les 9 741 membres survivants de l'échantillon initial, le taux de succès de l'enquête de suivi s'est donc élevé à 87,2 %.

La présence de sources d'information différenciées et l'étendue des questionnaires administrés ont pour conséquence que l'échantillon longitudinal du Wisconsin décrit de façon très complète l'origine socio-économique, les aspirations durant la jeunesse, les études accomplies, le service militaire éventuel, la formation de la famille, la trajectoire suivie sur le marché du travail et la participation sociale des membres de l'échantillon original. Aux mesures de performances scolaires ou d'aptitudes intellectuelles (pour les répondants et 2 000 de leurs frères et sœurs) ont aussi été ajoutées des caractéristiques contextuelles relatives aux communes de résidence, aux établissements scolaires et universitaires fréquentés, aux employeurs et aux firmes. Les données recueillies pour les membres de l'échantillon sont aussi reliées à celles de trois de leurs amis, du même sexe, au cours des études secondaires et qui appartiennent à la population étudiée. Considérées ensemble, les informations collectées en 1992-1993 (membres de l'échantillon original) et 1993-1994 (échantillon des frères ou sœurs) ont permis de recueillir des biographies professionnelles détaillées. Elles décrivent aussi les caractéristiques des emplois, les revenus perçus, les biens possédés et les transferts effectués entre ménages. Elles renseignent également sur les caractéristiques sociales et économiques des parents, des frères et sœurs et des enfants des répondants ainsi que sur les relations que ceux-ci entretiennent avec eux. Une partie substantielle du questionnaire était enfin consacrée à la santé et au bien-être, du point de vue physique et mental.

Selon ses concepteurs, la *Wisconsin Longitudinal Study* qui porte sur une cohorte née principalement en 1939 peut être considérée comme largement représentative de la population des hommes et femmes américains non hispaniques qui ont accompli avec succès leurs études secondaires. Parmi les Américains âgés de 50 à 54 ans en 1990 et 1991, environ 66 % sont blancs, non hispaniques et ont accompli au moins douze années d'études et les concepteurs de la *Wisconsin Longitudinal Study* précisent aussi que, selon leurs estimations, ce sont les trois-quarts des jeunes du Wisconsin qui, à la fin des années cinquante, obtenaient un diplôme de fin d'études secondaires. Près de 19 % des membres de l'échantillon original sont d'origine agricole, ce qui est conforme aux estimations nationales pour les cohortes nées à la fin des années trente. Le spectre des années de naissance pour l'échantillon des frères ou sœurs est évidemment plus large que pour les répondants et s'étend pour l'essentiel de 1930 à 1948. On relèvera enfin qu'en 1964, en 1975 et encore en 1992-1993, ce sont à peu près les deux tiers des membres de l'échantillon qui vivaient dans le Wisconsin contre un tiers dans le reste des États-Unis ou à l'étranger.

Les concepteurs de la *Wisconsin Longitudinal Study* présentent le corpus de données comme une ressource publique de valeur pour des études portant sur le déroulement du cycle de vie, les transferts et les relations entre générations, le fonctionnement de la famille, la stratification sociale, le bien-être physique et mental ainsi que la mortalité. Alors même que les données de la dernière vague d'enquête sont loin d'être totalement exploitées, on soulignera que l'échantillon longitudinal du Wisconsin a d'ores et déjà donné lieu à un nombre considérable de publications, qu'il s'agisse d'ouvrages ou de contributions à des ouvrages, d'articles de revues scientifiques ou de monographies. Certaines de ces publications sont des classiques de la sociologie empirique américaine[8] et l'échantillon longitudinal a contribué à faire de l'Université du Wisconsin un

[8]. On ne citera ici pour exemple que l'ouvrage suivant :

département de sociologie de réputation internationale dans le domaine de la stratification sociale, sous l'impulsion des professeurs David L. Featherman, Robert M. Hauser et William H. Sewell. Sans insister davantage, on ne détaillera ici que deux apports importants à la sociologie étroitement liés à des caractéristiques du dispositif d'enquête.

Dès la première vague d'enquête, la *Wisconsin Longitudinal Study* a mis l'accent sur la mesure des aspirations scolaires et professionnelles des jeunes interrogés ainsi que sur l'appréhension des « influences sociales » – notamment le fait que leurs parents (respectivement leurs professeurs) les encourageaient ou non à entrer à l'université ou encore que la plupart de leurs amis avaient l'intention de le faire. Des questions sur la satisfaction au travail, les aspirations professionnelles ou les projets d'avenir ont aussi été posées dans les enquêtes ultérieures. Cette dimension psycho-sociologique des données collectées s'est traduite dans l'élaboration du « modèle socio-psychologique de l'acquisition du statut professionnel », forme très enrichie du *status attainment model* proposé par Blau et Duncan dans *The American Occupational Structure* (1967). Ainsi, dans leur article de 1980, Sewell, Hauser et Wolf représentent (et estiment statistiquement) le processus d'acquisition du statut professionnel à l'aide d'un modèle récursif constitué de plusieurs groupes de variables où les variables antécédentes sont susceptibles d'affecter toutes celles qui les suivent dans l'ordre temporel suivant : l'origine socio-économique (8 variables) ; les capacités intellectuelles (1 variable) ; les performances scolaires (1 variable) ; les influences sociales (3 variables) ; les aspirations exprimées (2 variables) ; le niveau d'éducation atteint (1 variable) ; le statut professionnel en début de vie active (1 variable) ; le statut professionnel en 1975 ou au dernier emploi (1 variable).

L'originalité du dispositif d'enquête qui, dans les années soixante-dix et quatre-vingt-dix, a recueilli systématiquement des informations sur deux membres des mêmes familles – celui appartenant à l'échantillon original et un frère ou une sœur du premier – a permis également des avancées significatives dans la sociologie de la stratification américaine par l'estimation de « modèles de ressemblance au sein de la fratrie » (*models of sibling resemblance*). On sait en effet que les analyses des niveau d'éducation et statut professionnel atteints qui représentent l'origine familiale par un ensemble de variables explicitement mesurées risquent de sous-estimer l'influence de celle-ci puisqu'elles omettent de prendre en considération des caractéristiques non observées de la famille d'origine. Disposer d'observations appariées pour des individus des mêmes fratries constitue dès lors une stratégie d'estimation de l'influence *totale* du milieu d'origine et le département de sociologie de l'Université du Wisconsin s'est aussi construit une réputation internationale dans ce domaine grâce aux données de la *Wisconsin Longitudinal Study*[9].

Sewell W. H., Hauser R. M., 1975. – Education, Occupation, and Earnings: Achievement in the Early Career, New York, Academic Press.

[9]. On consultera par exemple les articles suivants :

Hauser R. M., Mossel P. A., 1985. – Fraternal resemblance in educational attainment and occupational status, American Journal of Sociology, 91(3), pp. 650-671.

Hauser R. M., Sewell W. H., 1986. – Family effects in simple models of education, occupational status, and earnings: findings from the Wisconsin and Kalamazoo studies, Journal of Labor Economics, 4, pp. S83-S115.

Hauser R. M., 1988. – A note on two models of sibling resemblance, American Journal of Sociology, 93(6), pp. 1401-1423.

La National Child Development Study britannique

La *National Child Development Study* réalisée par le *National Children's Bureau* de Londres a consisté dans le suivi temporel, durant leurs vingt-trois premières années, d'un ensemble d'individus issus de la cohorte de naissance britannique de 1958 : les enfants nés en Angleterre, Écosse et Pays de Galles durant la semaine du 3 au 9 mars 1958 ; des informations ont été collectées dans ce pays à leur sujet à cinq périodes qui correspondent aux âges de 7 ans (immédiatement avant le passage de l'école enfantine à l'école élémentaire), 11 ans (immédiatement avant l'entrée dans l'enseignement secondaire), 16 ans (à l'âge de fin de la scolarité obligatoire), 20 ans et 23 ans[10].

La *National Child Development Study* a trouvé en réalité son origine dans la *Perinatal Mortality Survey* qui portait sur les 17 733 bébés nés en Grande-Bretagne durant cette première semaine de mars 1958. En 1965, le *National Children's Bureau* a conduit une étude des mêmes enfants et de leurs familles : des interviews approfondis ont été réalisés auprès de l'un des parents (ou tuteurs), le plus souvent la mère ; des informations ont été collectées auprès du directeur d'école et de l'instituteur de chaque enfant ; un examen médical a été conduit par les services de santé scolaires et une batterie de tests a été administrée aux enfants de l'échantillon.

En 1969, un ensemble comparable de données a été recueilli à partir des mêmes sources. Dans la mesure du possible, les mesures effectuées étaient identiques ou au moins très similaires à celles réalisées à l'âge de 7 ans. Pour la première fois cependant, des informations ont été collectées auprès des enfants eux-mêmes en sus de l'administration d'une batterie de tests.

Lors de la troisième vague d'enquête conduite en 1974, la collecte directe de données auprès des jeunes s'est faite plus complète : batterie de tests, mais aussi mesures des aspirations et des projets d'avenir, sur les plans de l'éducation et de la profession comme ceux du mariage et de la formation de la famille. Toutes les sources d'information habituelles ont été de nouveau utilisées : interview d'un parent, questionnaires complétés par le personnel scolaire et examen médical.

En 1978, pour les jeunes de l'échantillon, des informations ont été recueillies auprès des établissements scolaires concernant leur âge de sortie de l'enseignement secondaire et leurs performances aux examens auxquels ils s'étaient présentés. Enfin, en 1981, des interviews approfondis ont été conduits auprès des membres de la cohorte. Ils portaient surtout sur les éventuelles études post-secondaires, les activités d'apprentissage professionnel et les expériences du marché du travail, quoique bien d'autres informations étaient aussi recueillies. Les jeunes nés en 1958 étaient invités à fournir ces informations de façon détaillée depuis l'âge de 16 ans, complétant ainsi les années qui s'étaient écoulées depuis leur dernière interrogation par l'équipe de recherche.

Lors des quatre premières vagues d'enquête, les établissements scolaires ont occupé une place centrale dans le dispositif de collecte et, en raison même du critère retenu pour constituer l'échantillon, la grande majorité des écoles britanniques se sont trouvées impliquées. À la naissance, la taille de l'échantillon s'élevait à 17 733. Des données ont pu être collectées pour 85,2 % de ces individus à l'âge de 7 ans, 83,8 % à 11 ans, 78,9 % à 16 ans, 77,4 % à 20 ans et 68,4 % à 23 ans. Raisonner sur les seuls « survivants », i.e. ceux qui, au moment d'une vague d'enquête, n'étaient pas décédés et demeuraient encore en Grande-Bretagne, fournit des proportions plus élevées : 91,6 %, 91,6 %, 86,7 %, 85,1 % et 77,1 % respectivement. Les diverses études statistiques conduites sur l'attrition de l'échantillon ont par ailleurs conclu qu'elle n'affectait que de biais minimes les estimations et conclusions qui pouvaient être tirées sur la base de cette observation longitudinale de la cohorte de naissance.

La *National Child Development Study* a donc recueilli en Grande-Bretagne des informations systématiques sur le milieu d'origine, les aptitudes, performances et trajectoires scolaires durant l'enfance, l'adolescence et la jeunesse ainsi que l'insertion sur le marché du travail d'un échantillon représentatif d'individus nés en 1958. C'est sur la base de cette information que le sociologue américain Alan C. Kerckhoff de Duke University a étudié la « divergence des itinéraires » : les caractéristiques structurelles des institutions scolaires et du marché du travail affectent les profils

[10]. On s'appuie ici sur l'ouvrage suivant :
Kerckhoff A. C., 1993. – Diverging Pathways: Social Structure and Career Deflections, Cambridge, Cambridge University Press.

individuels de réussite qui pouvaient être prédits sur la base des seules aptitudes mesurées ; l'observation longitudinale permet de mettre en évidence le résultat cumulé des positions avantageuses ou désavantageuses qu'occupent les individus aux différents moments ; c'est par là même la compréhension de l'interaction entre caractéristiques personnelles et arrangements institutionnels dans son effet sur le devenir des individus qui s'en trouve enrichie[11].

