# 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 États-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 Science 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[1], 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

1. 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[2] 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

2. 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*[3]). 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

3. 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 données.*

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

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[1].

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

1. 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).

*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.*

***e) Gestion du droit d'usage***

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 statistique : 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.

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

*Une charte constituera le cadre de la transaction entre le Centre d'archivage 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.