IHSN

Réseau International pour les Enquêtes auprès des Ménages

Diffusion des fichiers de microdonnées

Principes, procédures et pratiques

Olivier Dupriez and Ernie Boyko

IHSN Document de travail n°005
Août 2010

www.ihsn.org

Diffusion des fichiers de microdonnées
Principes, procédures et pratiques

Olivier Dupriez et Ernie Boyko

IHSN Document de travail n°005

Août 2010

Résumé

Les producteurs de données de tous pays sont confrontés à une demande croissante de microdonnées. Décider de la meilleure façon de diffuser ces données constitue un véritable défi. Ce défi est d'ordre technique, puisqu'il s'agit de mettre en place des procédures de documentation, de catalogage et de diffusion des données, mais également de nature juridique et éthique. Si les producteurs de données sont souvent parfaitement conscients du pouvoir et de l'importance des microdonnées, ils doivent cependant jongler entre cette demande et la nécessité d'assurer la confidentialité des informations fournies par les personnes interrogées. Cette exigence est imposée par les législations nationales en matière de statistiques et de confidentialité des données, et fait souvent l'objet d'un engagement fourni aux répondants au moment de la collecte des informations. Cela implique l'établissement de politiques et de procédures définissant les conditions d'accès aux microdonnées. Le présent document contient une description générale de ces politiques et de ces procédures, et recense les bonnes pratiques en la matière.

Les auteurs

Ernie Boyko a longtemps œuvré au sein de Statistique Canada, où il a successivement dirigé les divisions de la statistique agricole, de la planification intégrée et des systèmes de gestion, et de la diffusion informatique des données, avant de prendre la tête des opérations du recensement de 1991. Il a notamment supervisé les travaux de l'*Initiative de démocratisation des données* (IDD). Ernie Boyko est membre actif de l'Association canadienne des utilisateurs de données publiques (ACUDP), ainsi que de l'Association internationale pour les services et techniques d'information en sciences sociales (IASSIST).

Olivier Dupriez, économiste-statisticien confirmé, est membre du Groupe de gestion des données sur le développement de la Banque Mondiale. Il est également coordinateur du Réseau international pour les Enquêtes auprès des Ménages (IHSN). Il coordonne ainsi les programmes d'assistance technique d'un grand nombre de pays dans les domaines liés à la documentation et à la diffusion de microdonnées.

Remerciements

Le présent document a été élaboré par le réseau IHSN avec le soutien financier du mécanisme DGF (Development Grant Facility) de la Banque Mondiale, dont la subvention (n° 4001009-06) est gérée par le Secrétariat de PARIS21 à l'OCDE.

Il a été préparé par Ernie Boyko et Olivier Dupriez avec l'aide des personnes suivantes, qui ont servi de sources d'information ou ont formulé des remarques et des suggestions sur le document : François Fonteneau (PARIS21, OCDE), Julia Lane (*National Opinion Research Center*, Université de Chicago), Johan Mistiaen (Banque Mondiale), Dennis Trewin et Wendy Watkins (Université de Carleton, Canada).

Le présent document intègre également les discussions menées avec de nombreux collègues des agences membres du réseau IHSN, ainsi qu'avec les statisticiens officiels d'un certain nombre de pays. Il a été préparé à la publication par John Wright.

Le présent document (ou une copie à jour de celui-ci) est disponible sur le site Internet d'IHSN à l'adresse suivante : www.ihsn.org.

Référence

Dupriez, Olivier et Ernie Boyko. 2010. Diffusion des fichiers de microdonnées. Définition des politiques et des procédures, Réseau International pour les Enquêtes auprès des Ménages (IHSN), document de travail n°005.

Table des matières

Annexes

Liste des figures

Liste des encadrés

Sigles et acronymes

ABS	Australian Bureau of Statistics : bureau australien de statistique
ACS	American Community Survey : enquête annuelle menée aux Etats-Unis auprès d'échantillons de la population
API	Application programming interface : interface de programmation
ASCII	American Standard Code for Information Interchange : format normalisé pour l'échange d'informations
CDR	Centre de données de recherche (Statistique Canada)
CDS	Contrôle de la divulgation statistique
CEE-ONU	Commission économique pour l'Europe des Nations unies
CENEX	Centre of Excellence for Statistical Disclosure Control : projet européen visant à dresser la liste des pratiques courantes en Europe en matière de protection des données
CESSDA	Council of European Social Science Data Archives : conseil européen des archives de données en sciences sociales
CSO	Central Statistics Office : office central de statistique irlandais
CURF	Confidentialised Unit Record Files : fichiers d'enregistrements unitaires confidentialisés
DCMI	Dublin Core Metadata Initiative : organisation soutenant les activités relatives à l'établissement d'un schéma de métadonnées générique
DDI	Data Documentation Initiative : projet visant à établir des normes de documentation technique en sciences sociales
DHS	Demographic and Health Surveys : programme mondial des enquêtes démographiques et de santé
DSNU	Commission de statistique des Nations Unies
EU	Etats-Unis d'Amérique
FMGD	Fichier de microdonnées à grande diffusion
GPS	Global Positioning System : système de géolocalisation fonctionnant au niveau mondial
HTML	HyperText Markup Language : format de données conçu pour représenter les pages Web
HTTP	Hypertext Transfer Protocol : protocole de transfert hypertexte
ICPSR	Inter-University Consortium for Political and Social Research : consortium interuniversitaire pour la recherche en sciences politiques et sociales
IDD	Initiative de démocratisation des données (Statistique Canada)
IHSN	International Household Survey Network : Réseau International pour les Enquêtes auprès des Ménages
ISO/CEI	Organisation internationale de normalisation / Commission électrotechnique internationale
JSI	Job Submission Interface : interface permettant de publier et de consulter des offres d'emploi
LIS	Luxemburg Income Study : base de données résultant d'une étude sur les revenus
MCRDC	Michigan Census Research Data Center : centre de recherche sur le recensement
MICS	Multiple Indicator Cluster Surveys : méthodologie d'enquête à indicateurs multiples
MIT	Institut de technologie du Massachusetts
NCHS	National Center for Health Statistics : centre national des statistiques de santé (EU)
NCSA	National Center for Supercomputing Applications : centre américain de recherche et d'exploitation des applications haute performance
NDE	National Data Enclave : centre de données sécurisé
NORC	National Opinion Research Center (Université de Chicago)
NSD	Norwegian Social Science Data Services : archive norvégienne
NU	Nations unies
OAIS	Open Archival Information System : modèle de référence pour un système ouvert d'archivage d'information
OCLC	Online Computer Library Center : organisme mondial de recherche à but non lucratif offrant des services aux bibliothèques favorisant l'accès de ces dernières à l'information partout dans le monde
OCDE	Organisation de coopération et de développement économiques
ONG	Organisation non-gouvernementale

ONS	Office national de statistique
PDF	Portable Document Format : format de fichier développé par Abobe
PUMA	Public Use Microdata Areas : unités de microdonnées à grande diffusion
PUMS	Public Use Microdata Sample : échantillons de microdonnées à grande diffusion
SAS	Statistical Analysis System (logiciel)
SSN	Système statistique national
SNZ	Statistics New Zealand : institut de statistique néo-zélandais
SOAP	Simple Object Access Protocol : protocole orienté objets bâti sur XML
SQL	Structured Query Language : langage de base de données
UPE	Unité primaire d'échantillonage
UKDA	United Kingdom Data Archive : centre d'archivage du Royaume-Uni hébergé par l'Université d'Essex
UNF	Universal Numeric Fingerprint : signature électronique universelle
UNICEF	Fonds des Nations unies pour l'enfance
URL	Uniform Resource Locator : chaîne de caractères utilisée pour l'adressage des ressources sur le Web
USB	Universal Serial Bus : norme relative à un bus informatique en transmission série servant à connecter des périphériques à un ordinateur
XML	eXtensible Markup Language : langage de balisage extensible
XSL	Extensible Stylesheet Language : langage de structuration de données

Introduction

La collecte de données statistiques, destinée à soutenir les processus décisionnels privés et publics d'un pays, est une énorme entreprise généralement financée sur des fonds publics. Il incombe à tous les producteurs de données ainsi subventionnés, aux chercheurs et aux commanditaires d'assurer un retour sur investissement maximal en promouvant l'exploitation de ces données.

Les données socio-économiques, qui constituent l'axe principal de ce document, résultent de recensements, d'enquêtes par sondage et de systèmes d'enregistrement administratifs. Ces activités génèrent des microdonnées (niveau de définition le plus bas pouvant être observé, à savoir celui des personnes interrogées). Ces microdonnées peuvent ensuite être traitées (éditées, analysées et compilées) avant d'être mises à disposition des utilisateurs. Habituellement, on obtient ainsi des données agrégées se présentant sous forme de tableaux, graphiques, dossiers, rapports descriptifs et analyses. Le contenu de ces tableaux et rapports dépend de leur importance pour les producteurs de données et les commanditaires. La majorité des activités de collecte de données sont menées à des fins précises, dont la satisfaction est la priorité souvent absolue du producteur ou du commanditaire. Néanmoins, les microdonnées recueillies dans un but particulier peuvent souvent être utiles à beaucoup d'autres personnes, qui ne sont guère prises en compte au moment de la collecte. En d'autres termes, elles peuvent servir à d'autres fins que celles initialement prévues. En ouvrir l'accès à la communauté de chercheurs est un moyen économique et efficace de multiplier et de diversifier l'analyse et l'exploitation des informations recueillies. Une exploitation en profondeur de ces données offre des possibilités quasi-illimitées d'accès à de nouvelles connaissances.

La puissance croissante des ordinateurs et des logiciels depuis les années 1980 a rendu les microdonnées plus attractives pour les chercheurs. Les producteurs de données de tous pays sont confrontés à une demande croissante d'accès aux microdonnées servant de base aux statistiques publiques. L'accès aux microdonnées permet non seulement d'effectuer des recherches inédites et plus diversifiées, mais également de développer des méthodes innovantes d'exploitation, de traitement et d'affichage des informations, sans oublier la création de nouveaux ensembles de données combinant plusieurs sources.

Décider de la meilleure façon de diffuser les microdonnées constitue cependant un véritable défi pour leurs producteurs. Ce défi est à la fois d'ordre technique et organisationnel, puisqu'il s'agit de mettre en place des procédures adaptées de documentation, de catalogage et de partage des microdonnées. Des normes internationales et des principes de bonnes pratiques en la matière ont été élaborés par la communauté des archives de données pour résoudre ces questions. Le défi est néanmoins aussi de nature juridique et éthique. Ainsi, si les producteurs de données sont parfaitement conscients du pouvoir et de l'importance du partage des microdonnées, ils doivent cependant jongler entre la demande et la nécessité d'assurer la confidentialité des informations fournies par les personnes interrogées. Cette exigence, imposée par les législations nationales en matière de statistiques et de confidentialité des données, fait souvent l'objet d'un engagement fourni aux personnes interrogées au moment de la collecte des informations. Les instituts de statistique et autres producteurs de données doivent faire en sorte de préserver la confiance des répondants, sous peine de voir diminuer la volonté de coopérer à leurs enquêtes et la qualité des statistiques. La diffusion de microdonnées implique donc l'élaboration de politiques et de procédures définissant formellement les conditions d'accès à celles-ci.

Le cadre prévu à cet effet varie d'un pays à l'autre. Cela dit, « indépendamment des différences qui peuvent exister entre les pratiques et politiques en matière de partage de données, ainsi que des restrictions légitimes auxquelles peut être soumis l'accès aux données, un partage plus systématique serait bénéfique pour pratiquement tous les types de recherche ». [17]

Le présent guide est destiné à aider les producteurs et les conservateurs de microdonnées à élaborer leurs propres politiques et procédures en matière de diffusion des fichiers de microdonnées. Il est primordial que ces politiques et ces procédures soient définies de façon formelle et transparente. Une diffusion adéquate des microdonnées implique non seulement la mise à disposition des données et de la documentation correspondante, mais également des conditions régissant l'exploitation de ces données. Ces informations doivent être publiques et facilement accessibles, de préférence via l'internet.

Si la majeure partie de ce document présente un caractère générique, il est néanmoins destiné en premier lieu aux producteurs de données officiels – ONS et assimilés – des pays en développement. Les données dont il est question correspondent généralement aux microdonnées tirées d'enquêtes par sondage, de recensements et de systèmes de collecte de données administratifs.

Le présent guide a été élaboré sous l'égide du Réseau International pour les Enquêtes auprès des Ménages (IHSN). Il se fonde essentiellement sur les travaux de la Commission économique des Nations Unies pour l'Europe (CCE-ONU), et notamment du comité de réflexion de la Conférence des statisticiens européens chargé de plancher sur la gestion de la confidentialité et l'accès aux microdonnées, ainsi que de l'Office statistique de l'Union européenne (Eurostat) [5] [24] [25] [26]. Il s'appuie en outre sur l'expérience des instituts de statistique situés dans des régions du monde où la mise à disposition de fichiers de microdonnées est une pratique de longue date (plus de 40 ans parfois), mais aussi sur celle de divers centres de données universitaires.

Les informations contenues dans le présent document répondent aux douze grandes questions qui se posent dans le cadre de la formulation d'une politique de diffusion de fichiers de microdonnées :

1. Qu'est-ce que les microdonnées ?
2. Qu'est-ce que les métadonnées ?
3. Quels sont les arguments en faveur de la diffusion des microdonnées ?
4. Quels sont les coûts et les risques liés à la diffusion de microdonnées et comment peuvent-ils être maîtrisés ?
5. A qui les microdonnées sont-elles destinées ?
6. Quelles sont les conditions régissant la diffusion des microdonnées ?
7. Qu'entend-on par « anonymisation » des microdonnées ?
8. L'accès doit-il être payant ou gratuit ?
9. A quel moment du cycle de diffusion les microdonnées doivent-elles être rendues publiques ?
10. Quelles sont les exigences à remplir en termes d'infrastructure technique ?
11. Quelles sont les exigences institutionnelles relatives à la diffusion de microdonnées ?
12. Comment promouvoir l'utilisation de microdonnées ?

Le présent guide traite principalement des aspects politiques liés à la diffusion de microdonnées. Une diffusion appropriée et sûre des microdonnées implique toutefois également des solutions techniques pour la documentation, l'anonymisation, le catalogage et la conservation des données et des métadonnées. Ces questions ne sont que brièvement abordées ici, mais traitées plus en détails dans d'autres documents publiés par le réseau IHSN ou par d'autres organismes.

1. Qu'est-ce que les microdonnées ?

1.1 Notion de microdonnées

Les enquêtes / recensements ou la collecte de données administratives permettent aux instituts de statistique ou à d'autres producteurs de données de recueillir des informations sur des unités d'observation : ménages, individus, entreprises, exploitations agricoles, écoles, établissements de santé, etc. Les *microdonnées* dont il est question ici réfèrent aux fichiers de données électroniques qui contiennent les informations sur chacune de ces unités d'observation. A cet égard, les microdonnées s'opposent donc aux *macrodonnées* (ou *agrégats de données*), qui constituent une synthèse des informations recueillies sous forme de moyennes, pourcentages, fréquences ou autres statistiques de synthèse.