La Cohorte néerlandaise du Nord-Brabant

On évoquera enfin brièvement cet échantillon longitudinal constitué d'enfants nés autour de 1940. Son ampleur demeure plus limitée que les expériences précédemment évoquées, par la taille plus réduite de l'échantillon comme le caractère moins systématique du dispositif de collecte[12].

En 1952, dans le cadre d'une étude sur la qualité des écoles primaires, un échantillon d'un quart des enfants scolarisés en dernière année d'enseignement élémentaire dans la province du Nord-Brabant firent l'objet d'une enquête. On y enregistra en particulier la profession du père, les aptitudes intellectuelles – mesurées par le test Progressive Matrices de Raven – et le parcours scolaire déjà effectué de 5 771 enfants. Deux enquêtes de suivi furent réalisées dans la période 1957-1959 et atteignirent au total 2 830 d'entre eux. La première concernait les élèves qui avaient obtenu un score supérieur à la moyenne à un test de performance scolaire ; la seconde portait sur les seuls fils d'agriculteurs et d'ouvriers (ainsi qu'un groupe contrôle formé de garçons des autres origines).

En 1983 et à la suite d'un effort conjoint d'économistes et de sociologues néerlandais, les adresses de 4 706 individus, soit près de 82 % des membres de l'échantillon original, sont localisées et il leur est envoyé un questionnaire par voie postale. Selon les indications fournies par Dronkers (1998), le taux de réponse s'élève à 58 % (2 641 individus), puis à 46 % (2 397 individus) au terme d'une opération analogue conduite en 1993. Les interrogations de 1983 et 1993 concernaient le niveau d'études atteint par le répondant et son conjoint, leurs activités professionnelles et leurs revenus. Il a été possible d'établir que les non-réponses de 1983 n'affectaient pas la représentativité du sous-échantillon masculin, comparativement à l'échantillon original de 1952 et, depuis le milieu de la décennie quatre-vingt, les données de la cohorte du Nord-Brabant ont été utilisées intensivement, tant par les économistes que par les sociologues néerlandais.

---

[11]. Cette recherche n'est d'ailleurs pas achevée : dans les dernières pages de son ouvrage, Kerckhoff précise que les membres de la cohorte ont été encore interrogés en 1991 ; des données relatives aux trajectoires professionnelles ont été recueillies, mais l'auteur ne fournit pas d'indications supplémentaires à ce sujet.

[12]. On s'appuie ici sur le texte suivant :
Dronkers J., 1998. – The importance of cognitive abilities at primary school for educational and occupational success in the life course of a Dutch generation, born around 1940, paper presented at the Research Committee 28 Social Stratification of the International Sociological Association World Congress, Montréal, Canada.

Les échantillons longitudinaux du ministère de l'Éducation nationale :

une base naturelle pour une expérience française analogue

On voudrait dans cette dernière section argumenter l'idée simple suivante : les échantillons longitudinaux d'élèves suivis par la Direction de l'Évaluation et de la Prospective, puis aujourd'hui la Direction de la Programmation et du Développement, peuvent constituer la base d'un dispositif périodique d'observation sur le long terme des mêmes individus au cours de leur vie adulte. On constituerait ainsi progressivement, pour la France, un (des) corpus de données d'intérêt et d'ampleur comparables à ceux de la *National Child Development Study* et de la *Wisconsin Longitudinal Study* et dont les pages précédentes ont souligné la fécondité pour la recherche empirique en sciences sociales. Plusieurs éléments plaident en faveur d'une telle perspective : l'intérêt et la diversité des données d'ores et déjà recueillies par le ministère de l'Éducation nationale, la qualité de la collecte d'informations effectuée et la très faible attrition des échantillons suivis au cours des études secondaires. Il ne s'agirait évidemment pas de prévoir la poursuite d'une observation en continu – i.e. sur une base annuelle – des mêmes individus. Un tel dispositif serait en effet par trop coûteux. En revanche, le suivi périodique – i.e. avec un intervalle compris entre cinq et dix ans entre deux interrogations successives – pourrait constituer, comme le prouvent les expériences étrangères, une stratégie pertinente de collecte. On conclura cette note en rappelant succinctement les caractéristiques principales des panels nationaux 1989 et 1995 d'élèves du second degré suivis par le ministère de l'Éducation nationale.

Le panel national 1989 d'élèves du second degré est constitué par l'ensemble des enfants nés le 5 d'un mois quelconque qui, en septembre 1989, étaient scolarisés dans une classe de sixième ou de première année de section d'éducation spécialisée dans un collège public ou privé de France métropolitaine ou des départements d'outre-mer. Ces élèves – dont l'effectif total excède 26 000 – ont été suivis jusqu'au terme de leurs études secondaires et un sous-échantillon d'entre eux fait également l'objet d'une observation dans l'enseignement supérieur. Le panel national 1989 a repris le principe des panels antérieurs du même ministère : une enquête de recrutement la première année, des enquêtes de suivi les années suivantes, toutes renseignées par les chefs d'établissement et auxquelles viennent s'ajouter divers questionnements étalés dans le temps dont, en particulier, des enquêtes sur l'insertion des jeunes sortis du système éducatif. En outre et pour la première fois, le dispositif du panel 1989 a inclus, au printemps 1991, une enquête complémentaire auprès des familles des élèves et, pour une partie de l'échantillon, il s'est aussi enrichi des résultats aux épreuves standardisées d'évaluation en français et mathématiques administrées à l'entrée en classe de sixième[13].

Le panel national 1995 reprend un dispositif général largement analogue au précédent. Il est constitué des élèves scolarisés en sixième ou entrant en SES à la rentrée scolaire 1995-1996 dans un établissement public ou privé de France métropolitaine et nés le 17 d'un mois (à l'exception de mars, juillet et octobre). Les résultats aux épreuves d'évaluation de l'entrée en sixième ont, cette fois, été systématiquement collectés et les élèves ont aussi répondu à deux questionnaires individuels intitulés « Comment je travaille » et « La vie en société ». Enfin, l'enquête complémentaire auprès des familles, très enrichie par rapport à celle conduite en 1991, a été réalisée au printemps 1998. Au total, ce sont 19 770 élèves qui seront suivis jusqu'au terme de leur formation initiale. Il a aussi été prévu, lors de la mise en place du panel, que les mêmes élèves feront l'objet d'une observation, conduite par le Céreq, dans les premières années de leur insertion professionnelle.

[13]. Pour un aperçu plus complet des recherches possibles à partir du panel national 1989, on pourra consulter l'exemple suivant : Vallet L.-A., Caille J.-P., 1996. – Les élèves étrangers ou issus de l'immigration dans l'école et le collège français. Une étude d'ensemble, Les dossiers d'Éducation et Formations, 67, Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, Direction de l'Évaluation et de la Prospective, 153 p.

3. Note sur les enquêtes électorales en France

Gérard Grunberg, Directeur de recherche au CNRS

Bruno Cautrès, Chargé de Recherche au CNRS

Les grandes enquêtes électorales par sondage sont devenues depuis les années soixante un outil scientifique indispensable pour l'étude du comportement électoral et plus largement des systèmes de valeurs et de représentation. L'avance prise dans les années cinquante et soixante par les pays anglo-saxons dans le domaine de la sociologie électorale a été due en grande partie à l'existence de ce type de matériaux, renouvelé à chaque élection, et qui permettait ainsi de suivre les évolutions des opinions et des votes. Aux États-Unis et en Grande-Bretagne, les chercheurs disposent de moyens financiers publics pour réaliser lors de chaque élection importante (présidentielle aux États-Unis et législatives en Grande-Bretagne) une grande enquête qui est à la disposition ensuite de l'ensemble de la communauté scientifique. En France, on reste encore dans le bricolage. Depuis le début de la cinquième République, la communauté des politologues et des sociologues n'a pu mener que quatre enquêtes électorales lourdes : en 1978 (législatives), 1988 (présidentielle) 1995 (présidentielle) et 1997 (législative). À chacune de ces occasions, les chercheurs spécialisés sur ces questions ont dû se livrer à des acrobaties et profiter de leurs relations personnelles pour trouver des crédits, limités, crédits gouvernementaux et provenant de journaux, mais aussi des crédits américains. La Fondation des sciences politiques a consenti des efforts mais ses moyens dans ces domaines sont très limités. Le CNRS n'a pu financer, même modestement, la réalisation de ces enquêtes car il n'y a pas de crédits disponibles pour ce type d'enquêtes périodiques lourdes. Or, dans les sciences sociales et dans la science politique et la sociologie en particulier, il n'y a pas de besoins très importants comme dans les sciences de la matière et de la vie en matière d'investissements lourds, mais il y a des besoins de données, et en particulier de données d'enquêtes. En outre, l'absence de financement public régulier donne un caractère aléatoire à la réalisation de ces enquêtes. Or le suivi fin de l'évolution des opinions et des votes nécessite des enquêtes rapprochées dans le temps et seules des séries longues nous fournissent le matériau nécessaire. L'absence d'enquêtes en 1981 et 1986 a été très dommageable de ce point de vue. Par ailleurs, il faut souligner que la faiblesse de certains financements contraint souvent à réduire les ambitions en termes de procédés de sondage et de tailles des échantillons.

L'existence d'une telle série d'enquêtes permettrait aux chercheurs français d'occuper une position meilleure dans ce domaine de recherche au niveau international, en offrant à la fois plus de possibilités d'études, en particulier comparatives, et plus d'occasions d'échange de données avec les équipes étrangères, nombreuses et de qualité, travaillant dans le même domaine.

Dans les régimes représentatifs modernes où les questions relatives au rapport des citoyens à la politique et aux gouvernants, à la citoyenneté, française et européenne, et à celle de l'évolution des grandes tendances politiques dans l'opinion sont centrales, un dispositif permanent d'enquêtes par sondage permettrait de les éclairer.

4. Une enquête internationale, l'ISSP, et le projet European Social Survey

Bruno Cautrès (CIDSP-BDSP), Alain Degenne, Michel Forsé (Lasmas-IdL)

L'*International Social Survey Programme* (ISSP) a été fondé en 1985 par des équipes de recherche australienne (ANU, Canberra), américaine (NORC, Chicago), britannique (SCPR, Londres) et allemande (ZUMA, Mannheim) afin d'effectuer une enquête commune annuelle sur des échantillons représentatifs de ces différents pays. Le thème change chaque année (inégalités, réseaux, famille, etc.) mais revient en principe au bout de huit ans. En 1993, 22 pays avaient rejoint ce programme, dont tous les pays du G7, sauf la France. Pensant que cette situation était dommageable, l'association France-ISSP a été créée dans le but d'intégrer la France à ce programme scientifique. Elle réunit des chercheurs de laboratoires du CNRS (Lasmas, OSC, CIDSP) et de l'Insee (CREST, LSQ). Le groupe international a reconnu cette association comme son interlocuteur français en 1994. Les financements pour réaliser l'enquête ISSP en France n'ont pas été faciles à trouver, mais grâce aux efforts de différentes institutions (CNRS, FNSP, Insee, IUF), la première enquête sur le rôle de l'État (n = 1300) a pu être effectuée en 1997. La seconde enquête, sur les attitudes à l'égard du travail s'est faite début 1998. La troisième enquête sur la religion a été réalisée fin 1998. En 1999, l'enquête portera sur les inégalités sociales. L'environnement et les réseaux sociaux sont les thèmes d'ores et déjà retenus pour 2000 et 2001. Les données sont archivées et diffusées par le Zentralarchiv de Cologne. Chaque enquête produit un CD-Rom sur lequel sont enregistrés les fichiers de tous les pays dans lesquels elle a été réalisée. Les analyses secondaires comparatives sont ainsi très aisées à réaliser.