Les microdonnées sont généralement structurées en fichiers de données, dont chaque ligne (ou *enregistrement*) renferme des informations sur une unité d'observation. Ces informations sont enregistrées sous forme de *variables* de différents types (numérique ou alphanumérique, discrète ou continue, etc.). Elles peuvent être obtenues directement de la personne interrogée, via un questionnaire, ou par observation, mesure (localisation par GPS p. ex.), imputation ou calcul.

Les informations contenues dans les fichiers de microdonnées statistiques sont codées. Le sexe de la personne interrogée peut ainsi être enregistré sous la forme d'une variable de type *H01a*, à laquelle peut correspondre la valeur 1 ou 2 (1 pour masculin, 2 pour féminin). Par conséquent, les microdonnées doivent être accompagnées d'un *dictionnaire de données* contenant la liste des variables utilisées, une description de leur contenu et la signification de chaque code. Ces *métadonnées* constituent la documentation minimale requise. Cependant, comme nous le verrons au chapitre 2, le nombre de métadonnées nécessaire est en réalité beaucoup plus important.

Un ensemble de données d'enquête ou de recensement comprend généralement plusieurs fichiers, souvent issus de plusieurs niveaux d'observation d'une seule et même opération de collecte de données. Dans la majorité des cas, les recensements et autres enquêtes réalisées auprès des ménages permettent de recueillir des données à au moins deux niveaux : le ménage (avec des variables décrivant les caractéristiques du logement p. ex.) et l'individu (avec des informations sur l'âge, l'état civil, le niveau d'instruction et activité p. ex.). Un ensemble de données peut être constitué d'un ou de plusieurs fichiers pour chacun de ces niveaux. Ceux-ci contiennent des *variables clés* permettant aux utilisateurs de faire le lien entre les informations contenues dans des fichiers différents. Les ensembles de données ainsi organisés sont dits *hiérarchisés*.

1.2 Format de stockage et de diffusion des fichiers de microdonnées

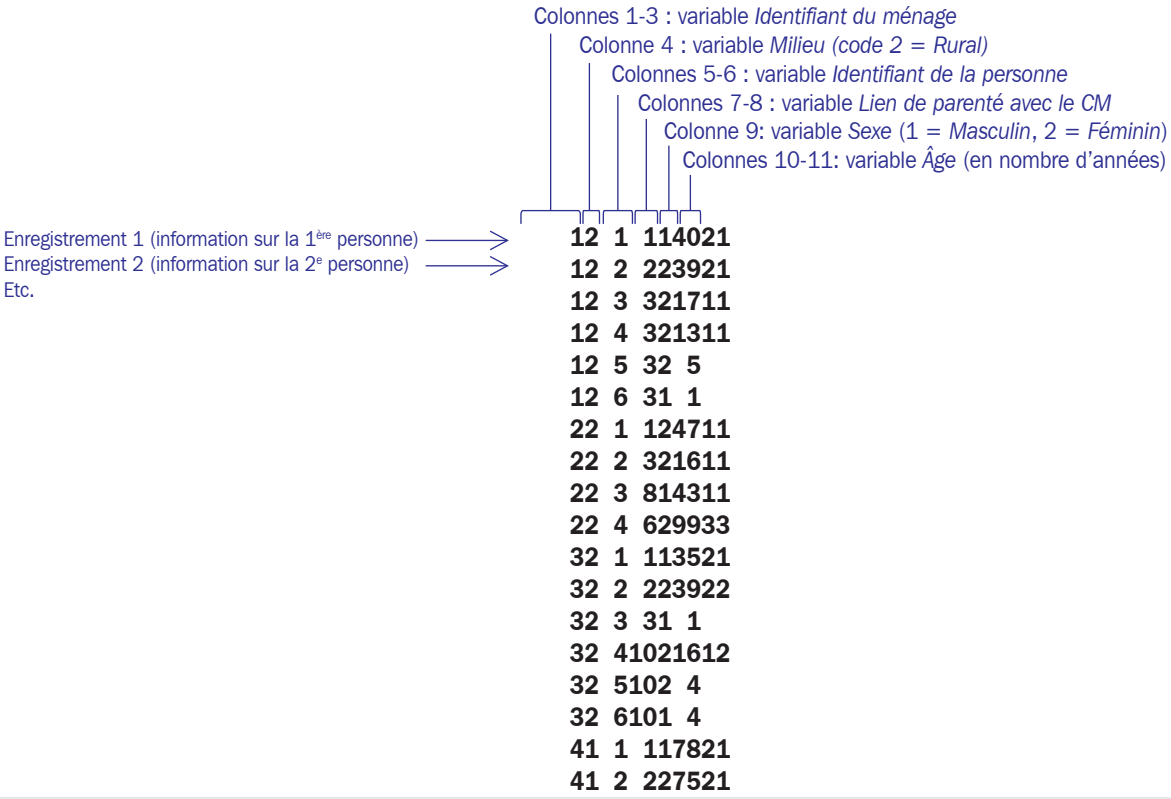
Les fichiers de microdonnées peuvent être stockés sous différents formats. Le format ASCII non propriétaire et les formats propriétaires créés par les logiciels de statistique spécialisés, comme SAS, SPSS et Stata, comptent parmi les plus couramment utilisés. Les microdonnées peuvent aussi être sauvegardées au format SQL ou dans d'autres formats de base de données. Ces formats sont toutefois moins courants et moins adaptés aux données d'enquête et de recensement, dans la mesure où les logiciels de base de données ne sont pas spécialement destinés à la création de tableaux et d'analyses statistiques.

Le format de fichier ASCII n'est pas propre à un logiciel ou à une plate-forme en particulier. Les fichiers ASCII contiennent des données lisibles avec la plupart des logiciels. N'étant pas liés à un logiciel menacé d'obsolescence, ils constituent la solution optimale pour garantir une conservation des données à long terme. Ils sont néanmoins indéchiffrables ou inutilisables sans un dictionnaire de données (fichier ou document séparé). La figure 1 présente un extrait de fichier de données statistiques type au format ASCII fixe[1].

Pour créer un tableau statistique ou analyser des données ASCII, il faut d'abord importer ces dernières dans un autre logiciel. Tous les logiciels de statistique et de base de données proposent des outils et des commandes à cet effet. La page suivante contient un exemple de script Stata permettant d'importer les données ASCII de la figure 1 et d'associer des étiquettes

1 Les fichiers ASCII peuvent être *fixes* ou *délimités*. Dans les fichiers ASCII fixes, les données associées à une variable occupent toujours la même position (colonne). Dans les fichiers ASCII délimités, les informations relatives à chaque variable sont séparées par un caractère spécial (point-virgule, tabulation, virgule ou autre caractère défini par l'utilisateur). A titre d'exemple, dans un fichier ASCII *délimité par des virgules*, chaque variable sera séparée par une virgule.

Figure 1 **Extrait d'un fichier de données au format ASCII fixe**



Exemple de configuration de l'importation de données ASCII dans Stata *(set-up)*

```
* · Read the ASCII data found in file test.dat and import the values in new variables;
* · Read the ASCII data found in file test.dat and import the values in new variables;
· · infix hhid 1-3 area 4 pid 5-6 relat 7-8 sex 9 age 10-11 using test.dat;

* · Add a label to describe each new variables;
· · label variable hhid "Identifiant du ménage";
· · label variable area "Milieu de résidence";
· · label variable pid "Identifiant de la personne";
· · label variable relat "Lien de parenté avec le CM";
· · label variable sex "Sexe";
· · label variable age "Âge au dernier anniversaire";

* · Add label to each code used by the variables;
· · label define relatcod 1 "Chef de ménage" 2 "Epoux/se" 3 "Fils/Fille" 4 "Belle-fille/Beau-fils "
· · 5 "Petit fils/Petite fille" 6 "Parent" 7 "Beau-parent" 8 "Frère/sœur" 9 "Autre parent"
· · 10 "Aucun lien", add;
· · label values relat relatcod;
· · label define areacod 1 "Urbain" 2 "Rural", add;
· · label values area areacod;
· · label define sexcod 1 "Masculin" 2 "Féminin", add;
· · label values sex sexcod;

* · Save the file as a Stata file;
· · save "test.dta", replace;
```

(*labels*) aux variables et aux codes pour rendre le contenu plus convivial pour l'utilisateur. Il est clair que pour rédiger ce type de script, l'utilisateur doit disposer d'un dictionnaire de données décrivant le contenu et la structure du fichier de données ASCII.

Une fois importé dans Stata après exécution du script, le fichier ASCII de la figure 1 apparaîtra tel qu'il est représenté à la figure 2. Les formats de fichier propriétaires de SAS, SPSS, Stata ou de logiciels équivalents comprennent à la fois les données, les variables et les étiquettes correspondantes.

Figure 2 **Extrait d'un fichier de données Stata**

	Hhid	area	Pid	Relat	Sex	Age
1	1	Rural	1	Chef de ménage	Masculin	40
2	1	Rural	2	Epoux/se	Féminin	39
3	1	Rural	3	Fils/Fille	Féminin	17
4	1	Rural	4	Fils/Fille	Féminin	13
5	1	Rural	5	Fils/Fille	Féminin	5
6	1	Rural	6	Fils/Fille	Masculin	1
7	2	Rural	1	Chef de ménage	Féminin	47
8	2	Rural	2	Fils/Fille	Féminin	16
9	2	Rural	3	Frère/sœur	Masculin	43
10	2	Rural	4	Parent	Féminin	99
11	3	Rural	1	Chef de ménage	Masculin	35
12	3	Rural	2	Epoux/se	Féminin	39
13	3	Rural	3	Fils/Fille	Masculin	1
14	3	Rural	4	Aucun lien	Féminin	16
15	3	Rural	5	Aucun lien	Féminin	4
16	3	Rural	6	Aucun lien	Masculin	4
17	4	Urbain	1	Chef de ménage	Masculin	78
18	4	Urbain	2	Epoux/se	Féminin	75

Les ensembles de données d'enquête peuvent contenir des centaines de variables, voire des milliers. La rédaction de scripts pour importer et documenter de tels fichiers de données à partir du format ASCII prend du temps et peut être source d'erreurs. Pour minimiser le risque d'erreur et pour un meilleur confort d'utilisation, il convient que les producteurs de données livrent leurs fichiers, soit au format ASCII, avec des modèles de script SPSS, SAS et Stata, soit aux formats statistiques propriétaires les plus courants. Il existe des logiciels spécialisés tels que StatTransfer (Stata Corporation) pour convertir les fichiers de données automatiquement d'un format de paquetage à un autre.

1.3 Quelle version des fichiers de données fautil diffuser ?

Les producteurs de données créent souvent plusieurs versions d'un même fichier de microdonnées. Ces versions se distinguent du point de vue qualitatif, de leur contenu et du nombre d'enregistrements. Elles vont de fichiers de microdonnées brutes – contenant toutes les réponses fournies par chaque répondant, obtenues directement après saisie des données – à des fichiers à grande diffusion, nettoyés et modifiés.

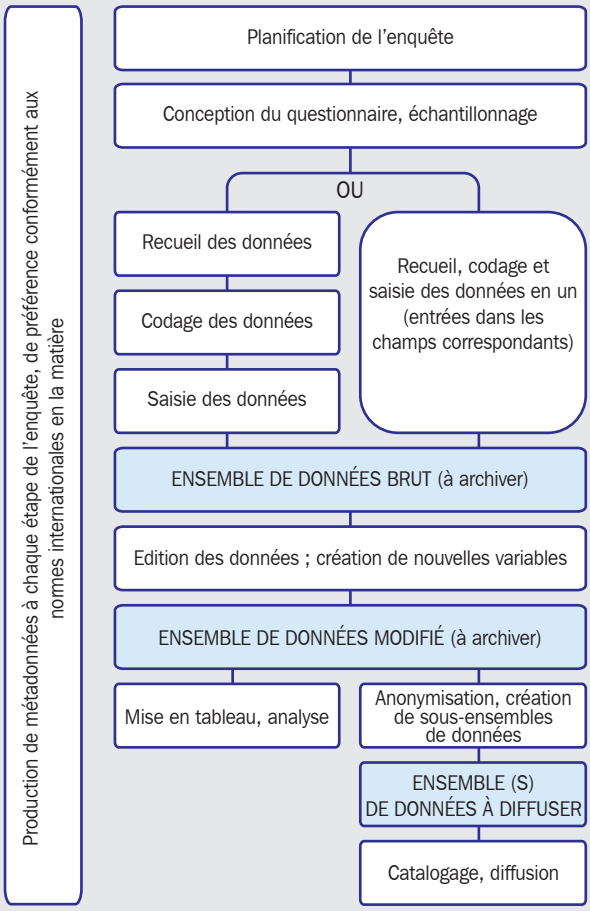
La figure 3 représente le cycle de vie type d'une enquête ou d'un recensement.

1.4 Quels sont les éléments sensibles du contenu des microdonnées ?

Les données issues de recensements et d'enquêtes par sondage sont utilisées exclusivement à des fins statistiques ou pour la recherche. Pour des raisons pratiques, le nom et l'adresse des répondants sont souvent mentionnés sur les questionnaires, mais figurent rarement dans les fichiers de données correspondants. En règle générale, ces derniers ne contiennent donc pas de variables constituant des *identifiants directs*. En revanche, les fichiers de données administratifs comprennent fréquemment des noms, adresses, numéros de téléphones, numéros de sécurité sociale, etc.

La plupart des ensembles de données comportent néanmoins des *identifiants indirects*. Ainsi, la situation géographique, la composition du ménage (âge et sexe) et l'activité professionnelle peuvent servir à identifier les répondants.

Figure 3 **Cycle de vie d'une enquête**



Ces informations sont considérées comme sensibles, car elles peuvent permettre d'identifier les unités d'observation. D'autres variables sont sensibles de par la nature des informations qu'elles renferment, comme l'état de santé d'un individu, ses habitudes sexuelles, ses revenus, etc. Les enquêtes réalisées auprès des entreprises, par exemple, sont sensibles par nature, dans la mesure où les informations recueillies peuvent être exploitées par la concurrence.

1.5 Quels sont les principaux types de fichiers de microdonnées à diffuser ?

Les fichiers de microdonnées destinés à être diffusés sont presque toujours différents de ceux réservés à l'usage du personnel des organismes producteurs. La préparation des fichiers de microdonnées brutes en vue de leur diffusion comprend des procédures d'ajustement du contenu et/ou du nombre d'enregistrements. La modification du *contenu des enregistrements* des fichiers de microdonnées à diffuser consiste à supprimer les identifiants directs et indirects pour protéger l'anonymat des répondants. Cela ne se traduit pas forcément par une suppression de variables. Il suffit parfois de regrouper certaines variables dans des catégories moins détaillées. Cela peut passer par une réduction du *nombre d'enregistrements* contenu dans le fichier de microdonnées diffusé (cas des données de recensement de population p. ex.). Les procédures visant à protéger l'identité des répondants sont communément désignées par les notions d'*anonymisation* ou de *contrôle de la divulgation statistique* (CDS).