Une trentaine de pays participent aujourd'hui à ce programme. Il y a donc là une banque de données extrêmement riche. Malheureusement, le financement trouvé en France est beaucoup trop faible pour produire des données de la même qualité que celle de la plupart des autres pays : 60 KF en regard des 200 KF apparaissant comme un minimum pour réaliser un sondage de cette nature auprès de 1500 personnes. La garantie de régularité de ce financement est aussi une condition nécessaire de la continuité de l'engagement dans ce projet.

L'*European Science Foundation* (ESF) a décidé en 1996 la création d'un groupe d'experts nommés par leurs organismes nationaux et chargés de rédiger un document sur la faisabilité d'une « enquête sociale européenne ». Les travaux de ce groupe d'experts ont duré près de deux ans et leur rapport final vient d'être approuvé (mai 1999) lors d'un récent *Standing Committee for Social Sciences* de l'ESF. Son secrétaire général pour les sciences sociales, M. John Smith ainsi que par le Professeur Max Kaase, du WZB de Berlin, président du groupe d'experts réunis par l'ESF autour de ce projet ont présenté ce document lors de cette réunion. Ce rapport final (64 pages) est le fruit de nombreuses réunions de travail ; les experts nommés auprès de l'ESF, répartis en un *Steering Committe* et un *Methodology Committee*[14], ont examiné tous les aspects théoriques, pratiques et méthodologiques que poserait la réalisation d'une enquête comparative européenne sur grand échantillon (au minimum 2500 questionnaires). Leurs conclusions générales valident ce projet et en posent les grandes lignes. Plusieurs pays ont déjà fait savoir qu'ils participaient à ce projet : le Royaume-Uni, l'Allemagne, le Portugal (la Hollande, la Belgique, l'Espagne, la Grèce en principe aussi).

Les données ainsi collectées constitueront une base de données européennes sur les comportements et valeurs sociales et politiques des européens. Les thèmes des modules seront également susceptibles de répétition dans le temps. Le rapport du groupe d'experts cite certains thèmes de modules possibles : exclusion sociale et inégalités, chômage et orientations vis-à-vis du travail, crime et victimation, la personnalité post-moderne, société civile et confiance, groupes d'intérêts et partis, unification européenne et identités nationales, attitudes à l'égard de la science et de la technologie, entre autres.

[14]. La France y est représentée par Bruno Cautrès (CR1 CNRS au CIDSP Grenoble) pour le Steering Committee et par Nonna Mayer (DR2 CNRS au Cevipof, Paris) pour le Methodology Committee.

Les coûts de production de cette enquête sont de deux sortes :

– coûts fixes du dispositif international, à la charge de l'ESF, évalués à 260 000 Euros par an

– coûts de production de chaque enquête nationale, à la charge de chacun des pays membres (évaluation pour la France entre 350 000 et 380 000 Euros, tous les deux ans sur la base de 2500 individus).

La situation de la France par rapport à ces deux grands dispositifs d'enquête comparative internationale illustre bien les thèses développées dans ce rapport.

L'ISSP existe depuis 15 ans et notre pays n'y participe que depuis 4 ans avec des moyens et un budget particulièrement réduits, qui ne permettent pas de réaliser la collecte dans des conditions normales de nature à garantir la représentativité de l'échantillon. Cependant cette enquête existe et permet de comparer entre eux du point de vue d'attitudes très générales un grand nombre de pays dont les plus industrialisés.

Le second projet est centré sur l'Europe et intéresse plus directement les sciences du politique. Il n'a pas encore connu de mise en œuvre et se montre beaucoup plus ambitieux du point de vue des conditions de réalisation et des coûts qu'elles induisent.

Il est évidemment plus que souhaitable que la France participe dans de bonnes conditions à ces deux programmes d'enquête, ce qui suppose d'une part qu'elle dégage les moyens pour normaliser les conditions de réalisation de l'enquête annuelle ISSP (200 KF par an) et d'autre part qu'elle envisage dès maintenant le montage financier que suppose l'entrée dans le dispositif *European Social Survey*.

Annexe V : Les questions juridiques

1. Études statistiques et droit d'auteur

Philippe Chevet, Cecoji, CNRS

Plusieurs étapes de la « vie » des statistiques sont à différencier, qui comportent chacune ses propres conséquences juridiques. Nous en distinguerons principalement trois, que nous examinerons plus précisément sous l'angle du droit d'auteur :

– la constitution des données,

– l'utilisation des données,

– l'exploitation des données.

1. La constitution des données

La constitution de données soulève différentes questions, de nature ou à incidence juridique, qui concernent tant la collecte elle-même (avec le respect impératif de certaines normes juridiques et déontologiques bien définies par ailleurs) que la constitution de fichiers qui lui fait logiquement suite (avec notamment la prise en compte de la réglementation concernant les fichiers nominatifs).

Une question, qui concerne, elle, l'ensemble de l'opération (de la collecte au traitement des informations statistiques), est propre au droit d'auteur. Il s'agit en effet de déterminer qui peut légitimement revendiquer des droits sur l'enquête une fois constituée. Pour cela, il faudra auparavant répondre à la question suivante[15] : s'agit-il d'une œuvre de l'esprit et pourquoi ?

1a. La qualification juridique de l'enquête statistique

L'on pourra donc appliquer le régime du droit d'auteur en matière d'enquêtes statistiques uniquement si ces dernières peuvent être considérées comme des œuvres de l'esprit. Deux points sont à différencier : la collecte et le traitement.

S'agissant de la collecte en elle-même, il est permis de se demander s'il peut s'agir véritablement d'une « œuvre de l'esprit ». À première vue, il manque le critère indispensable d'originalité pour accéder à cette qualification. En effet, l'on se borne dans une enquête, du moins au départ, à constater et vérifier une situation préexistante, et dans ce cas il n'y a apparemment pas de véritable « création », et donc de droit d'auteur la protégeant. Il faut cependant pousser plus loin la réflexion, tant au niveau de la collecte des informations que de leur traitement.

Au niveau donc de la simple collecte, l'on peut a priori douter de l'originalité de l'opération, et donc de la protection par le droit d'auteur. Il faut préciser qu'aujourd'hui le code de la propriété intellectuelle intègre dans le champ d'application de ce dernier la base de données[16]. L'article L.112-3 dispose en effet que « *les auteurs de traductions, d'adaptations, transformations ou arrangements des œuvres de l'esprit jouissent de la protection instituée par le présent code sans préjudice des droits de l'auteur de l'œuvre originale. Il en est de même des auteurs d'anthologies ou de recueil d'œuvres ou de données diverses, tels que les bases de données, qui, par le choix ou la disposition des matières, constituent des créations intellectuelles. On entend par base de données un recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen.* » Cette nouvelle donne vient de façon radicale modifier le raisonnement initial. Il est évident que ces informations statistiques, qu'elles soient collectées et/ou stockées de façon manuelle ou automatisée, constituent bien une base de données, et en tant que telle sont protégées par le droit d'auteur. Il ne faut cependant pas se tromper, car le droit de la base de données est quelque peu dérogatoire du droit commun.

---

[15]. En effet, pour revendiquer des « droits d'exploitation » au sens du droit d'auteur, encore faut-il que l'on soit bien en présence d'une œuvre de l'esprit (en dépit de quoi le droit commun de la propriété s'appliquera).

[16]. Depuis la loi de transposition de la directive 96/9 du 11 mars 1996 sur les bases de données. Nous conseillons sur ce point la lecture de l'article de Philippe Gaudrat, publié à la revue trimestrielle de droit commercial 1998 (juillet/septembre), p. 598 et suivantes.

Au niveau du traitement, la question est moins pertinente, car la protection ne posera évidemment aucun problème. Elle sera même plus étendue que celle relative à la simple collecte (ce qui peut s'expliquer par le travail effectué). L'ensemble, qui constitue une œuvre de l'esprit, sera protégé par le droit d'auteur. On appelle « traitement » la mise en forme[17], le commentaire, la production donc d'un document. Cet effort intellectuel sera récompensé. C'est par ce biais que les données en elles-mêmes pourront être le mieux protégées[18].

1b. La titularité des droits d'auteur sur l'enquête statistique

Qui est titulaire des droits sur une œuvre de l'esprit ? L'auteur, nous répond tout naturellement l'article L.111-1 du code de la propriété intellectuelle[19]. Est-ce toujours le cas ? Non, nous répond ce même code. A priori, et c'est heureux, même notamment dans un cas de création après conclusion d'un contrat de commande ou bien dans le cadre d'un contrat de travail, l'auteur reste bien titulaire originaire des droits d'auteur. On l'aura compris, il s'agit d'un principe important de notre droit d'auteur. Mais dans certains cas le principe est oublié, sans doute car le législateur entend donner raison, dans tout domaine (pourquoi le nôtre y échapperait ?), et de façon constante, à Pascal, pour qui « *il n'est pas de principe… qui n'ait quelque exception* » (« Pensées »). Quelles sont ces exceptions ?

Tout d'abord, citons l'œuvre collective, qui, de façon spectaculaire[20], permet à une personne, physique ou morale, de récolter les droits (soit les droits d'exploitation plus le droit moral en cadeau) sur l'ensemble de l'œuvre[21]. L'œuvre collective est définie par l'article L.113-2, alinéa 3, comme « *dite collective l'œuvre créée sur l'initiative d'une personne physique ou morale qui l'édite, la publie et la divulgue sous sa direction et son nom et dans laquelle la contribution personnelle des divers auteurs participant à son élaboration se fond dans l'ensemble en vue duquel elle est conçue, sans qu'il soit possible d'attribuer à chacun d'eux un droit distinct sur l'ensemble réalisé.* » Donc, dans ce type de création, l'œuvre est « *sauf preuve contraire, la propriété de la personne physique ou morale sous le nom de laquelle elle est divulguée. Cette personne est investie des droits d'auteur* » (article L113-5). L'on s'épargnera ici l'énumération de l'ensemble des conditions pour qu'une œuvre soit qualifiée de « collective », ainsi que l'étude des nombreuses controverses que suscite la notion[22]. L'on se bornera à mentionner la principale caractéristique de l'œuvre collective (outre le fait qu'il s'agisse d'une œuvre « d'équipe »), à savoir la direction de l'ensemble par une personne physique ou morale (qui peut ou non participer à la création, peu

---

[17]. Qui obéit à certaines règles, comme par exemple l'obligation de rendre anonymes les données.

[18]. Notamment en raison du droit de reproduction qu'a l'auteur sur son œuvre (en son ensemble comme en ses éléments).

[19]. « L'auteur d'une œuvre de l'esprit jouit sur cette œuvre, du seul fait de sa création, d'un droit de propriété incorporelle exclusif et opposable à tous. » Il existe toutefois une présomption de « paternité », figurant à l'article L.113-1 : « la qualité d'auteur appartient, sauf preuve contraire, à celui ou à ceux sous le nom de qui l'œuvre est divulguée » (présomption qui bien entendu peut être renversée en cas « d'usurpation » de cette qualité d'auteur par un tiers).

[20]. Spectaculaire car notre droit d'auteur est profondément « personnaliste » (le cordon ombilical liant l'auteur à son œuvre ne peut être rompu de façon totale ou définitive, il restera au moins cette « paternité » et cette « renommée » après sa mort). Dans un tel système, l'on ne peut par exemple concevoir qu'une autre personne que l'auteur détienne un droit moral (il est inaliénable) sur l'œuvre, du moins lorsque celui-ci est vivant et reconnu en sa qualité d'auteur.

[21]. Le principe dans une « création plurale » est que « » (article L.113-3). Concrètement, le régime est celui de l'indivision (toutefois quelque peu dérogatoire au droit commun, droit d'auteur oblige). Il s'agit de l'œuvre de collaboration donc, qui est définie comme « » (article L.113-2).

[22]. Nous renvoyons ici le lecteur à la lecture du traité de MM. Lucas, Propriété littéraire et artistique, Litec, 1994, p189 et suivantes, pour la présentation complète de l'œuvre collective. Les controverses sont courantes dans la littérature juridique. Elles concernent par exemple le point de savoir si l'œuvre collective ne peut être (ou pas) invoquée seulement en matière littéraire et pour certaines œuvre seulement (l'article emploie les termes « édition » et « publication », et l'on sait que l'œuvre collective a été à l'origine créée pour répondre aux problèmes que posaient les journaux, dictionnaires et encyclopédies notamment, en raison du fort nombre de personnes concourant à leur création). En pratique, nous le savons, cette conception est aujourd'hui dépassée (les « créations informatiques » reçoivent de façon quotidienne cette qualification d'œuvre collective).

importe, une coordination en tout cas suffit). C'est cette dernière qui détiendra les droits d'auteur sur l'œuvre en entier[23].

La seconde exception est celle de l'œuvre créée par le fonctionnaire dans le cadre de ses fonctions. On oublie très souvent de la présenter, et pour cause, le code de la propriété intellectuelle ne prévoit rien sur ce point. Le principe est donc bien affirmé[24], mais il n'est pas pour autant très clair dans tous les cas (par exemple, l'exploitation commerciale du cours d'un professeur d'université lui appartient en propre, même si ce cours fait partie de ses obligations[25]).

L'on ne peut, en ce qui concerne les études statistiques, privilégier une qualification à une autre. A priori, toutes peuvent s'appliquer. Il suffit d'observer le mode d'élaboration de l'œuvre pour la qualifier et connaître son régime.

S'il s'agit d'une étude effectuée par une seule personne, pas de problème : celle-ci sera considérée comme auteur, et de ce fait sera titulaire des droits à titre originaire[26]. S'il est fonctionnaire, et s'il a créé l'œuvre dans l'exercice de ses fonctions et pour les besoins de la mission qui lui est attribuée, alors (c'est l'exception qui joue) son administration de rattachement peut détenir à titre originaire les droits d'auteur sur celle-ci. La pratique diffère selon les lieux où le fonctionnaire travaille.

La chose se complique s'il s'agit d'une « création d'équipe ». Notamment, il faudra déterminer s'il s'agit d'une œuvre de collaboration ou bien d'une œuvre collective. La chose n'est pas toujours simple en pratique mais, en théorie, si une personne physique ou morale est à l'origine de la création, si elle la dirige, et la publie sous son nom, cela peut suffire pour qualifier l'œuvre de « collective ». En cas de création d'équipe entre fonctionnaires, la réponse semble plus claire, en principe, en raison de la titularité « automatique » (en fait, elle ne l'est pas toujours selon les cas) de l'administration.

2. L'utilisation des données

Distinguons deux points : le cadre général tout d'abord, un cas particulier ensuite.

2a. Le cadre juridique général de l'utilisation des enquêtes

S'il s'agit d'une œuvre de l'esprit (et a priori la réponse est positive) l'enquête statistique est soumise au droit d'auteur. Dès lors, des droits et obligations encadrent son utilisation.

L'article L.122-5 du code de la propriété intellectuelle en livre les principaux. Notamment, un droit « d'analyse et de courtes citations » est accordé à toute personne à condition qu'il soit justifié « *par le caractère critique, polémique, pédagogique, scientifique ou d'information de l'œuvre à laquelle elles sont incorporées* ». L'auteur de l'œuvre première ne peut donc empêcher un tel usage de celle-ci. La notion de « courte citation » a été fort heureusement précisée par les tribunaux (notamment elle doit être brève[27], comporter le nom de l'auteur[28],…).

Il ne saurait être question de la reproduire intégralement (du moins sans autorisation explicite de l'auteur), car il s'agirait dès lors d'une contrefaçon. Seule dans ce cas peut être tolérée la « copie privée » (pour un usage, comme son nom l'indique, strictement privé, et non collectif).

---

[23]. Il faut en effet préciser que si sa contribution est identifiée, l'auteur détiendra pour sa part un droit moral et des droits d'exploitation sur celle-ci (mais ses prérogatives seront limitées par le droit d'auteur qui porte sur l'ensemble de l'œuvre. Cette concurrence des droits peut entraîner certains conflits, qui le plus souvent se résoudront au profit de la personne qui détient les droits sur l'entière œuvre collective, sauf si précisément cette dernière nie les droits sur les contributions.)

[24]. Voir sur l'ensemble de la question la thèse de C. Blaizot-Hazard, Les droits de propriété intellectuelle des personnes publiques, LGDJ, 1991.

[25]. Cour d'appel de Paris, 24 novembre 1992, Revue internationale du droit d'auteur, janvier 1993, p.191.

[26]. Le contrat de travail (s'il s'agit d'un salarié) peut-il cependant prévoir la « cession » automatique des droits d'exploitation à l'employeur ? Non, car il s'agirait alors d'une « cession globale d'œuvres futures » prohibées à l'article par le code de la propriété intellectuelle. Par contre, une « cession » organisée pour une œuvre au cas par cas peut être bien entendue envisagée.

[27]. Cour d'appel de Paris, 15 juin 1901, Dalloz, 1903, 2, p.273.

[28]. T.G.I. Paris, 5 janvier 1983, Revue internationale du droit d'auteur, avril 1983, p.210.

2b. La question de « l'œuvre seconde »

On appelle « œuvre seconde » l'œuvre qui « s'inspire » d'une œuvre première, dite principale. Notre code de la propriété intellectuelle prévoit bien entendu cette possibilité. Il s'agit de l'article L.113-2 alinéa 2, qui définie « l'œuvre composite » (c'est le terme utilisé par le code) comme : « *l'œuvre nouvelle à laquelle est incorporée une œuvre préexistante sans la collaboration de l'auteur de cette dernière* ». L'une des principales caractéristiques de l'œuvre est donc que l'auteur de l'œuvre première ne participe en rien à l'œuvre seconde (il n'y a pas de « collaboration »). Le principe est que l'auteur de cette dernière (l'œuvre composite) est seul investi tant du droit moral que des droits d'exploitation (si bien sûr le critère d'originalité est respecté). Mais il y a une limite : celle fixée par les droits de l'auteur de l'œuvre principale (il y a en effet concurrence des droits).

Bien entendu, ces régimes d'œuvres peuvent trouver à s'appliquer dans notre domaine. L'on peut même dire que ce sera souvent le cas en pratique, car les études statistiques demandées par l'utilisateur (prenons le cas d'un chercheur) le sont par exemple pour développer et consolider ses travaux et, dans ce cas, il est évident que l'on retrouvera « trace » de ces enquêtes dans ses futures publications.

Il faut donc respecter certains principes, définis pour la plupart par la jurisprudence elle-même, car le code est fort peu disert en ce domaine[29]. Notamment, l'autorisation de l'auteur de l'œuvre première[30] doit être donnée pour toute exploitation de l'œuvre composite[31]. Si ce n'est pas le cas, l'œuvre seconde est une contrefaçon[32]. Autre exemple, l'auteur de l'œuvre première peut en poursuivre ou en reprendre l'exploitation, malgré son incorporation dans une autre œuvre[33].

3. L'exploitation des données

« Exploiter » les données, c'est plus que les utiliser. Cette exploitation, effectuée par une personne autre que l'auteur, doit faire l'objet d'un contrat, qui est minutieusement réglementé par le code de la propriété intellectuelle. L'article L.131-3 en effet prévoit que « *la transmission des droits de l'auteur est subordonnée à la condition que chacun des droits cédés soit délimité quant à son étendue et à sa destination, quant au lieu et quant à la durée.* » Pour céder ses droits d'exploitation (que sont le droit de reproduction[34] et le droit de représentation[35], en plus des droits autorisant l'adaptation, la traduction,… qui sont différents des premiers) sur son œuvre (par exemple à son éditeur), un tel contrat devra être signé. Il doit donc comporter précisément la liste des droits cédés, et les délimitations de l'exploitation quant à son étendue, sa destination, au lieu et à la durée.

Un accent particulier doit être mis sur la destination de l'œuvre. Elle doit être en effet très précisément décrite dans le contrat d'exploitation. Par exemple, si l'auteur entend que son œuvre soit seulement utilisée à des fins de recherche et qu'il découvre par la suite que ce n'est pas le cas (son œuvre fait l'objet d'une exploitation commerciale), il pourra agir sur le fondement de son droit de destination, que lui accorde très nettement la jurisprudence et même indirectement cet article L.131-3 du code de la propriété intellectuelle[36].

L'on doit donc passer obligatoirement par l'accord de l'auteur, ou celui de toute personne, physique ou morale autorisée. Cela peut être le cas du Lasmas qui, « recueillant » les droits d'exploitation sur les études statistiques (auprès des titulaires des droits), peut les accorder, exclusivement ou non, c'est au choix (mais cela doit être mentionné dans les contrats), à nouveau à un tiers, qui pourra alors, plus que les utiliser, les exploiter à son tour.

---

[29]. L'article L.113-4 se borne en effet à préciser (sans surprise) que « l'œuvre composite est la propriété de l'auteur qui l'a réalisée, sous réserve des droits de l'auteur de l'œuvre préexistante ».

[30]. Ou de la personne habilitée par lui à le faire (exemple : son éditeur).

[31]. T.G.I. Paris, 8 mai 1969, Dalloz 1970, sommaires commentés, p.7.

[32]. Cour d'appel de Paris, 13 janvier 1993, Juris-Data n°20603.

[33]. Cour de cassation, 10 mars 1993, Juris-Data, n°558.

[34]. La reproduction est la « fixation matérielle de l'œuvre par tous procédés qui permettent de la communiquer au public d'une manière indirecte » (article L.122-3 du code de la propriété intellectuelle).

[35]. La représentation est la « communication de l'œuvre au public par un procédé quelconque » (article L.122-2 du code de la propriété intellectuelle).

[36]. Voir plus généralement sur cette question, Pollaud-Dulian, Le droit de destination, le sort des exemplaires en droit d'auteur, LGDJ, 1989.

2. Droit d'auteur et accès aux données

René Padieu

Commission de déontologie, Société Française de Statistique

Les réflexions qui suivent sont une contribution à la mission de Mme Silberman. Elles ont été en partie inspirées par la lecture de Ph. Chevet (Cecoji) : « Études statistiques et droit d'auteur » ; mais elles n'en sont pas un commentaire. Elles incorporent aussi mon intervention lors de la réunion du 8 avril 1999 autour de Mme Roxane Silberman.

I - Contenu du droit

Plutôt que de se centrer sur un « droit d'auteur », on devrait analyser en quoi consiste le droit sur une donnée (son contenu) :

1 - intégrité :

– droit d'altération : modifier ou détruire (« droit d'abuser »), qui est un droit objectif,

– droit au respect de l'intégrité, qui est un droit subjectif *erga omnes* ;

2 - usage :

– droit d'user, soi-même,

– droit de décerner l'usage, d'autoriser autrui, nommément ou généralement, à utiliser ;

3 - contrepartie : droit de convenir d'obligations à la charge du bénéficiaire d'une cession (par « cession », on entend aussi bien la remise matérielle d'une copie des données, l'organisation de leur consultation ou le transfert de tout ou partie des droits évoqués ici) :

– droit de restriction (non-transmission à un tiers, usage défini, temps défini, dispositions de sécurité, etc.),

– contre-prestation : paiement (achat ou location), compte rendu de l'usage fait, etc.