Les fichiers de microdonnées élaborés en vue d'établir des statistiques officielles sont diffusables, à condition que l'anonymat des répondants puisse être correctement garanti. Trois types de fichiers sont à envisager dans le cadre de la définition d'une politique de diffusion : les fichiers à grande diffusion, les fichiers sous licence et les fichiers accessibles uniquement dans des centres sécurisés. Ils se distinguent par leur niveau d'accessibilité et leur degré d'anonymisation.

« Aucun individu (…) ne peut prétendre obtenir ou accéder à des données identifiables (…) en vertu de son seul statut professionnel. L'accès à des données identifiables n'est pas déterminé uniquement par le statut professionnel, par l'affiliation à un organisme ou par un engagement financier. Le besoin de données identifiables, l'usage qui en sera fait, ainsi que le rôle et la responsabilité du demandeur en matière de collecte de données sont des critères plus importants. Dans la mesure où l'accès à des données identifiables n'est jamais sans risque, il fera l'objet d'une évaluation et d'un suivi approfondis après octroi de l'autorisation correspondante. » [14] (Traduction de l'anglais)

Fichiers de microdonnées à grande diffusion (FMGD)

Les fichiers de microdonnées à grande diffusion (FMGD) peuvent être consultés par quiconque accepte de se conformer à quelques conditions de base faciles à remplir. Il s'agit des conditions d'utilisation (vente interdite p. ex.) et non pas d'accès aux données. Certains FMGD sont diffusés sans condition ; ils sont alors généralement disponibles en ligne. Ces données sont facilement accessibles, car le risque d'identification des répondants est considéré comme minime. Cela implique la suppression de l'intégralité du contenu pouvant permettre d'identifier directement les répondants — noms, adresses et numéros de téléphones notamment. Il faut

aussi éliminer les identifiants indirects. Ceux-ci varient selon la méthodologie de l'enquête, mais certains éléments d'information sont couramment retirés, comme les renseignements géographiques plus précis que la strate d'échantillonnage. Certains enregistrements sont parfois supprimés des FMGD, tels que les variables extrêmes ou caractérisées par une distribution très asymétrique. Il existe toutefois d'autres méthodes CDS permettant de minimiser le risque de divulgation tout en maximisant le contenu informatif des données (regroupement des valeurs extrêmes supérieures et inférieures, suppression des variables provenant de certains répondants ou encore techniques de perturbation des données[2]). Les FMGD sont généralement créés à partir de fichiers de données de recensement (sous-ensembles d'enregistrements plutôt que fichiers complets) et d'enquêtes réalisées auprès des ménages. S'il est techniquement possible de créer des FMGD à partir d'enquêtes auprès des entreprises, cela représente toutefois des défis particuliers qui seront décrits séparément.

Les FMGD doivent présenter le caractère le plus informatif possible. Selon le centre national des statistiques de santé américain (NCHS), « l'objectif est de mettre à disposition les microdonnées le plus largement et sous la forme la plus détaillée possible suivant les seules restrictions imposées par les ressources disponibles, par les exigences qualitatives, par les contraintes techniques et par la nécessité de protéger la confidentialité des données. » [14] (Traduction de l'anglais)

Fichiers sous licence

Les *fichiers sous licence* – ou *fichiers de recherche* – se distinguent des FMGD par le fait que leur diffusion est limitée aux utilisateurs bénéficiant d'une autorisation d'accès octroyée après le dépôt d'une demande dûment justifiée et la signature d'un accord régissant l'utilisation des données. En règle générale, les fichiers sous licence sont également anonymisés afin de réduire au minimum le risque d'identification des individus en cas d'utilisation isolée. Ils peuvent toutefois contenir des données potentiellement identifiables en association avec d'autres fichiers[3].

Les identifiants directs comme les noms des répondants doivent être supprimés des ensembles de données sous licence. Ces fichiers contiennent néanmoins parfois des variables indirectes pouvant servir à identifier les répondants lorsqu'elles sont recoupées avec d'autres ensembles de données (listes de votants, registres fonciers, dossiers scolaires p. ex).

La diffusion des fichiers sous licence s'appuiera de préférence sur l'élaboration et la signature d'un contrat entre le producteur de données et des utilisateurs *dignes de confiance* – c'est-à-dire dont le besoin d'accès aux données est légitime. Il convient que cet accord régisse l'accès et l'utilisation des fichiers de microdonnées. Les contrats de licence sont parfois signés uniquement avec les utilisateurs membres d'un organisme commanditaire approprié (centre de recherche, université ou partenaire de développement).

Il est en outre recommandé aux producteurs des données qu'ils demandent aux requérants, préalablement à la signature du contrat d'utilisation et d'accès, de remplir un formulaire de demande légitimant leur besoin de consulter le fichier sous licence (plutôt que le FMGD correspondant le cas échéant) à l'une des fins statistiques ou de recherche mentionnées. Des modèles de contrat et de formulaire de demande de fichiers sous licence sont fournis au chapitre 6. Ils permettent d'appréhender les conditions d'octroi de l'autorisation d'accès aux fichiers de microdonnées.

Fichiers consultables dans des centres de données sécurisés

Certains fichiers peuvent être proposés à la consultation dans des centres sécurisés (ou « enclaves de données »), sous des conditions très strictes. Il s'agit d'un service équipé d'ordinateurs qui ne sont connectés ni à Internet, ni à un réseau externe, et à partir desquels aucune information ne peut être téléchargée

2 Une présentation et une description des techniques CDS visant à minimiser le risque de divulgation sont fournies au chapitre 7.

3 Pour des informations plus détaillées sur la définition et les différences entre les fichiers à grande diffusion et les fichiers sous licence, se reporter au travail effectué par la Commission économique des Nations Unies pour l'Europe (CEE-ONU) dans le cadre de la Conférence des statisticiens européens [24].

via des ports USB, sur CD/DVD ou sur d'autres types de lecteurs. Ces « enclaves » renferment des données particulièrement sensibles ou permettant d'identifier directement et aisément les répondants.

Il peut s'agir, par exemple, d'ensembles de données complets issus de recensements de population, d'enquêtes réalisées auprès d'entreprises ou encore de dossiers de santé contenant des informations hautement confidentielles. Les utilisateurs qui souhaitent consulter de tels ensembles de données n'y ont pas nécessairement accès en intégralité – mais uniquement de manière restreinte (accès au sous-ensemble de données nécessaire). Ils sont invités à remplir un formulaire de demande justifiant du fait que l'accès aux données est destiné à des fins statistiques ou de recherche (voir exemple au chapitre 6). Les résultats obtenus doivent faire l'objet d'un examen minutieux dans le cadre d'une procédure complète de contrôle de divulgation.

L'exploitation d'un centre de données sécurisé a un coût non négligeable. Elle implique en effet l'aménagement de locaux et l'achat d'équipements informatiques spéciaux. Elle exige aussi du personnel disposant des compétences et du temps nécessaires pour procéder aux contrôles visant à éliminer tout risque de divulgation. Les personnes concernées doivent être familiarisées avec les méthodes d'analyse de données, savoir traiter les demandes et gérer des serveurs de fichiers.

Compte tenu des coûts d'exploitation substantiels qu'elles représentent et des compétences techniques élevées requises, certains instituts de statistique ou d'autres producteurs de données officiels ont choisi de collaborer avec des établissements universitaires ou des centres de recherche pour établir et gérer les « enclaves de données ». En voici quelques exemples avec mention de l'adresse du site Web de présentation correspondant : le MCRDC (centre de recherche sur le recensement), un projet commun du *Census Bureau* américain et de l'Université du Michigan (www.isr.umich.edu/src/mcrdc/) ; le *National Opinion Research Center* (NORC) de l'Université de Chicago (www.norc.org/DataEnclave) ; le programme CDR de Statistique Canada (www.statcan.gc.ca/rdc-cdr/index-fra.htm) ; et le *Research Data Center* du centre national des statistiques de santé américain (http://www.cdc.gov/nchs).

Encadré 1 **Une enquête, plusieurs produits**

Les producteurs de données peuvent choisir de générer plusieurs produits à partir d'un même ensemble de données de recensement ou d'enquête. Des FMGD peuvent ainsi être créés à partir de petits échantillons ou de sous-ensembles de variables. Ces fichiers peuvent être largement diffusés sans risque de divulgation de l'identité des répondants. Une version plus étendue portant sur un échantillon plus vaste et dont l'accessibilité est soumise à l'obtention d'une licence est envisageable. Enfin, le fichier complet (avec ou sans identifiants) peut être consulté dans un centre sécurisé.

Le *Census Bureau* américain a ainsi produit deux fichiers à grande diffusion distincts à partir de l'ensemble des données issu du recensement de 2000, avec des taux d'échantillonnage de 1 % et de 5 %.

> *Compte tenu de l'évolution rapide des technologies informatiques et de l'accessibilité croissante des données de recensement pour la communauté d'utilisateurs, le Census Bureau s'est vu contraint d'adopter des mesures plus contraignantes pour protéger la confidentialité des microdonnées à grande diffusion, en recourant à des techniques de limitation de la divulgation des informations. Le Census Bureau reconnaît par ailleurs que les utilisateurs ont besoin de données plus détaillées et plus resserrées sur le plan géographique. Deux ensembles de fichiers seront donc fournis: l'un contenant un plus grand nombre de caractéristiques détaillées (fichier national, taux d'échantillonnage de 1 %) et l'un comprenant de données géographiques plus précises, mais des caractéristiques moins détaillées (fichiers des Etats ; taux d'échantillonnage de 5 %). (Traduction de l'anglais)*
>
> Source: http://www.census.gov/population/www/cen2000/pums/index.html, site consulté le 8 avril 2010.

L'intégralité des microdonnées peut être consultée dans différents centres de données sécurisés aux EU, notamment au MCRDC de l'Université du Michigan.

> *Le centre de recherche sur le recensement du Michigan (Michigan Census Research Data Center, MCRDC) permet aux chercheurs qualifiés travaillant sur des projets approuvés par le Census Bureau des Etats-Unis d'exploiter les données non publiées recueillies dans le cadre des programmes économiques et démographiques du Census Bureau, ainsi que par le centre national des statistiques de santé (NCHS). Toutes les recherches du MCRDC sont menées au sein de son laboratoire sécurisé, situé à Ann Arbor au sein de l'Institute for Social Research de l'Université du Michigan. (Traduction de l'anglais)*
>
> Source: http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/restricted/index.jsp (site consulté le 7 mai 2010).

1.6 Existe-t-il des alternatives à la communication des microdonnées ?

Dans les cas susmentionnés, l'utilisateur se voit offrir un accès direct aux fichiers de microdonnées. Il existe cependant d'autres formes d'accès aux microdonnées, notamment la soumission de tâches à distance et le téléaccès. Les exigences spécifiques aux politiques régissant ce type d'accès ne sont pas décrites ici, mais simplement mentionnées, avec des renvois aux sources d'informations complémentaires. Il faut souligner que ces systèmes sont généralement coûteux et techniquement complexes.

Soumission de tâches

L'un des moyens dont disposent les utilisateurs pour analyser des données confidentielles est de créer une procédure leur permettant de soumettre des programmes de traitement et d'analyse de données confidentielles à distance à leur dépositaire. L'utilisateur obtient alors un ensemble de données synthétisé reproduisant la structure et le contenu des originaux. Dans ce cadre, les chercheurs peuvent élaborer des programmes à l'aide d'outils tels que SAS, SPSS ou Stata. Ces programmes sont ensuite transmis au personnel du dépositaire des données, qui exécute l'application avec l'ensemble de données original. Les résultats obtenus sont vérifiés et ajustés avant d'être retransmis à l'utilisateur.

Atitred'exemple,citonsleprojetLIS(*Luxembourg Income Study*), qui offre la possibilité d'accéder à des bases de microdonnées via un système de soumission de tâches automatisé appelé LISSY (Encadré 2).

Certes la confidentialité des données est protégée, mais le coût de la prise en charge des services de soumission de tâches peut être élevé. De plus, si les ressources affectées à ces services sont insuffisantes, la procédure sera lente pour les utilisateurs.

Téléaccès

Ici, les utilisateurs ont accès à des logiciels de classification et d'analyse de données, mais ne peuvent télécharger aucun ensemble de données,

Encadré 2 **Luxemburg Income Study - LISSY**

LISSY est un système d'exécution de tâches à distance entièrement automatisé fonctionnant 24 heures sur 24 et 7 jours sur 7. Il permet aux chercheurs de soumettre des programmes de traitement statistique par lots (conçus avec SAS, SPSS ou Stata) sans se déplacer. LISSY exécute automatiquement les tâches demandées, puis transmet les résultats sous une forme synthétisée, en moyenne en quelques minutes.

Les bases de microdonnées ne peuvent pas être téléchargées et aucun accès direct aux données n'est possible. Seuls les résultats des requêtes statistiques sont envoyés aux utilisateurs.

Inscription obligatoire

Les chiffres clés du système LIS (*key figures*) sont mis à disposition du grand public, mais l'accès aux bases de microdonnées est réservé aux utilisateurs inscrits. L'autorisation d'accès est accordée pour une période d'un an seulement, renouvelable chaque année.

Deux modes de soumission de tâches

LISSY offre un accès à distance sécurisé aux microdonnées via deux modes de soumission :

- logiciel JSI *(Job Submission Interface)*;
- logiciel de messagerie tel que Outlook, Thunderbird, etc.

Les deux systèmes aboutissent aux mêmes résultats, mais le site du LIS recommande fortement aux utilisateurs de passer par JSI pour accéder à LISSY. L'interface est plus conviviale et propose des fonctionnalités supplémentaires telles que l'accès à l'historique des tâches de l'utilisateur.

Instruction de soumission

Quel que soit le mode de soumission des tâches, il existe quelques spécificités par rapport à la syntaxe de programmation standard, qui doivent être maîtrisées pour que les requêtes des utilisateurs puissent être correctement traitées par LISSY.
Les logiciels de statistique actuellement disponibles dans le système LISSY sont SAS 9.2, Stata 11.0 et SPSS 11.5.

Assistance technique

Toutes les questions relatives à l'utilisation et au contenu des bases de données LIS doivent être adressées au service d'assistance technique (*LIS User Support*) plutôt qu'à des membres du personnel du LIS. L'objectif est de pouvoir recenser de façon coordonnée l'ensemble des questions posées.

Source: http://www.lisproject.org/data access/data access.html (site consulté le 9 avril 2010).

ni créer de tableaux qui dévoileraient des informations individuelles ou un petit nombre d'enregistrements.