La détention de données ne confère pas nécessairement l'ensemble de ces droits et privilèges. La transmission des données ne transfert pas *ipso facto* tous les droits dont le cédant était titulaire (infra § III).

II - Divers types de données

S'agissant d'information construite à partir de données personnelles, on peut distinguer trois cas :

– la donnée individuelle, rattachable à une personne précise,

– un ensemble de données individuelles,

– une information synthétique (statistique) déduite d'un ensemble de données individuelles mais non rattachable à aucune personne en particulier.

Pour la deuxième catégorie, on relèvera la définition donnée par l'art. L.112-3 (cité par Ph. Chevet) : « *On entend par base de données un recueil d'œuvres ou de données ou d'autres éléments indépendants et disposés de manière systématique et individuellement accessibles.* » Cette définition est peut-être plus satisfaisante que celle de « fichier » évoquée dans les textes de protection des données.

Deux termes[37] y sont notables :

– *« individuellement accessibles »* : c'est en effet la possibilité d'un usage personnalisé qui est ici préservé, sous condition toutefois que la personne y soit identifiée, identifiable ou réidentifiable. C'est cet accès individuel qui est en cause dans la protection des données. Parfois, cet accès est la finalité même de la base de données : gestion administrative, par exemple. Parfois, il n'est pas dans la finalité de la base de données mais résulte de sa constitution. Ainsi, dans l'utilisation statistique, les données individuelles sont un passage obligé, mais l'individualité de l'information n'est pas visée. (Noter néanmoins que la finalité statistique ne recouvre pas uniquement l'élaboration de résultats synthétiques : elle englobe quelquefois aussi la possibilité de réobserver les mêmes personnes, pour vérification, actualisation ou enquête additionnelle.)

– *« systématiquement »* marque la finalité du recueil : on s'intéresse à l'ensemble d'une population ou catégorie. Ceci englobe le cas d'utilisations personnalisées, potentiellement étendues à toutes les personnes couvertes, et le cas d'utilisations impersonnelles (statistiques).

III - Fondement des droits

Il convient ensuite de décider ce qui met quelqu'un en capacité d'exercer tout ou partie des droits ci-dessus. Je serais d'avis d'éviter le terme de « propriété », qui semble porteur de plus d'équivoques que de clarté.

Pour les données personnelles isolées, on admet généralement que la personne concernée dispose a priori de l'ensemble des droits et privilèges énoncés au § I. Il est admis qu'un motif d'ordre public peut la contraindre à déclarer ces données, voire même à en publier certaines (comme les dirigeants et le bilan des sociétés commerciales). Reste à préciser ce que la personne ou l'organisme dépositaire peut en faire, ce qui peut être défini aussi par référence à la liste du § I.

Dans le cas d'une base de données personnelles, les personnes concernées par celles-ci sont distinctes de l'auteur ou détenteur : il n'a donc pas un droit premier sur elles. De même, cette information n'est pas sa création : ce n'est pas une « œuvre de l'esprit »[38]. Il ne devrait donc pas pouvoir se prévaloir d'un droit : ni de les altérer, ni de décider d'une cession ni d'exiger des contreparties à celle-ci. En revanche, il a des obligations tant envers les personnes concernées qu'envers la puissance publique s'il tient d'elle le pouvoir de collecter, conserver et utiliser les données. Le détenteur n'est certes pas propriétaire et, s'il a quelques droits, il a certainement des obligations.

L'une de ses obligations peut être justement de mettre les données en cause à la disposition de tiers déterminés : une disposition, d'ordre publique aussi, peut par exemple poser que la recherche ou la statistique pourra mobiliser ce stock de données. Dans ce cas, le détenteur aurait l'obligation de fournir les données sans être en état de stipuler d'autres conditions ou contreparties que celles qu'il a expressément mandat d'exiger. (Et des obligations seraient aussitôt à la charge du tiers bénéficiaire.)

Une question à ce stade est de savoir si les obligations à la charge du détenteur (primitif ou secondaire) permettent de dessaisir les personnes concernées de tout ou partie de leur droit premier.

---

[37]. « indépendants » combiné avec « individuellement » ouvre une troisième perspective, qui n'était sans doute pas présente au législateur. Il n'y a pas nécessairement identité entre les « éléments » de la base et les « individus » concernés, bien que cette correspondance soit le cas largement dominant. Néanmoins, on peut concevoir une transformation qui, à plusieurs individus, associe une donnée (impersonnelle, donc), et qui puisse le faire de différentes façons : de sorte qu'à n individus correspondent m éléments transformés. (C'est le principe de l'hologramme.) Si ceux-ci, ensemble, contiennent toute l'information des n individus originaux, mais mêlée, une utilisation statistique reste possible, tandis que l'accès individuel aux éléments transformés ne dévoile rien des individus. (Seule la transformation inverse le permettrait.)

[38]. On pourrait toutefois observer que si les informations élémentaires ne sont en effet pas une œuvre de l'esprit, celui qui a constitué la base a créé une telle œuvre par leur assemblage. Il peut aussi avoir contrôlé la qualité ou la cohérence des données personnelles ou encore, avoir rapproché des données de deux sources relatives aux mêmes personnes, créant ainsi en partie une information les concernant. Toute peine méritant salaire, on peut alors lui accorder une protection ou un privilège quant à l'utilisation par des tiers de ces données ou de leur assemblage.

La position générale des protecteurs de données est que les personnes concernées conservent certains droits (être informées des cessions ou utilisations non annoncées initialement, pouvoir les refuser, avoir un droit d'accès et de rectification, etc.). Ceci semble bien justifié lorsque ces utilisations ultérieures visent les personnes. En revanche, on peut arguer que, dès lors que l'utilisation est uniquement impersonnelle (anonyme, statistique), il n'y a plus de justification à maintenir les personnes dans les droits en cause : « point d'intérêt, point d'action ».

Cette vue pourrait invoquer l'art. 7 f) de la Directive du 24 octobre 1995 : l'intérêt et les droits et libertés fondamentaux de la personne n'étant pas menacés, ne sauraient prévaloir sur l'intérêt légitime poursuivi par l'utilisateur. De même, on invoquerait l'art. 8 (relatif aux données sensibles) § 3, 4 et 5 dans la mesure où la personne effectuant le traitement est soumise à des obligations de secret appropriées. Enfin, on relèvera surtout l'art. 13 §2 : lorsqu'on peut exclure « que les données puissent être utilisées aux fins de mesures ou de décisions se rapportant à des personnes précises » et « où il n'existe manifestement aucun risque d'atteinte à la vie privée ». Bien entendu, redisons-le, ce dessaisissement implique que des obligations strictes, juridiques et techniques, pèsent sur le détenteur. Fondé sur l'absence de grief lorsqu'une utilisation personnalisée est exclue, il ne tient qu'autant que celle-ci est effectivement empêchée.

Si ce raisonnement était suivi, on pourrait, au prix de protections exigeantes, faire l'économie d'autres procédures et recours qui par contre gênent considérablement la recherche.

3. La déontologie des statisticiens

Philippe Amblard, Cecoji, CNRS

La statistique est une science dont les méthodes mathématiques ne sont pas sans conséquence sur la société en général et sur ces sujets d'études en particulier. C'est pourquoi, très tôt, les statisticiens, conscients des dangers, se sont pourvus de codes déontologiques.

Quelle est donc l'origine et la légitimité de cette démarche éthique (I) ?

Et ces codes sont-ils effectifs ? Comment sont-ils concrètement appliqués (II) ?

I - Origine et légitimité des codes déontologiques en matière statistique

La déontologie des statisticiens est issue d'une démarche ancienne de leurs associations professionnelles (A) qui s'inscrit dans le respect de la réglementation nationale et internationale (B), protectrice des trois acteurs de l'enquête statistique (la personne interrogée, le commanditaire et l'institut) (C).

A/ Une démarche éthique des associations professionnelles

Primordial à l'activité statistique (1), la démarche éthique s'est concrétisée par l'adoption de codes. Mais quelles sont leurs finalités ? (2)

1/ Importance de la démarche éthique

Comme l'explique Mme Marcia Freed Taylor[39], « *tout manquement* – aux engagements de respect des personnes, sujets des enquêtes statistiques – *risque de compromettre gravement, pour les futurs chercheurs, la possibilité de collecter des données sûres et utiles ; il risque en effet de limiter l'accès aux données statistiques collectées par d'autres.* »

Tout repose donc sur la confiance du public : confiance dans le fait que les enquêtes statistiques sont réalisées en toute honnêteté, en toute objectivité. Une véritable enquête statistique, dont les résultats sont représentatifs de la réalité, suppose une collaboration volontaire des répondants. Ceux-ci ne doivent donc pas être importunés ou éprouvés une gêne.

C'est pourquoi les associations professionnelles du milieu de la statistique tels que l'ESOMAR[40] ou l'IIS[41] sont les initiateurs d'une démarche éthique, afin que la statistique, par la qualité de ces méthodes, garde la confiance de la société.

2/ Finalités des codes déontologiques.

Le premier code a été publié par l'ESOMAR en 1948. Mais ce n'est que depuis 1976 que l'ESOMAR publie conjointement avec la Chambre de Commerce Internationale le Code International CCI/ESOMAR. La version la plus récente de ce code date de 1995.

En France, les sociétés membres de l'association Syntec Études Marketing et Opinions[42] appliquent ce code. L'objet du code est d'exposer, synthétiquement *les principes déontologique de base qui doivent gouverner la pratique des études de marché et d'opinion.*

Le deuxième texte de référence en matière de déontologie de la statistique est l'œuvre de l'Institut International de Statistique qui réunit tous les grands instituts nationaux de statistique. Adopté à Amsterdam, en 1985**, la Déclaration de l'IIS sur l'éthique professionnelle** se veut un guide qui exprime les valeurs professionnelles reconnues par la profession. L'objet de cette déclaration n'est pas de réglementer, mais d'avoir un rôle éducatif, pédagogique.

---

[39]. Marcia Freed Taylor, Considérations éthiques dans la recherche transnationale européenne, Revue Internationale des Sciences Sociales, n° 142, 1994.

[40]. Association Européenne pour les Études d'Opinion et de Marketing.

[41]. Institut International de Statistique.

[42]. Cette associations professionnelle regroupe 44 sociétés privées qui représentent 60 % du marché des études en France. Pour plus d'informations voir son site Internet : http://www.syntec-etudes.com

B/ Le respect de la réglementation nationale et internationale

Outre des valeurs éthiques, le milieu statistique se doit de respecter un cadre juridique d'origine internationale ou nationale (1). Cela est d'autant plus vrai pour les organismes de statistique du secteur public (2).

1/ Cadre juridique commun au milieu de la statistique

La pratique statistique, comme le rappelle les codes déontologiques, doit respecter certaines règles juridiques.

Mis à part les droits des « fichés »[43], la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés impose principalement le respect du principe de finalité, une obligation de sécurité, de confidentialité et une interdiction de collecte des données dite sensibles.

La Cnil[44], en collaboration avec les principaux acteurs de la statistique, a aménagé les obligations légales. Seuls les traitements statistiques dont la finalité est légitime et correspondant aux missions de l'organisme, peuvent être autorisés par la Cnil.

Concrètement, après concertation avec le SYNTEC, les instituts de sondages bénéficient d'une procédure allégée de déclaration : une déclaration globale par type d'enquête reconductible annuellement. Par contre, il leur est interdit de collecter toutes informations faisant apparaître les origines raciales, les opinions politiques, philosophiques ou religieuses et les appartenances syndicales. De même, la confidentialité sur l'identité des interrogés, comme la sécurité des données collectées s'imposent aux instituts de sondages.

La directive européenne du 24 octobre 1995 reprend les mêmes garanties que la loi française de 1978 quant à la protection des personnes physiques à l'égard du traitement des données à caractère personnel. Mais elle reconnaît les spécificités des traitements à des fins statistiques. Des aménagements sont concédés (traitement ultérieur des données, exception à l'obligation d'information, limitation à l'exercice du droit d'accès)[45], sous couvert de garanties appropriées et l'absence d'atteinte à la vie privée.

Ces garanties pourraient être l'œuvre conjointe des états par des lois générales ou sectorielles[46] et des organisations professionnelles.

Dans le cadre de la transposition de la directive au sein des législations nationales, l'article 27 de la directive *« encouragent l'élaboration de codes de conduite destinés à contribuer, en fonction de la spécificité des secteurs, à la bonne application, des dispositions nationales prises par les états membres en application de la directive »*. Selon les instances européennes, *« les codes de conduite… peuvent être un instrument utile pour fournir des indications sur les moyens par lesquels les données peuvent être rendues anonymes et conservées sous une forme ne permettant plus l'identification de la personne concernée. »*

C'est pourquoi, en France, dans le cadre de l'aménagement des exceptions de la directive européenne, la SFdS[47] a créé une commission de déontologie, en vue d'élaborer un code qui a pour ambition d'être le corpus de garanties apportées par l'ensemble des statisticiens. Ce code est cours d'élaboration.

2/ Le cadre juridique spécifique au milieu de la statistique publique

En contrepartie de l'obligation de réponse aux enquêtes statistiques publiques[48], la loi du 7 juin 1951 impose aux agents de l'Insee et des services ministériels de statistique le secret professionnel

[43]. En vertu de l'article 27 de la loi « Informatique et libertés », les personnes auprès desquelles sont recueillies des informations nominatives doivent être informées du caractère obligatoire ou facultatif des réponses, des conséquences en cas de défaut de réponse, et disposent d'un droit d'accès et de rectification.
[44]. Commission Nationale Informatique et Libertés, chargé de veiller au respect de la loi du 6 janvier 1978.
[45]. Pour plus de précisions, se reporter à la note relative à la vie privée.
[46]. Cf. Considérant 23 de la Directive.
[47]. Société Française de Statistique.
[48]. Article 3 de la Loi n° 51-711 du 7 juin 1951

sur les données dont ils ont à connaître et en particulier celles permettant l'identification des personnes physiques ou morales.

Il est à noter que le code international CCI/ISOMAR impose aussi cette règle au secret à l'ensemble de ses membres, en grande partie privés.

Enfin les normes simplifiées n° 19 et 26 de la Cnil ont aménagé le régime juridique des enquêtes statistiques publiques. La norme n° 19[49] concernant les enquêtes par sondages effectuées par l'État et les établissements publics, interdit que soient collectées ou traitées des données dite sensibles, c'est-à-dire celles révélant les origines raciales des personnes, leurs opinions politiques, religieuses ou syndicales, ainsi que leur mœurs.

La norme n° 26[50] ne concerne que les enquêtes statistiques réalisées par les services ministériels et par l'Insee (sauf les enquêtes soumis à visa ministériel). Une déclaration simplifiée pour ce type d'enquête est permise, sous couvert du respect des droits du fiché (art. 27) et l'interdiction de traiter des données dite sensibles (art. 31).

C/ Une déontologie protectrice des trois acteurs de l'enquête statistique

L'activité statistique met en relation trois acteurs : l'interrogé ou le répondant, le statisticien ou le praticien, le commanditaire ou le client (privé ou public) à l'initiative de l'enquête. Tout l'enjeu du code déontologique est de régler les rapports entre eux, de définir leurs droits et obligations dans un souci de probité et de qualité de l'activité statistique et de respect de la personne humaine. Les deux codes internationaux, références déontologiques du milieu français de la statistique, ont adopté la même approche, en définissant clairement les droits et les obligations de trois acteurs dans des parties distinctes, sans compter le rappel du respect du Droit et des principes généraux, obligations envers la société.

Tout d'abord le Code International CCI/ESOMAR, après une définition des termes et des acteurs (études de marché, praticien, client, répondant) édicte les règles essentielles sous quatre chapitres :

– Les règles générales (conformité avec les principes scientifiques établis, respect des lois) ;

– Les droits du répondant (participation volontaire, respect de son anonymat sinon demande d'autorisation, possibilité de vérifier facilement l'identité et la bonne foi du praticien) ;

– Responsabilité professionnelle des praticiens (pas de comportement conscient ou non pouvant entraîner le discrédit sur la profession, pas de fausse déclaration sur ses qualités et expériences, pas de critique injustifiée de ses confrères, des études offrant un bon rapport qualité/prix, une sécurité assurée pour les documents en sa possession, ne laisser diffuser que des conclusions d'études conformes aux données recueillies, n'avoir pour seule activité que l'étude statistique pour éviter les conflits d'intérêts) ;

– Droits et responsabilités mutuels des clients et des praticiens (respect des principes de concurrence loyale, informer le client si l'étude est multi-clients ou sous-traitée, non diffusion auprès des tiers des documents restant la propriété du client (informations fournies par lui, résultats de l'étude si elle n'est pas destinée à plusieurs clients), non-divulgation à des tiers des noms des clients, possibilité pour le client de contrôler la qualité de l'étude, informer le client des détails techniques, distinction nette des résultats et de l'interprétation, ne pas permettre l'usage de son nom pour assurer à une étude la conformité au code, informer le client de l'existence du code).

De la même manière, **la Déclaration de l'IIS sur l'éthique professionnelle**, après un préambule sur sa finalité, expose les obligations du statisticien envers chacun des trois acteurs.

---

[49]. Délibération n° 81-28 du 24 mars 1981 concernant les traitements automatisés à des fins statistiques d'informations nominatives extraites d'enquêtes par sondages intéressant des personnes physiques effectués par l'État et les établissements publics à caractère administratif, J. O. du 14 mai 1981.

[50]. Délibération n° 84638 du 13 novembre 1984 concernant les traitements automatisés à caractère statistique effectués, à partir de documents ou de fichiers de gestion contenant des informations nominatives sur des personnes physiques, par les services producteurs d'informations statistiques au sens du Décret n° 84-628 du 17 juillet 1984, J.O. du 1er Décembre 1984.

– Obligations envers la société (prévenir les interprétations et utilisations erronées de ces travaux, œuvrer pour une diffusion de la statistique la plus large possible, employer des méthodes objectives et impartiales) ;

– Obligations envers les commanditaires (établir clairement à l'avance les rôles de chacun, fournir avec impartialité au commanditaire pour son choix les différentes méthodes possibles, ne jamais s'engager sur la teneur des résultats futurs, garder secrètes les informations fournies par le commanditaire) ;

– Obligations envers les collègues (avoir un comportement professionnel qui préserve la confiance du public dans la statistique, permettre à ses confrères de contrôler les méthodes et les résultats, informer les personnes étrangères à la statistique avec qui on collabore des principes éthiques) ;

– Obligations envers les sujets d'enquête (éviter les intrusions injustifiées dans l'intimité des personnes, obtenir leur plein consentement ; en cas d'impossibilité d'obtenir leur consentement : respecter leur vie privée et leur probable refus, troubler le moins possible les personnes interrogées, assurer la confidentialité des données collectées, prendre toutes les mesures pour éviter l'identification des personnes interrogées).

II - Effectivité et pratique des codes

Les codes n'énoncent pas seulement les principes éthiques du milieu statistique, ils ont aussi pour ambition de les faire appliquer. Concrètement, la pratique déontologique passe par des procédures de résolution de conflits (A), voire des sanctions (B). Mais surtout, les codes sont effectifs par la prise de conscience qu'ils déclenchent chez les statisticiens (C).

A/ La résolution des conflits

Les principes exprimés par les codes éthiques définissent la conduite de la recherche, son organisation, les méthodologies à employer... Naturellement, ce cadre déontologique est utilisé par les associations professionnelles comme référence à la résolution de conflits dans le milieu statistique.

La formulation ou la conduite d'un projet statistique peut donner lieu à des dilemmes éthiques non résolus ou difficiles, voire des conflits. L'assistance ou les conseils d'associations professionnelles sont alors des moyens de promouvoir le respect des principes édictés par le code.

Ainsi l'ESOMAR[51], soucieuse de l'application du Code International CCI/ESOMAR, propose de donner des avis pour la mise en œuvre concrète de ce code, à l'occasion de difficultés rencontrées par les professionnels des études statistiques.

Il est également possible de faire appel à l'ESOMAR pour résoudre des désaccords d'ordre technique ou autre relatifs à des études de marché. Une procédure d'arbitrage ou les services d'un expert sont alors proposés.

B/ Les sanctions en cas de violation du code

Classiquement pour s'assurer du respect des règles énoncées, certains codes d'associations professionnelles ou d'instituts de recherches prévoient des sanctions. Le code, dans ce cas, peut être assimilé à un « règlement intérieur ».

Selon la gravité de la violation du code, les mesures « disciplinaires » peuvent être le blâme, la suspension, voire la radiation. Bien que les formes et les procédures diffèrent selon l'environnement : public ou privé, professionnel, ministériel, académique, scientifique, …, le mode d'application du code est identique.

Par exemple, pour le secteur privé des études d'opinions et de marché, en tant qu'association membre de l'ESOMAR, le Syntec Études se doit de procéder à une enquête, prendre les mesures nécessaires et informer la CCI et l'ESOMAR, en cas d'infraction au code présumée.

Les sanctions prévues sont la suspension ou la radiation du Syntec.

[51]. Association Européenne pour les Études d'Opinion et de Marketing.

C/ La prise de conscience aux exigences éthiques de la statistique

Le code est souvent un guide, une aide pour faire plus facilement prendre conscience de ces exigences éthiques à un milieu. C'est la voie choisie **par la Déclaration sur l'Éthique Professionnelle** de **l'Institut International de Statistique**.

Pour promouvoir la pratique des principes éthiques, la Déclaration se veut « éducative ». Deux prémisses sont à l'origine de ce type de code :

– La plupart des questions éthiques échappent à toute réglementation catégorique,

– Les décisions éthiques appartiennent à l'individu, non au groupe.

Donc, l'enjeu est de faire prendre conscience aux statisticiens des problèmes éthiques et dans un deuxième temps de « *leur permettre de fonder leurs jugements éthiques et personnels sur des valeurs et une expérience partagées* »[52]. La déclaration offre « *un cadre à l'intérieur duquel* ***le statisticien consciencieux*** *devrait pouvoir, pour l'essentiel, travailler à l'aise* »[53].

L'effectivité des principes est assurée par l'information du code. Donc tout repose sur la connaissance par le milieu statistique de ce code. Depuis quelques années, le milieu statistique est informé des questions éthiques. L'action de la Cnil et la prochaine transposition de la directive européenne n'ont fait qu'accentuer cette évolution.

Conclusion

La démarche éthique dans le milieu statistique est utile et efficace. Elle a permis de faire reconnaître les spécificités de la statistique tout en garantissant le respect des droits fondamentaux des sujets d'enquêtes en particulier. Les aménagement juridiques avec la Cnil, entre autres, sont les fruits de cette démarche. **Le code reste un « outil » efficace pour sensibiliser et guider les statisticiens confrontés à des problèmes éthiques**, surtout si les aspects éthiques les plus sensibles sont réglementés (secret de la statistique, nature nominatives des données…).

[52]. Préambule de la Déclaration de l'IIS sur l'éthique professionnnelle.
[53]. Idem.