Plusieurs logiciels sont proposés sur le marché (Nesstar, Beyond 20/20, SuperCross, Redatam, PcAxis, etc.). Certains centres de données à la pointe du progrès développent leurs propres plate-formes. Le centre d'archivage UKDA gère ainsi un service de données sécurisé (*Secure Data Service*, SDS) « visant à promouvoir l'excellence de la recherche en offrant aux chercheurs dignes de confiance un accès à distance sécurisé à des données jugées trop sensibles, détaillées, confidentielles ou identifiables pour être mise à disposition via un contrat de licence ou des dispositifs de diffusion standard. » [12] (Traduction de l'anglais)

Ce système convient pour la création de tableaux statistiques (cas des recensements de population et des logements notamment), mais pas pour une analyse approfondie.

Recrutement d'un chercheur comme agent temporaire

Certains producteurs de données veillent à ce que les chercheurs aient accès aux microdonnées en les recrutant comme agents temporaires. Ils sont alors soumis aux mêmes règles relatives à la confidentialité que le personnel permanent de l'organisme. Il convient de n'avoir recours à cette formule que si l'activité du chercheur représente un véritable apport au travail du producteur de données concerné, sous peine d'y voir un simple simulacre. [24]

2. Qu'est-ce que les métadonnées ?

Les métadonnées sont généralement définies comme des « données décrivant d'autres données ». Le chapitre précédent mentionne l'importance que revêt la mise à disposition d'un dictionnaire de données approprié décrivant le contenu de toutes les variables comprises dans un ensemble de données. Mais les métadonnées de qualité fournissent des informations allant bien audelà de celles d'un simple dictionnaire de données.

Les *métadonnées* sont destinées à aider les chercheurs à **comprendre** ce que mesurent les données et la façon dont elles ont été obtenues. Sans une bonne description du concept de l'enquête et des méthodes employées pour recueillir et traiter les données, l'utilisateur risque fortement de mal interpréter ces dernières, voire de les utiliser à mauvais escient.

Une documentation appropriée réduit également le volume de travail des statisticiens en termes d'assistance technique aux utilisateurs externes de leurs microdonnées.

Les métadonnées sont par ailleurs destinées à aider les utilisateurs à évaluer la qualité des données. Il est déterminant que tout chercheur souhaitant pouvoir juger de l'intérêt de certaines données pour son travail connaisse les normes en vigueur en matière de collecte de données – et soit à même d'identifier les écarts par rapport à ces normes.

Enfin, les métadonnées sont nécessaires au développement d'outils de **recherche de données** tels que les catalogues d'enquêtes, qui aident les chercheurs à trouver des ensembles de données pertinents.

Il faut souligner que les données doivent faire l'objet d'une documentation même quand elles ne sont pas destinées à être diffusées. Produire des métadonnées de qualité favorise la constitution d'une « mémoire institutionnelle » en matière de collecte de données, peut contribuer à la formation du nouveau personnel et améliorer la cohérence des données dans le temps.

2.1 Qu'est-ce que des métadonnées de qualité ?

La description suivante de métadonnées de qualité est tirée du guide *Good Practices in Data Documentation* (bonnes pratiques en matière de documentation de données) publié par l'UKDA. [20] Le site Internet du Réseau International pour les Enquêtes auprès des Ménages (www.ihsn.org) et son guide pratique à l'intention des archives de données (*Quick reference Guide for Data Archivists*) renferment aussi des informations très intéressantes sur le sujet. [4]

« Une étape déterminante de la création d'un ensemble de données de qualité et exploitable à long terme consiste à faire en sorte que les données soient faciles à comprendre et à analyser. Cela implique une description et une documentation des données qui soient conviviales, claires et détaillées tout en étant exhaustives. » (http://www.data archive.ac.uk, traduction de l'anglais)

La documentation idéale d'un ensemble de données comporte essentiellement trois types de documents :

1. Documents à caractère explicatif

C'est le minimum requis pour assurer la viabilité et la fonctionnalité des données à long terme – en l'absence de ces informations, il est impossible de bien comprendre l'ensemble de données et son contenu.

Informations sur les méthodes de collecte de données

Cette partie est consacrée au processus de collecte des données – enquête, recueil de données administratives ou transcription d'un document source. Elle décrit normalement les instruments utilisés, les méthodes employées et la façon dont ces dernières ont été élaborées. Le cas échéant, des détails sont fournis sur la méthode et le cadre d'échantillonnage. Les informations sur un éventuel système de surveillance de la collecte des données, ainsi que de contrôle de la qualité, sont en outre extrêmement utiles.

Informations sur la structure de l'ensemble de données

Il s'agit essentiellement d'un document détaillé décrivant la structure de l'ensemble de données, y compris les relations entre certains fichiers ou éléments d'information (enregistrements) qu'il contient. Il indique généralement les variables clés nécessaires pour identifier de façon univoque les sujets dans les différents fichiers. Normalement, le nombre de cas et de variables que contient chaque fichier, ainsi que le nombre total de fichiers constituant l'ensemble

de données sont également stipulés. Dans le cas des modèles relationnels, des renseignements sur la structure et sur les liens entre les enregistrements devraient être fournis.

Informations techniques

Ces informations se rapportent à l'infrastructure technique et indiquent généralement :

- le système informatique utilisé pour créer les fichiers ;
- les logiciels utilisés pour créer les fichiers ;
- le support de stockage des données ; et
- la liste complète des fichiers contenus dans l'ensemble de données.

Variables et valeurs, systèmes de codage et de classification

Il est souhaitable que la documentation comprenne la liste complète des variables (ou des champs) qui apparaissent dans l'ensemble de données, avec une description exhaustive et des renseignements détaillés sur les systèmes de codage et de classification retenus pour les informations correspondantes. Il est particulièrement important que les champs non renseignés et manquants soient mentionnés et pris en compte. Il est également utile d'indiquer quelles sont les variables auxquelles une nomenclature standard s'applique, en précisant la version du système de classification employé – de préférence avec les références bibliographiques correspondantes.

Informations sur les variables dérivées

De nombreux producteurs de données créent de nouvelles variables à partir de données originales. La méthode peut être très simple et se traduire, par exemple, par un regroupement de données par âge (en nombre d'années révolues), suivant les classes d'âge pertinentes pour l'enquête en question. D'autres méthodes plus complexes font appel à des algorithmes beaucoup plus élaborés. Il convient d'expliquer la logique qui préside au regroupement ou à une dérivation de variables. Un simple regroupement de données par âge peut être explicité dans le dictionnaire de données. Les dérivations plus complexes doivent en revanche faire l'objet d'une documentation séparée, contenant de préférence des organigrammes ou des déclarations booléennes précises. L'idée est que les informations fournies permettent de faire aisément le lien entre les variables principales et les variables résultantes. Il est par ailleurs recommandé que les algorithmes informatiques utilisés pour créer les variables dérivées soient conservés et consignés avec les informations relatives au logiciel employé.

Pondération et extrapolation

Une liste complète des variables pondérées et extrapolées doit être fournie, avec des renseignements sur le mode d'obtention de ces variables et en stipulant clairement les conditions d'applicabilité de celles-ci. Ceci est particulièrement important quand différents facteurs de pondération doivent être appliqués suivant l'objectif visé.

Source des données

Des informations détaillées sur la source des données sont normalement fournies. Dans le cas d'une source de données constituée des réponses à des questionnaires d'enquête, par exemple, chaque question devrait être minutieusement répertoriée dans la documentation. Idéalement, les références de chaque variable générée sont mentionnées. Il est en outre intéressant d'expliquer les conditions dans lesquelles les questions sont posées et, si possible, d'indiquer les cas de figure dans lesquels elles s'appliquent, avec de préférence une synthèse des statistiques de réponse.

Confidentialité et anonymisation

Il est important de préciser si les données contiennent des informations confidentielles sur des individus, des ménages, des organisations ou des institutions. Le cas échéant, il est recommandé de mentionner ces données, ainsi que tout contrat régissant les conditions d'utilisation de celles-ci (information des participants à l'enquête p. ex.). Les questions de confidentialité peuvent avoir un effet restrictif sur les analyses effectuées ou sur les résultats publiés, notamment si les données sont destinées à une utilisation secondaire. En cas d'anonymisation des données visant à protéger l'identité des personnes interrogées, la procédure retenue dans ce cadre et son impact sur les données devraient être précisés. Dans la mesure où de telles modifications sont susceptibles de limiter l'analyse qui s'ensuit, il est en effet utile d'en faire état.

2. Informations contextuelles

Il s'agit de fournir aux utilisateurs des informations sur le contexte de la collecte des données, ainsi que sur l'usage qui en a été fait. Ces renseignements enrichissent

et approfondissent la documentation. Ils permettent aux utilisateurs secondaires d'appréhender pleinement le contexte et les procédures de collecte de données. Qui plus est, ils constituent un témoignage historique fondamental pour les futurs chercheurs.

Description du projet d'origine

Il est souhaitable que la documentation comprenne des informations sur l'historique du projet ou sur le processus qui est à l'origine de l'ensemble de données, décrivant le cadre intellectuel et formel de référence. Exemples :

- Raisons ayant motivé la collecte de données ;
- Buts et objectifs du projet ;
- Sujet de l'étude ;
- Couverture géographique et cadre temporel ;
- Publications auxquelles l'étude a contribué ou politiques ayant été définies en réponse à celle-ci ; et
- Toute autre information jugée pertinente.

Provenance de l'ensemble de données

Il s'agit des informations relatives aux aspects tels que l'historique du processus de collecte des données, les changements intervenus et l'évolution des données proprement dites, ainsi que la méthodologie et les ajustements effectués. Les renseignements suivants peuvent également être fournis :

- Liste des données erronées ;
- Problèmes rencontrés au moment de la collecte, de la saisie, du contrôle et du nettoyage des données ;
- Conversion visant à assurer la compatibilité avec un autre logiciel ou système d'exploitation ;
- Références bibliographiques des rapports ou publications s'appuyant sur l'étude ; et
- Toute autre information jugée pertinente sur le cycle de vie de l'ensemble de données.

Ensembles de données chronologiques et chroniques, nouvelles éditions

Dans le cas de données chronologiques, de panel ou transversales chroniques, il est extrêmement utile de bénéficier d'informations supplémentaires, par exemple sur l'évolution de la formulation des questions, de l'intitulé des variables ou des procédures d'échantillonnage.

3. Systèmes de catalogage

L'objectif de ces systèmes est double. Premièrement, ils constituent les références bibliographiques de l'ensemble de données, ce qui permet de l'identifier et de le citer correctement dans des publications. Ils permettent également de l'enregistrer officiellement en vue de sa conservation à long terme. Deuxièmement, il s'agit de l'outil de base utilisé pour rechercher des ressources. L'ensemble de données peut ainsi être identifié parmi d'autres de manière univoque en précisant les informations qui aideront les utilisateurs secondaires à déterminer l'utilité de l'étude pour leur activité.

En l'absence d'intitulés, de résumés, de mots-clés et d'autres éléments de métadonnées importants, il sera difficile pour les chercheurs d'identifier les ensembles de données et les variables qui répondent à leurs besoins. Tous les systèmes de catalogage et de recherche de ressources, qu'ils soient physiques ou électroniques, reposent sur des métadonnées.

« Il est plus aisé d'obtenir une documentation de qualité quand sa constitution est planifiée dès le début du projet et pensée à toutes les étapes de la recherche (tout au long du cycle de vie des données). Une planification poussée peut permettre de réduire sensiblement le temps et le budget nécessaires à la préparation de la documentation. » (http://www.data-archive.ac.uk, traduction de l'anglais). Voir aussi [21].

2.2 Normes et bonnes pratiques en matière de métadonnées

« L'interopérabilité technologique et sémantique est essentielle pour faciliter et encourager l'accessibilité et l'utilisation des données de la recherche dans un contexte international et interdisciplinaire. Les dispositifs d'accès devraient tenir dûment compte des normes internationales pertinentes applicables en matière de documentation des données. » [17]

Soucieuse de favoriser l'échange de données entre les organisations et les systèmes logiciels et d'améliorer la qualité de la documentation statistique fournie aux utilisateurs, la communauté des archives a élaboré un ensemble de normes relatives aux métadonnées. Ces normes offrent un cadre structuré pour l'organisation et la diffusion d'informations sur le contenu et sur la structure des données statistiques.

ISO 11179 – Technologies de l'information- Registres de métadonnées (RM)

La norme internationale ISO/CEI 11179-1 a été mise au point par le comité technique JTC 1 de l'ISO, (technologies de l'information), sous-comité SC 32 (services de gestion des données).

« La norme ISO/CEI 11179 définit la standardisation et l'enregistrement des éléments de données dans le but de d'en faciliter la compréhension et l'échange. La standardisation et l'enregistrement des éléments de données tels qu'ils sont décrits par la norme ISO/CEI 11179 permettent de constituer un environnement de données partagées beaucoup plus rapidement et aisément qu'avec des méthodes de gestion des données classiques. » [9] (Traduction de l'anglais)

Certaines organisations ont fondé la conception de leurs bases de données de concepts et de définitions sur la norme ISO 11179. Il faut toutefois souligner que cette dernière ne fournit aucun outil de documentation et de diffusion pratique, contrairement aux normes ci-après, basées sur le langage XML.

Data Documentation Initiative (DDI)

Traditionnellement, les producteurs de données rédigeaient des dictionnaires de codes au format texte. Afin d'exploiter au mieux la technologie Internet, la majorité des normes sont désormais définies en langage XML. La spécification *Data Documentation Initiative* (DDI) est une norme relative à la documentation des microdonnées[4]. [13]

L'initiative DDI a permis de mettre au point des normes offrant un cadre structuré pour l'organisation du contenu, de la présentation, du transfert et de la conservation des métadonnées en sciences sociales et comportementales. Ce cadre permet de documenter les fichiers de microdonnées les plus complexes de façon à la fois souple et rigoureuse.

L'initiative DDI cherche à établir une norme internationale basée sur le langage XML pour la documentation des microdonnées. L'objectif est de proposer des moyens pratiques pour l'enregistrement et la communication des aspects essentiels des ensembles de microdonnées. La spécification DDI est une évolution majeure du dictionnaire de codes électronique traditionnel : en substance, elle offre les mêmes possibilités, mais améliore considérablement l'étendue et la rigueur des informations fournies. La spécification DDI pour les métadonnées a été initiée par l'ICPSR (consortium interuniversitaire pour la recherche en sciences politiques et sociales), une organisation réunissant plus de 500 universités et établissements d'enseignement supérieur du monde entier. Le projet est désormais celui d'une association d'institutions nord-américaines et européennes, dont les membres comptent des producteurs et des archives de données parmi les plus importants de la planète.