4. Les notions de données directement nominatives et indirectement nominatives

Fabrice Mollo, Cecoji, CNRS

La réalisation d'une enquête statistique se doit à chaque fois de concilier deux besoins qui peuvent, parfois, sembler contraires : la nécessité de recueillir des données et la protection de la vie privée. Dans ce domaine, la France a été un des premiers pays à se doter d'un cadre juridique. Dès 1951, par la loi du 7 juin, certains traitements statistiques sont soumis à l'obligation de secret. Mais c'est la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés qui va vraiment marquer le début de la prise en compte de possibles atteintes à la vie privée du fait d'une manipulation de données sans contrôle. Bien que les professionnels de la statistique n'aient pas toujours bien accepté les avis de la Commission Nationale de l'Informatique et des Libertés (Cnil), il a bien fallu que s'instaure un dialogue pour permettre à chacun d'assurer ses missions. Ainsi, après des discussions parfois difficiles, les normes simplifiées 19 et 26 déterminent les conditions dans lesquelles les services de l'État peuvent utiliser des données nominatives.

Cette entente « raisonnée » va devoir connaître de nouveaux développements du fait de la généralisation des nouveaux moyens de communications que sont les réseaux informatiques. Internet en est bien sûr l'exemple le plus actuel, mais ce réseau mondial ouvert ne constitue qu'une des nombreuses possibilités offertes par les nouvelles technologies de l'information. Cette forme d'échange au-delà des frontières a conduit à la rédaction de la Directive Européenne 95/46 sur les flux de données transfrontières. Bien que la France n'en ait toujours pas achevé la transposition, il nous faut voir quelles modifications seront nécessaires. Même si les positions françaises ont été souvent entendues par les autres pays européens, les définitions des concepts de base de la directive manquent souvent de précision. Le flou ainsi créé risque de mener à une multitude d'interprétations, loin de la convergence souhaitée.

Pour pouvoir traiter toutes les questions ayant trait à la protection, l'utilisation des données nominatives et indirectement nominatives, il nous faut tout d'abord chercher à mieux en cerner la définition car, la loi de 1978 mise à part, les autres textes se réfèrent « aux données à caractère personnel ». Cette différence de termes influe en outre sur ce que l'on cherche à protéger : doit-on privilégier un droit sur les données ou le droit à la vie privée ? Après ces problèmes de définition, il nous faudra nous interroger sur ce que recouvre l'article 11 paragraphe 2 de la directive européenne, qui prévoit un régime d'exception pour la production de statistiques, puis aborder les conditions de diffusions de ces données.

I - Des Concepts difficiles à définir

Comme nous venons de le voir, les notions de données nominatives et indirectement nominatives sont le fruit d'une évolution juridique due à l'adoption successive de textes nationaux et internationaux. Pour cette étude, nous rapprocherons la loi française de janvier 1978, les lignes directrices de l'OCDE régissant la protection de la vie privée et les flux transfrontières de données à caractère personnel de 1980, la Convention 108 du Conseil de l'Europe de janvier 1981 et la Directive 95/46.

a) Informations nominatives ou données à caractère personnel

L'article 4 de la loi « Informatique et Libertés » énonce : « *sont réputées nominatives au sens de la présente loi les informations qui permettent, sous quelque forme que ce soit, directement ou non, l'identification des personnes physiques auxquelles elles s'appliquent, que le traitement soit effectué par une personne physique ou par une personne morale.* » Les autres textes utilisent le vocable de données à caractère personnel. Selon le rapport rendu par le Conseiller Braibant au Premier Ministre sur la transposition de la directive 95/46, ce dernier terme semble mieux adapté aux nouveaux développements technologiques compte tenu des moyens d'identification indirecte (recoupement des données, profils de personnalité…). Mieux vaudrait donc reprendre les définitions du texte européen.

Pourtant, la notion française d'informations nominatives se divise selon les avis de la Cnil en deux catégories : celles qui sont manifestement nominatives et celles qui le sont indirectement. Cette dernière classe s'applique particulièrement à la production de données statistiques. L'anonymisation des données est la règle. Mais lorsque les croisements s'effectuent selon des critères de tri trop fin, la présentation des résultats ne peut éviter l'identification de certaines personnes interrogées. La qualification d'indirectement nominative est certes moins précise que son équivalent européen, mais permet une meilleure protection de l'individu. L'enjeu est de mettre en place des critères plaçant ou non un traitement dans le champ de protection de la loi, selon l'intérêt que l'on voudra privilégier.

b) Protection de la vie privée contre protection des données

Dès son article premier, la directive place comme un de ses principaux objectifs la protection de la vie privée. Quant à savoir ce que recouvre cette expression, aucune définition – ou tentative – ne vient éclairer le lecteur. Encore une fois la vie privée sert ici de borne qu'il ne faut pas franchir. Sur le plan national la question suscite déjà de nombreuses controverses. Il semble alors difficile de vouloir une unité avec des partenaires ayant leur conception propre. La fracture se situe surtout entre l'Europe du continent et la définition anglo-saxonne de la « privacy ».

La jurisprudence de la Cnil paraît sans équivoque sur ce point, l'important n'est pas la nature ou le contenu d'une information ; c'est sa finalité, ce pourquoi on l'utilise. Le « right of privacy » s'exerce différemment. Dans une situation donnée, il faut mettre face à face l'intérêt du citoyen et celui de l'État. Cela ressemble plus à un droit au secret. Cette différence d'appréciation se retrouve lorsqu'il faut définir l'objet de la protection. La directive n'arrivant à cerner précisément la protection des données personnelles, n'y a-t-il pas un risque que certains pays préfèrent appliquer une simple protection des données (secret des informations, exclusivité des droits) au détriment d'autres plus respectueux des libertés individuelles ?

Après avoir tenté de déterminer les changements que va connaître la définition de donnée nominative (ou indirectement nominative), il nous faut essayer de prévoir ce que sera le régime d'exception prévu à l'article 11 de la Directive, ainsi que les problèmes de diffusion internationale liés à l'exigence d'un niveau de protection adéquat.

II - Comment organiser le nouveau régime d'exception et la diffusion de ces données ?

a) Le régime d'exception prévu par la Directive 95/46

Depuis 1978, il s'est établi entre la Cnil et les professionnels de la statistique un dialogue, quelquefois tendu, qui a aboutit à une compréhension des besoins de chacun quitte à remettre en cause certaines pratiques ou à s'interroger sur l'utilité réelle de certains traitements (le bien fondé de certaines questions). Il s'est donc peu à peu établi une sorte de régime dérogatoire dont la traduction la plus concrète reste les normes simplifiées 19 et 26. Avec la transposition de la Directive que va-t-il advenir de ce régime ? Le texte, dans son article 11 paragraphe 2, mentionne bien une exception à des fin statistiques, sans toutefois apporter plus de précisions. Le rapport Braibant se contente d'affirmer que cela doit être repris dans le futur texte.

Puisque sur le plan national nous disposons de peu d'éléments, il faut alors regarder dans d'autres états membres comment se fait l'équilibre entre la vie privée et les besoins statistiques qu'ils émanent du secteur public ou privé. Entre autres cas, nous nous intéresserons au Danemark, car il a mis en place un système de protection sociale très développé dont le simple fonctionnement engendre un grand nombre de traitement de données nominatives.

Le rapprochement des diverses politiques nationales nous fournira des exemples utiles pour le débat qui ne manquera pas de s'engager lors de l'aménagement du régime dérogatoire visé à l'article 11. D'autant plus que si l'institution Cnil se voit confirmée, selon les souhaits exprimés par le Conseiller Braibant, son fonctionnement et ses attributions devront subir quelques aménagements pour être conforme à la Directive. Tout ceci influe naturellement sur les conditions de production et l'utilisation des données statistiques.

b) Flux transfrontières et niveau de protection « adéquat »

Un des principes clés de la directive, lié à la transmission des données entre état membre ou entre un autre état hors CEE, oblige le destinataire de la transmission à disposer d'un niveau de protection

équivalent. Les articles 25 et 26 font l'objet de vives discussions. Pratiquement tous les commentaires venant de divers pays européens insistent sur le fait que le texte pose plus de questions qu'il n'apporte de réponse, notamment sur la définition même du terme adéquat et sur ses conséquences.

Concernant la définition, quel sens donner au mot « adéquat » ? S'il faut entendre équivalent, se pose alors la question des critères. Si cela désigne plutôt des conditions appropriées, cela suppose d'établir différentes catégories de traitements avec les conditions de protection qui leur sont propres. Concrètement la directive recommande de dresser des listes de pays, blanche pour ceux qui sont jugés aptes à recevoir ou envoyer des données tout en respectant les libertés, noire pour les autres.

La production statistique se trouve alors dans une position délicate. Les pays ne disposant pas de régime protecteur risquent de se voir conférer un avantage économique pour la réalisation de traitements de certaines données « sensibles ». De ce fait comment vont s'établir les listes ? Quelles sanctions peut-on envisager pour rendre le système efficace, sous peine de voir se développer des paradis numériques pouvant acquérir tous types de données en vue d'un commerce alors difficilement maîtrisable.

5. Le statut des données contenues dans les archives publiques et les conditions de leur utilisation

Frédérique Cormier, Cecoji, CNRS

L'utilisation des données contenues au sein des documents versés auprès d'un service d'archives au titre de la loi du 3 janvier 1979 sur les archives, ou déposés conformément aux dispositions qui régissent le dépôt légal, n'absout aucunement l'utilisateur d'une certaine vigilance dans la mesure où le dépôt et le versement ne font pas obstacle à l'application des dispositions relatives à la propriété intellectuelle, au régime applicable à la presse ainsi qu'à la protection de la vie privée.

En effet, les archives sont définies comme « l'ensemble des documents, quels que soient leur date, leur forme et leur support matériel, produits ou reçus par toute personne physique ou morale, et par tout service ou organisme public ou privé, dans l'exercice de leur activité » (article 1er, al. 1er de la loi du 3 janvier 1979).

Ainsi tout document peut être a priori qualifié de pièce d'archives, cette qualification n'est cependant retenue que pour les documents qui répondent à un critère fonctionnel, qu'ils soient produits ou reçus par une unité organique, de nature publique ou privée, dans l'exercice de ses fonctions.

Cette définition de l'archive qui permet de différencier le fonds d'archives des notions de collection ou de documentation et ce, dans le respect des principes archivistiques, vise à établir les documents qui bénéficient du statut des archives et non celui des données qu'elles contiennent.

La loi du 3 janvier 1979 sur les archives ne prend en considération les données contenues au sein des documents d'archives que dans le cadre des dispositions sur leur communicabilité. La loi de 1979 instaure des délais de communication dont la durée dépend de l'information que l'on tend à protéger.

Ainsi les dispositions de 1979 instaurent un délai de droit commun de 30 ans, puis laissent place à 5 délais spéciaux ; de 150 ans à compter de la date de naissance pour les dossiers médicaux, de 120 ans pour les dossiers personnels, de 100 ans pour les dossiers des affaires portées devant les juridictions à partir de la date clôture du dossier, y compris les délais de grâce, de 100 ans pour les enquêtes statistiques qui mentionnent des éléments d'ordre privé, personnel et familial, et de 60 ans pour ceux qui concernent la vie privée.

La nature des données contenues au sein des archives induit au régime de communication des archives privées, l'article 9 de la loi de 1979 énonce en effet que les services d'archives qui reçoivent des documents d'origine privée doivent respecter les conditions de communication imposées par les propriétaires, ce qui, dans la pratique, correspond aux informations que l'on tend à protéger lorsqu'il s'agit d'archives de nature publique.

Ainsi, les dispositions de la loi de 1979 n'ont pas eu vocation à établir un statut des données contenues au sein des archives même si la vocation scientifique des archives, la mise en place de la loi du 17 juillet 1978 qui organise l'accès aux documents administratifs, instaurent une certaine transparence et permettent une communication plus ouverte du patrimoine archivistique.

Les observations que nous venons de faire au sujet des dispositions relatives aux archives sont valables quant au régime des documents collectés dans le cadre du dépôt légal. En effet, le dépôt légal vise à constituer un patrimoine culturel national composé des diverses créations intellectuelles de l'homme.

Le dépôt légal a vocation à permettre l'accès du public aux collections constituées par les organismes institués dépositaires, sous réserve des secrets protégés par la loi et dans des conditions conformes au droit d'auteur (article 2 de la loi du 20 juin 1992).

Lorsque sont visés, comme assujettis au dépôt légal, les documents ayant fait l'objet d'une communication auprès du public, cela ne signifie en aucune manière que les documents déposés et offerts à la consultation sont susceptibles d'être réutilisés librement. Pour l'exemple, l'utilisation des documents qualifiés d'œuvre de l'esprit par le droit d'auteur, est subordonnée au consentement de l'auteur ; l'accord du producteur d'une base de données doit être requis, lorsque l'on veut extraire ou utiliser les données qui la constituent.

Ainsi, le statut des données contenues au sein de documents que l'on peut qualifier d'archives au sens commun du terme comme source de notre mémoire, dépend essentiellement du statut que les documents possédaient avant même leur entrée au sein d'un service chargé de leur collecte et de leur préservation.

Il faut s'interroger sur la nature publique ou privée des données – ce qui ne présume pas nécessairement de la nature de l'établissement qui détient les données, il faut alors rechercher quel est l'organisme producteur – si les dispositions sur le droit d'auteur ont vocation à être appliquées, sur la nature de l'information collectée au travers de ces données, données personnelles, nominatives, etc.

Il faut en outre différencier l'accord de l'auteur, du propriétaire, du détenteur de l'archive, d'accéder à l'information, aux données contenues, de l'autorisation d'utiliser ces mêmes archives, ces données. Pour l'exemple, l'article 10 de la loi du 17 juillet 1978, qui institue l'accès aux documents administratifs, réserve quant à leur utilisation les droits afférents à la propriété intellectuelle et interdit toute utilisation commerciale ultérieure des données ou documents par les bénéficiaires du droit à la communication.

Cette question pourra être l'objet de thèmes approfondis :

– propriété matérielle des données et propriété intellectuelle,

– protection des données et droit à l'information,

– exploitation des données et respect des droits afférents.

Aspects de droit comparé :

Le caractère des données contenues dans les archives publiques est pris en compte dans l'ensemble des pays occidentaux lors de la communication des archives.

Si le délai trentenaire de droit commun est de règle dans les pays européens (Allemagne pour les Archives fédérales, Royaume-Uni, etc.), la protection des intérêts supérieurs et la sécurité de l'État, les intérêts individuels et ceux relatifs à la vie privée, ainsi que le secret bancaire et commercial et industriel imposent des délais spéciaux supérieurs.

La transparence administrative permet au sein des pays européens d'accéder librement aux archives mais ne permet pas, à l'instar du système français, une utilisation libre des données collectées.

Bibliographie

ADIJ, *L'information juridique : contenu, accessibilité et circulation. Défis politique, économique et technique*, Congrès international, Paris, 22-23 octobre 1998.

Hervé Bastien, *Droit des archives*, La Documentation française et Direction des Archives de France, 1996.

Guy Braibant, *Les archives en France*, rapport au Premier Ministre, collection des rapports officiels. La Documentation française 1997.

CADA, *L'accès aux documents administratifs*, 8ème rapport de la CADA, La Documentation française, 1995.

Anne-Marie Chabin, Réflexions, méthodes et prospectives – La Communicabilité des archives : l'information, le document, le dossier, *La revue administrative*, n° 286, p.415 - 422, *Panorama de la presse juridique*, octobre 1995, n° 60, p.9-15.

Cnil, *Voix, image et protection des données personnelles*, La Documentation française.1997.

Jean Driol, La connaissance et la diffusion des données publiques en France, *Revue française d'administration publique*, n° 72, octobre-décembre 1994, p.661-668.

Michel Duchein, *Les obstacles à l'accès, à l'utilisation et au transfert de l'information contenue dans les archives*, Étude RAMP, PGI-83/WS/20.

Les Études du Conseil d'État, *Pour une meilleure transparence de l'administration : étude sur l'accès des citoyens aux données publiques*, La Documentation française, 1998.

Jean Favier, *Les archives*, Que sais-je ? n° 805, 4ème édit. 1985.

La Gazette des archives, Droit à l'information, droit au respect : la communication des archives contemporaines, n° 130-131, 3e et 4e trimestres 1985.

Philippe Gaudrat, *Commercialisation des données publiques*, Rapport OJTI, La Documentation française, Paris, 1992.

Maurice Ronai, *Données publiques : accès, diffusion, commercialisation*, La Documentation française, Problèmes politiques et sociaux, n°773-774, 1er novembre 1996.

Dispositions législatives

– Loi n° 79-18 du 3 janvier 1979 sur les archives, J.O. du 5 janvier 1979 et rectificatif au J.O du 6 janvier 1979.

– Décret n° 79-1037 du 3 décembre 1979 relatif à la compétence des services d'archives publics et à la coopération entre les administrations pour la collecte, la conservation et la communication des archives publiques. J.O. du 5 janvier 1979

– Décret n° 79-1038 du 3 décembre 1979 relatif à la communication des documents d'archives publiques, J.O. du 5 décembre 1979

– Circulaire du 2 octobre 1997 relative à l'accès aux archives publiques de la période de 1940-1945 par Lionel Jospin à Paris, le 3 octobre 1997 in J.O. n° 14339 du 3 octobre 1997

– Loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal. J.O., 18 juillet 1978, modifiée par la loi n° 79- 587 du 11 janvier 1979, J.O. du 12 juillet 1979.

– Circulaire du 14 février 1994 relative à la diffusion des données publiques. J.O.19 février 1994, n° 2864.

– Arrêté du 8 août 1996 fixant la liste des documents administratifs non communicables au public, J.O.13 septembre 1996.

– Loi n° 92-546 du 20 juin 1992 relative au dépôt légal, J.O. du 23 juin 1992.

– Décret n° 93-1429 du 31 décembre 1993 relatif au dépôt légal, J.O. du 1er janvier 1994

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

– Délibération de la Cnil n° 88-52 du 10 mai 1988 portant adoption d'une recommandation sur la compatibilité entre les lois n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, et n° 79-18 du 3 janvier 1979 sur les archives.

– Directive 95/46/CE du Parlement européen et du Conseil du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation des données. JOCE (L) 281, 23 novembre 1995, P.0031

– Convention 108 du Conseil de l'Europe du 28 janvier 1981 sur les principes liés au traitement des données personnelles telles que la voix et l'image.

– Délibération de la Cnil n° 88- 52 du 10 mai 1988 portant adoption d'une recommandation sur la compatibilité entre les lois n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, et n° 79-18 du 3 janvier 1979 sur les archives.

Annexe VI : Une zone sécurisée - L’exemple de l’accès au Recensement 1999

Une zone sécurisée - L'exemple de l'accès au recensement 1999

Irène Fournier- Mearelli, Yolande Kan, Nicole Dausque, Alexandre Kych)

Le cadre général

Une institution d'archivage et de diffusion de grandes enquêtes pour les chercheurs peut être conduite à gérer des données à caractère personnel (panels, données infracommunales du recensement…) dans le cadre d'une autorisation de la Cnil. Elle devra nécessairement offrir des garanties de sécurité tant au niveau conservation des données, qu'au niveau accès et traitements, ces derniers étant examinés par la Cnil. Pour ce faire elle devra disposer d'une « zone sécurisée » (au sens physique : local et machines) et d'un « guichet » ; ceci permettra de créer un dedans et un dehors et de reconstituer les deux modes d'accès offerts actuellement par l'Insee.
L'Insee est très attaché à l'aspect juridique du montage : cela veut dire une institution, des fonctions et des utilisateurs parfaitement définis et une « traçabilité » des responsabilités.
Enfin, l'Insee est très prudent devant tout accès par réseau, pratique qui se développe à l'étranger. Pour ce faire l'architecture réseau, système et logiciel, devra tout particulièrement tenir compte de l'aspect sécurité et les moyens techniques appropriés devront être mis en œuvre.

Vers une architecture de réseau

L'implantation du site

Il paraît judicieux, dans le choix de l'implantation du site de s'orienter dès à présent vers une solution permettant un accès à RENATER 2 (déploiement à partir de l'été 1999).

Matériels

Une étude précise des différents types de services (données d'enquêtes accessibles, pré-traitements) que l'institution devra offrir, permettra de définir les caractéristiques et le nombre d'équipements réseaux et systèmes nécessaires. On peut d'ores et déjà prévoir de répartir des serveurs dans la zone sécurisée (pour le rp1999), accessibles uniquement à partir de machines clientes authentifiées et situées sur le réseau interne de l'institution ; toute intervention dans la zone sécurisée se fera sous contrôle manuel. Une zone à accès semi-ouvert (utilisation de moyens d'authentification par carte à puces et/ou mot de passe à usage unique) permettra aux chercheurs (après signature d'une charte de confiance) d'accéder à des données d'enquêtes considérées comme moins sensibles. Pour tous les transferts de données de la zone sécurisée vers la zone semi-ouverte, une intervention humaine sera obligatoire pour réaliser le transfert afin de valider à la fois le type de données et le traitement demandés. On s'orientera vers une recopie systématique du sous-ensemble demandé et fourniture d'un accès authentifié et temporaire pour accéder à partir de la zone semi-ouverte à ces données et traitements temporaires. Il est important, **dès la création de l'institution, d'offrir des services dans un contexte sécurisé**, les contraintes qui en découlent seront mieux acceptées.

Les types de matériels de connexion (routeurs, commutateurs,…) seront choisis après étude complète des besoins (quantifier les débits, définir les liens entre serveurs et clients, choisir les logiciels d'application …). Une architecture à base de *virtual local area network* peut tout à fait être envisagée.
Une épine dorsale à **100 Mbits/s** sera la base de départ.
Les machines serveurs et clientes seront choisies en fonction des besoins dans le monde système Unix, Win9x ou Win/NT.
Une attention toute particulière sera faite dans le choix des équipements sur lesquels seront stockées les enquêtes (redondance de disques, technologie « RAID »).

Dans le cas de bases réparties, les données d'enquêtes étant des données stables, il sera préférable d'envisager la duplication de la base plutôt que d'utiliser des mécanismes de canaux sécurisés (type « VPN ») complexes à mettre en œuvre et à gérer.
La sécurité absolue passe par une redondance des équipements, celle-ci peut être en partie réalisée par des sauvegarde sur des sites sûrs à l'extérieur.

Logiciels

Authentification des utilisateurs : SSH, S/Key
Logiciels sécurisés : HTTPS, SSL
Logiciels de service : SPSS, SAS, GLIM, LISREL, STATA, système SGBD, Compilateur , C, C++, logiciels de cartographie ...
Logiciel de sauvegardes (totales et incrémentales)

Quelques précisions

– un serveur réservé au WEB, pour l'interrogation de la base documentaire et au développement, plus disque de 3x9G pour le stockage de la documentation,
– un serveur de noms,
– un serveur pour les autres enquêtes (Emploi, FQP, PCV, etc.), connecter pour les gens du site, avec un « rack de 5x9 G ».

Personnels

– Un ingénieur réseau et système avec compétences sécurité,
– Un ingénieur spécialisé sécurité et bases de données,
en plus de
– Ingénieur documentaliste,
– Ingénieur développement,
– Assistant Ingénieur connexion PC.

Architecture :