La spécification DDI a été conçue pour prendre en charge tous les types de données issus d'enquêtes, de recensements, de fichiers administratifs, d'expérimentations, d'observations directes et d'autres méthodologies systématiques mises en œuvre pour obtenir des mesures empiriques. Autrement dit, l'unité d'analyse peut être constituée d'individus, de ménages, de familles, d'entreprises, de transactions, de pays ou encore d'autres sujets présentant un intérêt scientifique. De même, les observations peuvent consister en des mesures effectuées de manière ponctuelle dans un milieu donné (sur une semaine au sein d'un échantillon représentatif d'un pays par ex.). Il peut aussi s'agir d'observations répétées réalisées dans plusieurs milieux (données longitudinales et transversales de divers pays, ainsi que données chronologiques et agrégées). La spécification DDI contient également une description exhaustive de la méthodologie de l'étude (mode de collecte des données, méthodes d'échantillonnage le cas échéant, milieu, zones géographiques étudiées, organisme et personnes responsables, etc.).

Structure

La spécification DDI permet de décrire en détails tous les aspects d'une enquête : méthodologie, responsabilités, fichiers et variables. Elle met à disposition une liste structurée et exhaustive de plusieurs centaines d'éléments et d'attributs susceptibles de constituer la documentation d'un ensemble de données. S'il est peu probable qu'ils y figurent tous, certains éléments, tels que le titre (*Title*), sont néanmoins obligatoires – et

4 Pour une description de la norme DDI : http://www.ddialliance.org (site en anglais uniquement)

Encadré 3 **A qui s'adresse la norme DDI ?**

La norme DDI pour les métadonnées est utilisée par une vaste communauté d'archives, notamment des bibliothèques de données universitaires, les gestionnaires de données d'ONS et d'autres producteurs de données officiels, ainsi que des organisations internationales.

Exemples d'utilisateurs dans le milieu universitaire :

- DataFirst, Université du Cap (www.datafirst.uct.ac.za)
- UKDA, Université d'Essex (www.data archive.ac.uk/)
- ICPSR, Université du Michigan (www.icpsr.umich.edu)
- Universités canadiennes participant au programme CDR de Statistique Canada (http://www.statcan.gc.ca/rdc-cdr/network-reseau-eng.htm)
- Réseau DataVerse du Centre de données Harvard-MIT et de la bibliothèque de l'Université de Harvard (http://thedata.org/)
- Agences membres du CESSDA (http://www.cessda.org)

Producteurs de données officiels de plus de 50 pays, notamment :

- Statistique Canada, via l'Initiative de démocratisation des données (IDD) (http://www.statcan.gc.ca/dli-ild/dli-idd-fra.htm)
- Institut national de statistique de Bolivie (http://www.ine.gov.bo/anda/)
- Office national de statistique, bureau de statistiques du travail et de l'emploi (Labour and Employment Statistics) des Philippines (www.census.gov.ph, www.bas.gov.ph, www.bles.dole.gov.ph)
- Département du recensement et des enquêtes du Sri Lanka (http://statistics.sltidc.lk/)
- Agence statistique centrale d'Ethiopie (www.csa.gov.et)
- Et bien d'autres encore… (voir www.ihsn.org/adp)

Organisations internationales :

- UNICEF, pour sa méthodologie d'enquête à indicateurs multiples (MICS) (http://www.childinfo.org/mics3_ surveys.html)
- Banque Mondiale (http://data.worldbank.org/)
- Le Global Fund (http://www.theglobalfund.org/html/ 5YEdata/)

L'adoption de la norme DDI pour les métadonnées est grandement facilitée par l'existence de logiciels conviviaux, tels que DDI Metadata Editor et d'autres outils de catalogage compatibles DDI fournis par IHSN (voir www.ihsn.org/toolkit et www.ihsn.org/nada).

impérativement univoques. D'autres éléments comme l'auteur et le chercheur principal (*Authoring Entity* et *Primary Investigator)* sont optionnels et peuvent apparaître plusieurs fois, car ils concernent la/les personne(s) et/ou la/les organisation(s) responsables de l'enquête. Les éléments de la norme DDI (version 2.4) sont structurés en cinq sections :

Section 1.0 : description du document

L'agence qui se charge de la documentation et de la diffusion de l'étude (enquête, recensement ou autre) n'est pas toujours le producteur des données. Par conséquent, il est important de fournir des informations (métadonnées) non seulement sur l'étude proprement dite, mais aussi sur le processus de documentation. La section Description du document est une présentation générale décrivant le document XML au format DDI, autrement dit « des métadonnées sur d'autres métadonnées ».

Section 2.0 : description de l'étude

Cette section consiste en une présentation générale de l'étude. Elle précise la référence de l'enquête et l'identité de la personne qui a recueilli ou compilé les données, ainsi que de celle qui les a diffusées. Elle résume (synthèse) le contenu des données, informe sur les méthodes de collecte et de traitement, etc.

Section 3.0 : description des fichiers de données

Cette section décrit le contenu de chaque fichier de données, indique le nombre d'enregistrements et de variables, précise la version, le nom du producteur, etc.

Section 4.0 : description des variables

Cette section correspond aux informations sur chaque variable, notamment sur la formulation des questions, univers de l'enquête, l'intitulé des variables et des valeurs, les méthodes de dérivation et d'imputation, etc.

Section 5.0 : autres documents

Cette section permet d'inclure la description d'autres documents relatifs à l'étude, par exemple : ressources documentaires (questionnaires, notes de codage, rapports techniques et analytiques, manuel de l'interviewer, etc.), programmes de traitement et d'analyse des données, photos et cartes.

Norme de métadonnées du Dublin Core (DCMI)

Le contenu de cette section est tiré du site Internet de la DCMI (*Dublin Core Metadata Initiative*) : http://dublincore.org (en anglais uniquement)

> L'ensemble des éléments de métadonnées Dublin Core (norme ISO 15836), également appelé norme de métadonnées du Dublin Core, permet de décrire des ressources électroniques. Cette norme se révèle particulièrement utile pour décrire les ressources sous-jacentes aux microdonnées : questionnaires, rapports, manuels, scripts et programmes de traitement des données, etc. L'idée de l'élaboration d'une telle norme a vu le jour en 1995, sous l'impulsion de l'OCLC (*Online Computer Library Center*) et du NCSA (*National Center for Supercomputing Applications*), réunis à Dublin (Ohio). Progressivement, le Dublin Core est devenu la norme la plus répandue pour la description de ressources numériques sur le Web. Il a été élevé au rang de norme ISO en 2003. La DCMI assure la maintenance et travaille au développement de la norme. Cette organisation internationale a pour objectif de promouvoir l'établissement de normes de métadonnées interopérables.
>
> Le succès de la norme de métadonnées du Dublin Core est lié principalement à sa simplicité. Dès le départ, ses concepteurs ont veillé à ce que l'ensemble d'éléments reste aussi sommaire et simple que possible, afin que la norme puisse être utilisée par les non-spécialistes. L'objectif de la norme est de favoriser une création aisée et à moindre coût de notices descriptives simples pour les ressources pédagogiques, tout en permettant de rechercher efficacement ces mêmes ressources sur le Web ou dans tout autre environnement en réseau. Sous sa forme la plus simple, le Dublin Core comprend 15 éléments de métadonnées, tous optionnels et utilisables à plusieurs reprises. Ces 15 éléments sont :

• Titre	• Relation	• Droits
• Sujet	• Couverture	• Date
• Description	• Créateur	• Format
• Type	• Editeur	• Identifiant
• Source	• Collaborateur	• Langue

Encadré 4 **A qui s'adresse le Dublin Core ?**

Le Dublin Core est une norme de métadonnées extrêmement souple et facile à appréhender. Elle n'est pas aussi détaillée que la norme DDI (notamment, le DC ne contient pas d'éléments correspondant à des informations sur les fichiers de données et les variables), mais elle peut servir à décrire les caractéristiques générales d'un ensemble de données. Elle offre en outre une option permettant de décrire des ensembles de données dont les ressources ne peuvent faire l'objet d'une documentation détaillée. Une version adaptée du Dublin Core est utilisée par l'open data initiative aux Etats-Unis (www.Data.gov).

Encadré 5 **Le langage XML**

XML est l'acronyme de l'anglais e***X****tensible* ***M****arkup* ***L****anguage*. Ce langage offre le moyen de baliser un texte à l'aide de balises qui en déterminent le sens et non pas seulement la présentation. Autrement dit, XML permet de structurer un texte à l'aide de balises contenant des indications sémantiques. Sur le principe, en termes d'organisation, ces « balises » sont identiques aux « champs » d'une base de données. A la différence des fichiers de base de données, les fichiers XML peuvent cependant être visualisés et modifiés à l'aide d'un éditeur de texte standard. Des recherches peuvent être effectuées dans ces fichiers de la même manière que dans une base de données classique, à l'aide des outils appropriés. De la même façon que le contenu d'une base de données peut servir à créer un rapport, les documents XML peuvent être affichés et convertis à l'aide d'autres applications logicielles pour obtenir des formats plus conviviaux (tableurs, PDF ou pages Web).

L'exemple ci-dessous permet de se faire une idée de la façon dont des informations textuelles sur une enquête peuvent se présenter au format XML :

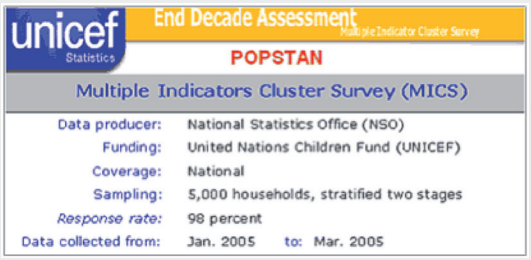
Enoncé : de janvier à mars 2005, l'office national de statistique (ONS) Popstan a mené une enquête à indicateurs multiples (MCIS) financée par l'UNICEF. Un total de 5 000 ménages représentatifs de l'ensemble de la population du pays a été sélectionné aléatoirement pour participer à l'enquête, suivant un plan d'échantillonnage stratifié à deux niveaux. 4 900 d'entre eux ont fourni des informations.
Les mêmes informations converties aux format DDI-XML apparaîtraient comme suit:

```
<titl>Enquête à indicteurs multiples 2005</titl>
<altTitl>MICS</altTitl>
<AuthEnty>Office national de statistique (ONS)</AuthEnty>
<fundAg abbr="UNICEF">Fond des Nations Unies pour l'enfance</fundAg>
<collDate date="2005-01" event="start"/>
<collDate date="2005-03" event="end"/>
<nation>Popstan</nation>
<geogCover>National</geogCover>
<sampProc>5 000 ménages, stratifié à deux niveaux</sampProc>
<respRate>98 pourcents</respRate>
```

Le recours aux balises est d'autant plus efficace quand un ensemble commun de balises est adopté par la communauté (normes DDI ou Dublin Core p. ex.). L'adoption d'un ensemble commun de balises XML présente des avantages majeurs pour la documentation des microdonnées :

- Création d'une liste exhaustive des éléments de métadonnées utiles ;
- Possibilité d'évaluer le contenu d'un fichier en vérifiant s'il contient ou non des balises données ;
- Création d'un catalogue pour les ensembles de données permettant de rechercher des éléments de métadonnées clés;
- Possibilité de convertir le fichier dans des formats plus conviviaux.

Les fichiers XML peuvent être convertis en HTML, au format PDF ou dans d'autres types de documents prenant en charge les transformations XSL. Ils peuvent donc s'échanger d'un réseau à l'autre ou sur Internet via des services Web ou via le protocole SOAP (protocole basé sur le langage XML permettant l'échange d'informations entre applications via HTTP). La page HTML ci-dessous est un exemple de transformation XSL de la version anglaise du fichier XML ci-dessus :



3. Quels sont les arguments en faveur de la diffusion de microdonnées ?

La diffusion est l'une des principales responsabilités d'un institut de statistique. Ce chapitre recense les nombreux avantages que présente la diffusion de microdonnées. L'étude des énoncés de mission, des politiques de diffusion et de l'expérience de plusieurs producteurs de données à travers le monde souligne l'importance et les arguments qui plaident en faveur de la possibilité d'accéder aux fichiers de microdonnées.

3.1 Soutenir la recherche

La principale raison (et souvent la seule motivation explicite) qui pousse les producteurs de données à partager leur microdonnées est de soutenir la recherche. A la fin d'une enquête, les responsables de la collecte des données produisent généralement une série de tableaux destinés à présenter les aspects essentiels et à offrir un aperçu général des résultats aux utilisateurs. On ne peut guère attendre de ces agences qu'elles identifient tous les sujets de recherche que ces données peuvent contribuer à traiter – elles ne disposent d'ailleurs pas de budget pour cela. Les fichiers de microdonnées offrent aux chercheurs une latitude considérable pour l'identification des relations et des interactions entre les phénomènes couverts par une enquête, ce qui favorise la diversité et la qualité des travaux de recherche[5].

Voici quelques exemples destinés à illustrer la façon dont certains instituts nationaux de statistique décrivent l'objet de leur politique de diffusion de données.

> « L'accès aux microdonnées soutient et favorise une prise de décision avisée en permettant un usage étendu des données de l'ABS [*Australian Bureau of Statistics*] aux fins de leur analyse et de la recherche économique et sociale. Depuis 1985, l'ABS met à disposition des microdonnées sous forme de fichiers d'enregistrements unitaires confidentialisés (CURF), sous certaines conditions et à des fins statistiques. Aujourd'hui, l'ABS enregistre de nombreuses demandes d'accès à des enregistrements unitaires plus détaillés, selon des modalités plus souples et provenant d'un éventail plus large d'ensembles de données (données d'entreprise et données longitudinales p. ex.). Toute impossibilité de satisfaire à ces demandes pèsera de plus en plus négativement sur les activités principales de l'ABS, sur l'intérêt de celles-ci et, au final, sur la cohérence du système de statistique national. Cette évolution, associée à une série d'autres facteurs justifiant un changement, notamment le risque croissant d'identification, ont conduit l'ABS à proposer une nouvelle stratégie d'accès à ses microdonnées pour l'avenir. » (Bureau australien de statistique, http://www.abs.gov.au/, traduction de l'anglais)

> « Le principal objectif du CSO [*Central Statistics Office*] en matière d'accès aux microdonnées est de soutenir la communauté de chercheurs et de faire en sorte que les données recueillies par nos soins soient exploitées au maximum. Cette approche préside à l'élaboration de politiques fondées sur des éléments probants. Elle peut en outre contribuer à réduire les coûts de la recherche et à éviter les doublons de données. » (Office central de statistique irlandais, http://www.cso.ie/, traduction de l'anglais)

> « La mission de l'archive nationale de microdonnées du Sri Lanka est de répondre aux besoins de la communauté nationale et internationale de chercheurs, qui s'efforce de trouver des réponses aux problèmes socio-économiques qui se posent dans le monde. Avec l'aide du *Department of Census and Statistics* (office central de statistique du Sri Lanka), LankaDatta diffuse des données statistiques pertinentes, fiables et récentes, produites par les agences du système de statistique national, tout en préservant strictement la confidentialité des répondants. » (*Department of Census and Statistics*, Sri Lanka http://statistics.sltidc.lk/)

> « L'IDD [Initiative de démocratisation des données] est un exemple d'une mise en valeur importante de la technologie de l'autoroute de l'information au Canada puisqu'elle permet aux

5 Le travail de Hamilton et Humphrey [6] illustre bien l'utilisation approfondie qui peut être faite des fichiers de microdonnées par les chercheurs. Il démontre que plusieurs centaines de projets de recherche ont été menés sur la base de l'Enquête nationale sur la santé de la population (ENSP) au Canada, après sa publication sous forme de fichier de microdonnées à grande diffusion. Voir aussi [30].

établissements d'enseignement postsecondaire d'offrir pour la première fois une gamme complète de services de données aux étudiants comme aux professeurs. Il apparaît aussi de plus en plus que l'Initiative contribue de manière importante à l'enseignement et à la recherche au Canada. (…) L'IDD a contribué à l'avènement d'une culture de données au Canada. » (Statistique Canada, http://www.statcan.gc. ca/dli-ild/about-apropos-fra.htm ; voir aussi Watkins [31])

3.2 Renforcer la crédibilité des statistiques officielles

L'accès élargi aux microdonnées témoigne de la confiance des producteurs dans leurs données, dans la mesure où ils autorisent leur reproduction ou leur rectification par des organismes indépendants.

3.3 Améliorer la fiabilité et la pertinence des données

Une relation plus étroite entre les producteurs de données et des utilisateurs avisés peut présenter d'autres avantages. Bien souvent, c'est l'utilisation des données qui permet de se faire une idée des améliorations envisageables, notamment sur le plan de la conception des enquêtes et de la diffusion des microdonnées. Le processus de diffusion de microdonnées peut intégrer une procédure formalisée de feed-back des chercheurs, comme c'est le cas au *Census Bureau* américain. Le feed-back des utilisateurs peut conduire à une amélioration progressive des enquêtes.

3.4 Réduire les doublons de données

Le fait que des fichiers de microdonnées soient mis à disposition des utilisateurs dissuade souvent ces derniers de recueillir eux-mêmes les données dont ils ont besoin. Les répondants sont ainsi moins sollicités et le risque d'obtenir des études incohérentes sur un même sujet est réduit.

3.5 Améliorer le retour sur investissement

« L'accès ouvert aux données de la recherche financée sur fonds publics et leur partage contribuent non seulement à maximiser l'impact des nouvelles technologies et des nouveaux réseaux numériques sur le potentiel de recherche, mais permettent aussi un retour plus important sur l'investissement public dans la recherche. (…)

Des volumes sans cesse croissants de données sont collectés par les chercheurs et établissements de recherche financés sur fonds publics. Cette somme de données de recherche en expansion rapide représente à la fois un investissement massif de fonds publics et l'une des sources potentielles des connaissances nécessaires pour relever les multiples défis auxquels est confrontée l'humanité. (…)

Pour promouvoir un meilleur rendement scientifique et social des investissements publics dans les données de la recherche, les pays membres de l'OCDE, [par exemple], ont établi diverses lois, politiques et pratiques sur l'accès aux données de la recherche au niveau national. Dans ce contexte, des lignes directrices internationales constitueraient un atout important pour encourager les échanges et l'utilisation à l'échelle mondiale des données de la recherche. » [17]

3.6 Lever des fonds pour les études statistiques

Plus les fichiers de données sont diffusés et exploités, plus ils ont de valeur aux yeux des bailleurs de fonds.

Cet argument peut encourager les commanditaires à financer la collecte de données. Certains bailleurs de fonds exigent en effet des preuves d'utilisation.

Une meilleure utilisation des données se traduit par un meilleur retour sur investissement pour les commanditaires, qui seront donc davantage enclins à soutenir les activités de collecte de données. De plus en plus souvent, le financement des enquêtes par des organismes internationaux est soumis à la diffusion des ensembles de données obtenus.

3.7 Réduire les coûts de la diffusion des données

Enfin, les agences de collecte pourraient tirer parti d'une efficience accrue, en ce sens qu'elles pourraient réduire la quantité de tableaux prédéfinis produits et se consacrer davantage à l'élaboration de données de synthèse. Mettre l'accent sur cet aspect – notamment dans les médias et dans le secteur de l'enseignement – peut permettre d'éveiller l'intérêt d'un plus large public et favoriser le financement des activités des agences de collecte, qu'il s'agisse d'ONS ou d'autres types d'établissements. Ces tableaux se révèlent toutefois insuffisants pour des analyses approfondies. Enfin, il faut placer la question de l'efficience dans le contexte

des coûts de production et de diffusion des fichiers de microdonnées. Ces questions sont abordées dans la partie suivante.

3.8 Respecter les obligations légales et contractuelles

Dans certains pays, les agences publiques sont obligées de diffuser une partie de leurs microdonnées. La collecte de données y est souvent financée par le contribuable et donc considérée comme relevant du domaine public. Ailleurs, les bailleurs de fonds sont les organismes commanditaires, qui exigent que les données obtenues soient mises à la disposition des chercheurs (voir Encadré 6). Cette obligation de diffusion des microdonnées n'est pas en contradiction avec l'obligation relative au respect de la confidentialité et de la vie privée d'autrui. Il incombe au statisticien en chef de chaque établissement, ou éventuellement au comité de divulgation, de décider du contenu des microdonnées à publier, ainsi que des procédures à mettre en place pour créer des fichiers à grande diffusion.

3.9 Promouvoir le développement de nouveaux outils d'utilisation des données

Une nouvelle tendance, dite des données ouvertes, se répand depuis quelques années. Le concept fondamental de cette tendance est que les données collectées grâce à des fonds publics ou sous la supervision d'une agence d'Etat relèvent du domaine public.

Plusieurs gouvernements y ont souscrit – voir l'*Open Government Initiative* aux Etats-Unis (www.data.gov) ou l'équivalent britannique (http://data.gov.uk). Ce type de tendance met le public au défi d'ajouter de la valeur aux données existantes. En ouvrant l'accès aux microdonnées sans restriction, de telles initiatives promeuvent le développement de nouveaux outils logiciels, en particulier des applications innovantes du Web 2.

Ces applications Web novatrices, qui utilisent des données ouvertes, sont souvent appelées *mashups* (ou *applications composites*) :

Dans le domaine du développement de sites Web, un *mashup* est une page Web ou une application qui utilise ou combine des données ou des fonctionnalités de plusieurs sources externes dans le but de créer un nouveau service. Cela implique une intégration aisée et rapide, généralement via des API et en utilisant des sources de données ouvertes, l'objectif étant de produire des résultats enrichissants qui ne correspondaient pas nécessairement à la motivation première de la production des données brutes initiales.

Afin de pouvoir accéder en permanence aux données d'autres services, les mashups sont généralement conçus comme des applications clientes ou hébergés en ligne. Les mashups jouent sans doute un rôle actif dans l'évolution des logiciels sociaux et du Web 2.0. (voir aussi http://fr.wikipedia.org/wiki/application_composite).

Afin de favoriser le développement des mashups, les producteurs de données ne se contentant pas de diffuser des données, mais fournissent également souvent des applications logicielles appelées *interfaces de programmation* (API). Ces dernières permettent aux développeurs d'utiliser les données plus facilement.

« Une interface de programmation (*Application Programming Interface* ou *API*) est une interface fournie par un programme informatique. Elle permet l'interaction des programmes les uns avec les autres, de manière analogue à une interface homme-machine, qui rend possible l'interaction entre un homme et une machine. » Les API sont mises en œuvre (ou implémentées) par une application, par une

Encadré 6 **Obligation légale de diffusion des microdonnées : exemple du NCHS (EU)**

Le centre national des statistiques de santé américain (NCHS), rattaché aux Centres de contrôle et de prévention des maladies (US Centers for Disease Control and Prevention), donne un exemple de législation régissant la diffusion de microdonnées.

« Du fait de sa qualité d'institut de statistique fédéral, le NCHS doit faire tout ce qui est en son pouvoir pour maximiser la disponibilité des données, notamment écourter l'intervalle entre la collecte et la diffusion des données, afin d'en optimiser la qualité, mais également réduire au minimum le risque de divulgation. La législation régissant les activités du NCHS stipule que les données doivent être mises à disposition le plus largement possible (…).

« Cependant, l'obligation de diffusion des données doit être appréhendée eu égard à la qualité d'institut de statistique fédéral du NCHS et être mise en correspondance avec le besoin d'assurer la protection de la confidentialité des répondants et de garantir la qualité des données. (…)

« La même loi qui contraint le NCHS à diffuser ses données lui impose également de préserver l'identité des individus ou des organisations qui apparaissent dans ses systèmes de données. » [14] (Traduction de l'anglais)

bibliothèque logicielle ou par un système d'exploitation. Elle permet de déterminer la syntaxe et les conventions d'appel que le programmeur doit respecter pour utiliser les services mis à disposition. Elle peut contenir des spécifications relatives aux routines d'exécution, aux structures de données, aux classes d'objets, ainsi qu'aux protocoles de communication entre l'utilisateur et le composant qui implémente l'API. (http://fr.wikipedia.org/wiki/Inter face_de_programmation, site consulté le 9 janvier 2011 ; voir aussi Encadré 7)

Encadré 7 **Promouvoir les mashups en diffusant des données ouvertes et des API**

Etats-Unis : data.gov – Découvrir, participer, s'engager

L'un des objectifs sous-jacents de l'*Open Government Initiative* est de modifier les habitudes de diffusion d'informations en institutionnalisant une mise disposition élargie des données fédérales dans des formats plus accessibles. Figure de proue de l'*Open Government Initiative*, Data.gov est destiné faciliter l'accès aux ensembles de données fédéraux qui permettent au grand public de mieux appréhender le fonctionnement des agences fédérales et leurs activités, mettent en exergue leurs missions, génèrent des opportunités économiques, et améliorent la transparence et l'ouverture du gouvernement fédéral (ensembles de données à forte valeur ajoutée).

Félicitations aux développeurs

Nous sommes très étonnés du nombre de téléchargements enregistré et d'applications innovantes intégrant des données gouvernementales qui sont apparues depuis le lancement de Data.gov. L'un des principaux objectifs de Data.gov est de favoriser l'innovation. A cet égard, notre communauté de développeurs s'est particulièrement distinguée. Nous encourageons tous les développeurs et programmeurs à explorer les ensembles de données répertoriés sur Data.gov et à participer activement à cette communauté dynamique en pleine croissance.

Source: http://www.data.gov/open, site consulté le 3 avril 2010, traduction de l'anglais).

Royaume-Uni : data.gov.uk – La clé de l'innovation – Montrez-nous la voie

Nous savons qu'il existe en dehors du gouvernement de nombreuses personnes comme vous qui ont les compétences et les capacités requises pour tirer le meilleur parti des données publiques. Tel est pour nous le point de départ de la relation de collaboration que nous souhaitons construire avec vous.

Source: http://data.gov.uk/, site consulté le 3 avril 2010, traduction de l'anglais

L'initiative de données ouvertes porte sur l'accessibilité. Avec le lancement de donnees.banquemondiale.org, nous intensifions les efforts de la Banque Mondiale pour ouvrir ses catalogues de données au moyen d'un accès facile et direct sur le Web. Au cours de 2010, nous allons déployer deux vagues de fonctions Web qui serviront de plate-forme à cet effort. (…)

La phase 2 sera axée sur la promotion de l'utilisation des données de la Banque Mondiale au moyen de l'API sur le Web. Cela nécessitera d'apporter des améliorations à l'API actuelle, mais également de fournir un endroit pour communiquer au sujet de l'API et un seuil plus faible d'entrée, surtout pour les utilisateurs qui ne sont pas développeurs informatiques.

Cette phase marquera le début d'un tournant qui fera passer l'API sur le Web d'une interface destinée exclusivement aux développeurs informatiques à une interface ciblant aussi bien les chercheurs et les décideurs politiques que les développeurs informatiques.

Source: http://donnees.banquemondiale.org/developpeurs, site consulté le 7 mai 2010

4. Quels sont les coûts et les risques liés à la diffusion de microdonnées et comment peuvent-ils être maîtrisés ?

Les ONS et autres producteurs / collecteurs de données doivent tenir compte d'un certain nombre de facteurs pour la définition et la mise en œuvre de leurs politiques et programmes de diffusion de microdonnées : coûts et compétences requises, questions relatives à la qualité des données, possibilité d'une exploitation abusive ou d'une mauvaise interprétation des données par les utilisateurs, aspects juridiques et éthiques, sans oublier la préservation de la confiance et du soutien des répondants.

4.1 Questions éthiques et préservation de la confiance des répondants

Quand ils recueillent des informations auprès d'individus, d'entreprises ou d'autres organisations, les instituts de statistique et autres producteurs de données garantissent généralement aux répondants que les informations fournies seront utilisées uniquement à des fins statistiques. Il s'agit d'une obligation morale et éthique.

« Les ONS doivent conserver la confiance des répondants pour que ceux-ci continuent de collaborer à leurs collectes de données. La protection de la confidentialité est l'aspect clé de cette confiance. Si les personnes interrogées pensent ou ont le sentiment que l'ONS n'est pas en mesure d'assurer la confidentialité des données, elles seront moins enclines à coopérer ou à fournir des informations précises. Même un incident isolé, en particulier s'il s'accompagne d'une forte couverture médiatique, peut avoir un impact significatif sur la collaboration des personnes interrogées et, par voie de conséquence, sur la qualité des statistiques officielles. La protection de la confidentialité est la principale préoccupation des ONS, mais pas la seule Ces derniers ont notamment à cœur de disposer de l'autorité suffisante, soit par le biais d'un mandat légal ou de toute autre forme d'autorisation, pour pouvoir appuyer l'accès des chercheurs aux microdonnées. » [24] (Traduction de l'anglais)

Pour maximiser l'utilisation des microdonnées, les ONS et autres producteurs de données doivent trouver un juste milieu entre la satisfaction des exigences de confidentialité et la fourniture de droits d'accès. L'une des solutions consiste à utiliser différents types de microdonnées (voir chapitres précédents). Il existe une autre option, qui n'est toutefois guère envisageable dans tous les cas de figure : obtenir de chaque répondant qu'il approuve formellement la communication des données recueillies.

Obtenir l'accord de personnes physiques

« L'accord peut être obtenu par signature ou par déduction. Dans le premier cas, le répondant se voit remettre un document écrit décrivant les modalités d'utilisation prévues des informations qui lui sont demandées, qu'il est invité à approuver en apposant sa signature. Le répondant n'est cependant pas toujours informé par écrit, mais parfois seulement oralement. S'il consent à fournir les informations demandées, [l'enquêteur] en « déduira » qu'il approuve l'utilisation envisagée et la communication des données aux parties mentionnées par écrit ou évoquées oralement. [L'enquêteur] sera alors autorisé exclusivement à utiliser les données selon les modalités décrites au répondant. » [15] (Traduction de l'anglais)

Obtenir le consentement d'une personne morale

« Dans le cas de personnes morales, la méthode varie selon que les informations sont demandées de vive voix ou par mail.

A. Si les informations sont demandées de vive voix par un membre du personnel ou un agent de [l'enquêteur], le contact cherche d'abord à établir qui est habilité à les fournir pour le compte de l'organisation. Quand cette personne répond aux questions après avoir été informée de l'utilisation qui sera faite des données recueillies, le représentant de [l'agence enquêtrice] considère que l'organisation approuve celle-ci.

B. Pour les données collectées par mail, la demande peut être adressée à l'organisation, à son gestionnaire ou à toute autre personne préalablement identifiée par [l'agence enquêtrice] comme étant habilitée à fournir les informations requises. La lettre contenant la

> demande d'informations précise l'utilisation qui sera faite des données. A la réception des données, le personnel de [l'agence enquêtrice] considère que l'organisation approuve leur utilisation selon les modalités dont il a été informé. » [15] (Traduction de l'anglais)

4.2 Aspects juridiques

Un producteur de données est-il légalement en droit de diffuser des fichiers de microdonnées ? Il n'y a pas de réponse unique à cette question. La législation qui régit les activités des producteurs de données est spécifique à chaque pays et à chaque programme cadre (Encadré 8).

Nous l'avons évoqué, la diffusion des microdonnées est parfois une obligation légale. Dans la majorité des cas, la législation prévoit toutefois de simples restrictions. Par conséquent, la politique d'un pays en matière de diffusion des microdonnées sera déterminée par son cadre législatif. Il est primordial pour les producteurs de données de « faire en sorte de disposer d'une base juridique et éthique solide (ainsi que des instruments techniques et méthodologiques nécessaires) pour protéger la confidentialité. Cette base juridique et éthique nécessite que l'on mette soigneusement en balance l'intérêt pour le public d'une solide protection de la confidentialité, d'une part, et les avantages que lui apporte la recherche, d'autre part. La décision de savoir s'il convient ou non de permettre à un chercheur d'avoir accès aux données pourrait dépendre des mérites de son projet de recherche et de sa crédibilité. Il faudrait en tenir compte d'une façon ou d'une autre dans les dispositions législatives. » [24] (Traduction de l'anglais)

« Les dispositifs d'accès aux données devraient respecter les droits et intérêts légitimes de tous les acteurs de l'activité de recherche publique. L'accès à certaines données de la recherche et leur utilisation seront nécessairement limités par divers types de prescriptions légales, qui peuvent imposer des restrictions pour raisons de :

- Sécurité nationale : certaines données relatives au renseignement, aux activités militaires ou à la prise de décision politique peuvent être classifiées et partant, soumises à un accès limité.

- Protection de la vie privée et confidentialité : les données relatives aux sujets humains et d'autres données personnelles font l'objet d'un accès limité en vertu des législations et des politiques nationales de protection de la confidentialité et de la vie privée. Il convient toutefois que les propriétaires de ces données envisagent des procédures d'anonymisation ou de confidentialité permettant d'assurer un niveau de confidentialité satisfaisant afin de préserver autant que possible l'utilité des données pour les chercheurs.

- Secrets commerciaux et droits de propriété intellectuelle : les données concernant les entreprises ou autres parties, ou provenant de celles-ci, qui contiennent des informations confidentielles peuvent être inaccessibles pour la recherche. (...)

L'adhésion à des codes de conduite professionnels peut faciliter le respect des prescriptions légales. » [17]

Principes fondamentaux de la statistique officielle des Nations Unies

De nombreux pays se sont appuyés sur les *Principes fondamentaux de la statistique officielle des Nations Unies* pour élaborer leur législation. Il est donc intéressant de se pencher sur ces principes, dans la mesure où ils ont trait à la confidentialité des statistiques.

Le sixième des *Principes fondamentaux de la statistique officielle des Nations Unies* est formulé comme suit : « Les données individuelles collectées pour l'établissement des statistiques par les organismes qui en ont la responsabilité, qu'elles concernent des personnes physiques ou des personnes morales, doivent être strictement confidentielles et ne doivent être utilisées qu'à des fins statistiques. » [27]

Tous les principes régissant l'accès aux microdonnées doivent satisfaire à cette exigence ou aux dispositions de la législation applicable par l'ONS. Les points suivants doivent être pris en compte dans le cadre de la gestion de la confidentialité des microdonnées.

Principe 1 : utilisation appropriée des microdonnées

> « On peut exploiter les microdonnées réunies dans le cadre de la statistique officielle, aux fins de l'analyse statistique, en vue d'étayer des recherches pour autant que la confidentialité de ces données soit protégée. (...)

La communication des microdonnées aux chercheurs n'est pas incompatible avec le sixième Principe fondamental de l'ONU tant qu'il demeure impossible d'identifier les données se rapportant à un individu. Le Principe 1 ci-dessus ne constitue pas une obligation de diffuser les microdonnées. C'est à l'ONS qu'il devrait appartenir de décider s'il convient ou non de fournir des microdonnées. D'autres considérations (la qualité des microdonnées p. ex.) peuvent rendre inopportun l'accès aux microdonnées. Il arrive aussi qu'il soit inapproprié de fournir des microdonnées à des personnes ou institutions bien précises. » [24] (Traduction de l'anglais)

Principe 2 : les microdonnées ne devraient être communiquées qu'à des fins statistiques

« Selon le principe 2 susmentionné, il convient de faire la distinction entre une utilisation des données à des fins statistiques ou analytiques et un usage administratif de celles-ci. Dans le premier cas, l'objectif poursuivi consiste à établir des statistiques se rapportant à un groupe (personnes physiques ou morales). A l'inverse, le but recherché dans l'utilisation administrative des données est d'obtenir des informations sur une personne physique ou morale en vue de prendre une décision susceptible de profiter ou de nuire à un particulier, comme par exemple des demandes de données individuelles par décision judiciaire. Afin de remporter la confiance du public dans le système statistique officiel, ces demandes, certes légales, sont incompatibles avec ce principe et devraient être systématiquement rejetées.

Si l'utilisation escomptée des microdonnées semble ne pas suivre des fins statistiques ou analytiques, l'accès aux microdonnées ne devrait pas être accordé. Des comités d'éthique – ou tout système analogue – peuvent prêter leurs concours lorsque le doute concernant l'accessibilité aux microdonnées s'installe.

Les chercheurs consultent les microdonnées aux fins de la recherche. Cela comprend l'élaboration d'agrégats statistiques de natures diverses, l'établissement de distributions statistiques, l'ajustement de modèles statistiques, ou l'analyse de différences statistiques entre sous-populations. Ces utilisations sont compatibles avec les objectifs statistiques. A cet égard, les microdonnées peuvent être considérées comme utiles pour appuyer la recherche. » [24] (Traduction de l'anglais)

Principe 3 : la fourniture de microdonnées devrait s'effectuer en accord avec les dispositions juridiques et autres qui garantissent la confidentialité des microdonnées communiquées

« En ce qui concerne le Principe 3, il faudrait idéalement mettre en place des dispositions juridiques visant à protéger la confidentialité avant de publier quelque microdonnée que ce soit. Toutefois, les dispositions légales doivent être complétées par des mesures administratives et techniques tendant à réglementer l'accès aux microdonnées et à assurer que les données personnelles ne puissent être divulguées. L'existence et la visibilité de telles modalités (qu'elles soient consacrées par la loi ou par des règlements supplémentaires, des arrêtés, etc.) sont indispensables pour susciter du public une plus grande confiance dans le bon usage qui sera fait des microdonnées. De toute évidence, l'instauration de dispositions légales est à privilégier. Dans les pays où ce n'est pas possible, il convient d'instituer une autre forme de dispositions administratives. Il faudrait également que, là où elles existent, les autorités compétentes en matière de confidentialité valident les dispositions légales (ou autres) avant qu'elles ne soient consacrées par la loi. En l'absence d'autorité de cette nature, certaines ONG peuvent exercer une fonction de « surveillance » pour les questions de confidentialité. Il serait judicieux d'obtenir qu'elles souscrivent à toute forme de disposition juridique ou autre mise en place, ou tout au moins de répondre à n'importe quelle inquiétude sérieuse qu'elles pourraient avoir.

Tous les pays ne sont pas dotés d'un fondement législatif. Au minimum, il convient que la divulgation des microdonnées bénéficie du soutien d'une forme quelconque d'autorité. Il est cependant préférable d'instaurer une législation habilitante. » [24] (Traduction de l'anglais)

Principe 4 : il faudrait assurer la transparence des modalités d'accès des chercheurs aux microdonnées, ainsi que des utilisations et utilisateurs de microdonnées, et les rendre publics

> « Le Principe 4 est important pour rassurer le public quant au fait que les microdonnées sont exploitées à bon escient et pour montrer que les décisions touchant à la diffusion des microdonnées sont prises sur une base objective. Il appartient à l'ONS de déterminer si les microdonnées peuvent être divulguées, selon quelles modalités et à quel utilisateur. Il convient néanmoins d'assurer la transparence de ses décisions. Le site Web de l'ONS constitue un instrument efficace pour garantir le respect des règles établies mais aussi pour fournir des informations sur les modalités d'accès aux rapports d'études fondées sur les microdonnées mises à disposition. » [24] (Traduction de l'anglais)

4.3 Exposition aux critiques et à la contradiction

« Certains ONS craignent que la qualité de leurs microdonnées ne soit pas suffisante pour faire l'objet d'une plus large diffusion. Si la qualité peut s'avérer suffisante pour permettre l'élaboration d'agrégats statistiques, elle est parfois insuffisante pour une analyse très détaillée. Dans certains cas, des ajustements ont été apportés aux statistiques agrégées, au stade de l'édition des résultats, sans que les microdonnées aient été modifiées. Par conséquent, des incohérences entre les résultats de recherches fondées sur les microdonnées et celles basées sur les données agrégées publiées peuvent apparaître ». [24] (Traduction de l'anglais.) Les parties d'ensembles de données considérées comme insuffisamment fiables peuvent être supprimées avant la diffusion. Il convient que les producteurs de données pratiquent l'ouverture et la transparence nécessaires sur la question de la qualité.

Par ailleurs, la communication des microdonnées aux chercheurs va de pair avec la possibilité de voir publier des résultats non conformes aux estimations de l'agence qui les a produites. S'il s'agit d'une agence de statistique officielle, cela peut entraîner un conflit entre des estimations (officielles et officieuses) contradictoires, et conduire à une remise en question des données, voire à des répercussions politiques. Plusieurs facteurs peuvent être à l'origine de tels écarts. Tout d'abord, il se peut que les estimations officielles soient erronées, auquel cas un contrôle extérieur est positif. Ensuite, les écarts peuvent être liés à l'utilisation de différentes versions des données (fichier principal complet ou version publique antonymie / restreinte, autres modifications effectuées par le chercheur, etc.). Les écarts devraient alors être minimes et trouver facilement une explication.

Enfin, les écarts peuvent résulter de différences méthodologiques. Ceci est souvent plus problématique pour les producteurs de données, dans la mesure où le grand public n'est pas toujours à même de saisir des explications techniques complexes. Il est important que les producteurs de données sachent défendre leurs estimations. Cela implique que la collecte, le traitement et l'analyse des données soient documentés de A à Z. Qui plus est, ces informations doivent être facilement accessibles. Il arrive que les résultats publiés aient été élaborés par ou avec le concours d'experts externes, qui ne sont plus là pour répondre aux questions soulevées. Les producteurs de données peuvent prévenir ce risque en adoptant et en mettant en œuvre des pratiques de documentation et de conservation très strictes, conformes à la norme de reproduction des données. En substance, cette norme peut être décrite comme suit : « (…) Le seul moyen d'appréhender et d'évaluer une analyse empirique est de connaître avec précision le processus de création et d'analyse des données. (…) La norme de reproduction prévoit que la quantité d'informations disponible soit suffisante pour comprendre, évaluer et se référer à des travaux antérieurs au cas où une tierce partie répliquait les résultats sans autre information de l'auteur. » [10] (Traduction de l'anglais, voir aussi [11]).

4.4 Coûts

« Les ONS peuvent également être préoccupés par la question des coûts, non seulement ceux qui impliquent la création et la documentation de fichiers de microdonnées, mais aussi ceux associés à la mise en place d'instruments d'accès et de protection, ainsi qu'au soutien et à l'autorisation d'enquêtes réalisées par la communauté des chercheurs ; en effet, les nouveaux utilisateurs de leurs fichiers de données ont besoin d'aide pour s'y retrouver dans les structures de fichier complexes et les définitions de variables. Bien que les coûts en question soient à la charge des ONS, ceux-ci ne disposent généralement pas d'enveloppe budgétaire pour financer les travaux supplémentaires à entreprendre dans ce contexte. Quant aux chercheurs, ils n'ont généralement pas les moyens d'assumer une

part substantielle de ces coûts. » [24] (Traduction de l'anglais.) Par conséquent, dans la mesure du possible, ces coûts doivent être intégrés au budget de l'enquête et servir à optimiser l'exploitation des résultats obtenus. Il est dans l'intérêt général que les conclusions tirées des données soient communiquées, afin d'en informer les décideurs politiques et le public. De plus, une plus large utilisation des données d'enquête constituera une barrière supplémentaire à la réduction des enveloppes budgétaires attribuées aux programmes statistiques. Les enquêtes peu utiles en termes d'aide à l'élaboration des politiques publiques sont davantage menacées.

4.5 Perte d'exclusivité

En diffusant les microdonnées, leurs propriétaires perdent leur droit exclusif de consultation des données. Cet aspect est plus problématique pour les chercheurs universitaires que pour les producteurs de données officiels, même si ces derniers (ou certains membres de leur personnel) profitent parfois de leur droit d'accès exclusif pour proposer des services de consultation. De plus en plus souvent, les commanditaires des enquêtes définissent une durée « raisonnable » d'accès exclusif aux données pour le producteur. A l'issue de cette période, l'accès aux données doit être ouvert à d'autres utilisateurs.

4.6 Capacités techniques

Certaines capacités techniques sont requises pour pouvoir assurer la diffusion de fichiers de microdonnées. Les fichiers doivent être accompagnés d'une documentation détaillée (de préférence conforme à la norme DDI pour les métadonnées) et conservés de façon appropriée. Qui plus est, ils doivent être examinés en vue d'évaluer le risque de divulgation d'informations personnelles - et la diminution de ce risque entraînée par le recours à diverses techniques. Les exigences techniques relatives à la diffusion des fichiers de microdonnées sont exposées plus en détails au chapitre 10.

Encadré 8 **Exemples de législation en matière de confidentialité**

Voici quelques exemples de lois sur la statistique et de modalités de diffusion des microdonnées :

1. Les activités du *Census Bureau* américain sont régies par les dispositions de la section 9 du titre 13 du Code des Etats-Unis. « En vertu des dispositions de la section 9 du titre 13 du Code des Etats-Unis, il est interdit de publier ou de diffuser toute information pouvant permettre d'identifier une organisation, un individu ou un ménage en particulier. Le contrôle de divulgation englobe les mesures prises pour s'assurer que les données mentionnées au titre 13 soient préparées avant publication, de façon à être protégées contre une divulgation abusive. Ces mesures comprennent des méthodes de restriction de divulgation et une procédure de contrôle visant à garantir que les méthodes employées offrent une protection adéquate des données. » (Traduction de l'anglais)

   Source: http://www.census.gov/srd/sdc/wendy.drb.faq.pdf

2. La Loi sur la statistique canadienne stipule la règle suivante : « aucune personne qui a été assermentée en vertu de l'article 6 ne peut révéler ni sciemment faire révéler, par quelque moyen que ce soit, des renseignements obtenus en vertu de la présente loi de telle manière qu'il soit possible, grâce à ces révélations, de rattacher à un particulier, à une entreprise ou à une organisation identifiables les détails obtenus dans un relevé qui les concerne exclusivement. »

   Source: http://www.statcan.gc.ca/about-aperçu/act-loi-fra.htm

   Statistique Canada diffuse des fichiers de microdonnées. Sa politique sur la diffusion des microdonnées stipule que la publication des fichiers de microdonnées est soumise aux conditions suivantes :

   (a) la diffusion améliore sensiblement la valeur analytique des données recueillies ; et
   (b) l'agence a pris toutes les mesures possibles pour empêcher l'identification d'unités d'observation de l'enquête.

3. En Thaïlande, la loi contient les dispositions suivantes : « Les informations personnelles recueillies sous le régime de la présente loi doivent être traitées avec la plus grande confidentialité. Toute personne qui a une obligation au titre des présentes ou qui est chargée de la conservation de ce type d'informations devra s'abstenir de les divulguer à quiconque n'a pas d'obligation au titre des présentes, excepté dans les cas suivants:

   (1) Les informations sont divulguées dans le cadre d'une enquête ou de poursuites relatives à une infraction au titre des présentes.
   (2) Les informations divulguées sont destinées à être utilisées aux fins de la préparation et de l'analyse de statistiques ou de la recherche statistique, sous réserve qu'une telle divulgation ne nuise pas aux propriétaires des données et que l'identité de ces derniers soit préservée. » (Traduction de l'anglais)

   Source: http://web.nso.go.th/eng/en/about/about0.htm section 15.

5. A qui les microdonnées sont-elles destinées ?

Les fichiers de microdonnées sont destinés à des spécialistes dotés de solides compétences quantitatives. Exemples :

- Décideurs politiques et chercheurs employés par des administrations et des services de planification.
- Organisations internationales et autres commanditaires.
- Instituts universitaires et centres de recherche impliqués dans la recherche économique et sociale.
- Personnel universitaire et étudiants ; et
- Autres utilisateurs impliqués dans la recherche scientifique.

L'un des fondamentaux de la diffusion de microdonnées est *l'équitabilité*. Lorsque la diffusion est juridiquement possible, les microdonnées issues d'opérations de collecte financées sur fonds publics doivent de préférence être proposées à tous les utilisateurs potentiels sur le principe d'un accès ouvert, avec les métadonnées les plus complètes possibles. « Par ouverture, on entend l'accès dans des conditions d'égalité de la communauté scientifique internationale, à un coût le plus bas possible, de préférence ne dépassant pas le coût marginal de la diffusion. L'accès ouvert aux données de la recherche financée sur fonds publics devrait être aisé, rapide, convivial et passer de préférence par l'internet. » [17]

Les ONS proposent généralement des produits différents, destinés à des publics tout aussi différents. Les données de synthèse (tableaux, graphiques, analyses) s'adressent généralement à un vaste public. Elles sont publiées et placées sur les sites Web des organismes. Les fichiers de microdonnées sont quant à elles destinées aux chercheurs de diverses institutions (agences gouvernementales et ministères, ONG, instituts de recherche, universités et organisations internationales p. ex.) communément appelée « communauté de chercheurs ».

« Que désigne la communauté de chercheurs ? Elle englobe bien sûr ceux qui travaillent dans les établissements universitaires, mais aussi les chercheurs officiant pour des organisations non gouvernementales (ONG) et des organisations internationales. Par ailleurs, certains des chercheurs désireux d'avoir accès aux microdonnées travaillent au sein d'agences ou d'institutions financées par le gouvernement. » [24] (Traduction de l'anglais)

De nombreuses lois sur la statistique stipulent que les données doivent être mises à disposition à des fins statistiques légitimes et pour soutenir la recherche. Une utilisation commerciale dans un cadre contractuel est généralement considérée comme incompatible avec cette règle. L'accent mis sur l'utilisation potentielle des microdonnées est un aspect fondamental pour l'adhésion aux lois qui régissent les activités des ONS.

L'Office statistique de l'Union européenne, Eurostat, définit l'accès aux microdonnées comme suit :

> « Les microdonnées sont principalement accessibles à la communauté scientifique, aux instituts ou aux universités, à des fins de recherche uniquement et sur présentation d'un projet de recherche. Dans des cas très rares, l'accès est étendu aux étudiants ou à un public non scientifique plus large. » Les ensembles de microdonnées anonymisés mis à la disposition du grand public sont appelés fichiers à grande diffusion. S'ils sont largement et gratuitement diffusés par les instituts de statistique, ils présentent toutefois un intérêt limité en termes de pertinence pour l'action publique. « En général, les entreprises privées n'ont pas le droit de travailler avec des microdonnées[6]. »

Si la diffusion des microdonnées est généralement axée sur la communauté de chercheurs, certains qualificatifs y sont fréquemment associés. Ainsi, il est souvent fait référence à des utilisateurs *dignes de confiance*, à savoir d'authentiques chercheurs officiant en toute bonne foi.

En réalité, la « propriété » des fichiers de données n'est pas transférée aux chercheurs : ces derniers bénéficient d'une licence d'utilisation non transférable. Cela permet à l'ONS de déterminer le but recherché. Il est souhaitable que l'ONS élabore une politique lui permettant de formuler les conditions d'accès à ses données.

---

6 Le site Internet d'Eurostat sert de plate-forme d'accès à des microdonnées . Le lien ci-dessous permet d'avoir un aperçu des procédures de confidentialité appliquées par Eurostat : http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/statistical_confidentiality/confidential_data/introduction, site consulté le 10 janvier 2011

Il convient que cette politique soit la plus générale possible et contiennent des informations pour tous les types d'utilisateurs. Des conditions spécifiques peuvent s'appliquer à certains utilisateurs en fonction de leur nationalité ou de leur organisme d'appartenance. Exemples :

Utilisateurs nationaux :

- Les représentants officiels explicitement concernés par la législation régissant les statistiques peuvent bénéficier d'un accès facilité aux microdonnées officielles, dans la mesure où ils sont soumis aux mêmes règles que l'ONS. Ils peuvent être assermentés et les sanctions sont plus faciles à exécuter – ce qui sert les objectifs du gouvernement.

- D'autres utilisateurs nationaux officiant sous le régime de la législation sur les statistiques ont toutefois la possibilité d'utiliser des fichiers de données, à condition qu'ils signent un contrat de licence adapté.

Utilisateurs internationaux :

- Dans de nombreux cas, il existe une obligation de communication des données à des organisations internationales. Cette obligation peut être liée à l'appartenance de l'agence à un groupe international ou à des contrats de financement de projets importants, ou encore à la participation du pays concerné à des projets de développement internationaux.

- L'octroi de licences d'utilisation à des chercheurs officiant pour des universités ou des centres de recherche étrangers est plus critique. Il est en effet plus difficile de faire respecter les dispositions applicables dans ces cas de figure. Il peut néanmoins être intéressant de partager des données avec ces chercheurs – qui représentent un riche pool d'expertise. Le risque peut être minimisé en accordant la licence à l'établissement (université p. ex.) plutôt qu'à un individu en particulier (lire les informations complémentaires fournies plus loin).

Assistants techniques :

Le personnel des ONS bénéficie souvent du concours de consultants, étrangers ou non, pour le traitement et / ou l'analyse des données d'enquête. Dans la mesure où cette coutume est compatible avec les objectifs de l'ONS, il est recommandé d'en faire mention dans la politique de diffusion – sous réserve que les consultants signent le même contrat que les autres chercheurs. Les clauses de ce contrat stipuleront de préférence que les données ne peuvent pas être communiquées sans l'accord préalable de leur producteur. Il convient en outre d'exiger une déclaration de confidentialité.

Les ONS peuvent transférer une partie des risques aux chercheurs en prenant les mesures suivantes :

(i) « Leur demander de prouver leur légitimité en qualité de chercheurs, de démontrer l'intérêt de leurs recherches pour le public et de fournir des preuves que l'usage des microdonnées est nécessaire dans ce cadre ;

(ii) Faire en sorte que les chercheurs signent un engagement juridiquement contraignant prévoyant des sanctions analogues à celles qui s'appliquent aux fonctionnaires de l'ONS s'ils enfreignent les règles en matière de confidentialité ;

(iii) Expliquer les raisons de la prudence dont font preuve les ONS. S'assurer que les chercheurs sont pleinement conscients de leurs obligations en les informant comme il se doit. Instaurer des procédures de suivi et de surveillance efficaces. Il pourrait s'avérer utile d'établir un code de conduite en collaboration avec la communauté des chercheurs ;

(iv) Lorsque des infractions sont commises, refuser au chercheur et éventuellement à l'organisme dont il relève de leur fournir tout service pendant un certain temps (p. ex. jusqu'à ce que l'organisme ait pris les mesures disciplinaires qui s'imposent à l'égard du chercheur en faute). Il est crucial de leur faire comprendre qu'une fronde du public pourrait hypothéquer la divulgation future de microdonnées à la communauté de chercheurs. Entreprendre une action en justice si cela est opportun. » [24] (Traduction de l'anglais)

Quand un chercheur dépose une demande d'accès pour le compte d'une organisation, il est souhaitable qu'il ne puisse le faire qu'au nom de celle-ci (ou de son agence, employeur, etc.) et non pas en tant qu'individu. Ces organisations ont une réputation à défendre et sont donc mieux à même de faire respecter les engagements pris. Les contrats en question sont difficiles à faire appliquer dans le cas de demandes d'accès transfrontalières. L'une des solutions à ce problème consiste à collaborer avec les archives de données ou les ONS du pays demandeur.

Encadré 9 **Exemple de déclaration de confidentialité**

Je m'engage :

1, à ne faire aucune copie partielle ou totale des fichiers auxquels je pourrais être autorisé(e) à accéder par le personnel du Centre national de données sécurisé (ci-après le Centre). J'ai conscience qu'aucune des données ou informations personnelles consultées ou obtenues en vertu de ma position de chercheur au Centre ne pourra être emportée à l'extérieur du Centre.
2. à restituer au personnel du Centre l'ensemble des documents confidentiels qui me seront confiés durant mes recherches par le Centre, ainsi que tout autre document demandé.
3. à ne rien entreprendre pour identifier une personne, une organisation ou une unité d'échantillonage dont l'identité ne serait pas mentionnée dans les fichiers de données à grande diffusion.
4. à garder la plus grande confidentialité sur l'identité des organisations ou des personnes découverte inopinément, que ce soit dans un document, à travers une discussion ou dans le cadre d'une analyse. Toute découverte d'identité involontaire dans le cadre de mes travaux d'analyse sera immédiatement signalée au personnel du Centre.
5. à ne pas emporter d'impressions, de fichiers électroniques, de documents ou tout autre support sans qu'ils aient été soigneusement examinés au préalable par le personnel du Centre en vue d'éliminer tout risque de divulgation. J'ai conscience que le Centre procèdera à un contrôle de divulgation et que tout retrait de données, que ce soit au format électronique ou papier, sera soumis à l'autorisation préalable du Centre.
6. à n'emporter aucune information manuscrite sur l'identité d'une organisation ou d'une personne ou sur des coordonnées géographiques établies dans le cadre de mes recherches au Centre.
7. à adopter un comportement conforme aux principes et aux normes de bonne conduite considérés comme adaptés à un établissement de recherche scientifique.

J'ai été informé du fait que toute infraction délibérée de l'une des conditions ci-dessus entraînera l'annulation de l'autorisation d'accès aux données. J'ai également conscience du fait qu'une telle infraction risque de me fermer définitivement les portes du Centre pour autant que son Directeur le juge nécessaire pour protéger l'intégrité la confidentialité du Centre.

_____ _____
Nom et signature du consultant Date

_____ _____
Nom et signature du témoin Date

Cette déclaration de confidentialité s'inspire de la version 2008 du contrat régissant les conditions d'accès aux données du Research Data Center du Centre national des statistiques de santé américain (NCHS) (*Agreement Regarding Conditions of Access to Confidential Data in the Research Data Center of the US National Center for Health Statistics*) [16]

6. Quelles sont les conditions régissant la diffusion de microdonnées ?

La majorité des producteurs de données financés sur fonds publics ont pour mission de diffuser les données qu'ils recueillent le plus largement possible. Tous ces organismes sont toutefois soumis à des obligations morales et légales qui leur imposent des modalités de diffusion restreintes. « Il est primordial de diffuser au plus grand nombre les statistiques utiles, mais il est tout aussi important de le faire selon des modalités qui ne nuiront pas aux organismes qui les ont produites. » [15] (Traduction de l'anglais)

« Un ensemble de mesures juridiques, administratives et techniques devra [donc] être mis en place pour gagner et conserver la confiance du public. » [24] (Traduction de l'anglais)

Les conditions régissant la diffusion des microdonnées doivent être formelles et transparentes. Idéalement, elles sont définies dans une politique ainsi que dans des notes de procédure et divers protocoles. Ce chapitre contient des informations sur la formulation de ces politiques et de ces procédures. Il faut toutefois souligner que ces informations ne constituent pas une norme recommandée. Il appartient à chaque producteur de données de définir sa propre politique et ses propres procédures, sur la base de considérations à la fois techniques, juridiques et éthiques.

Habituellement, une politique de diffusion contient des dispositions à caractère plutôt général, notamment :

- L'objet de la politique, qui a pour vocation de « définir le type de fichiers de microdonnées destinés à la diffusion, les fins auxquelles ils peuvent être utilisés, ainsi que les conditions régissant leur mise à disposition ».

- Une brève déclaration de principe, via laquelle l'établissement explique en quoi il considère qu'il est important de diffuser des microdonnées (lire le chapitre 3).

- La législation faisant autorité en matière de diffusion de microdonnées (voir section 6.1).

- Les grands principes qui président à la préservation de la vie privée et de la confidentialité des données.

- Les délais de mise à disposition des données. Exemple : « Conscient de l'importance des besoins des utilisateurs et de la nécessité de mettre à disposition des données récentes, l'ONS s'efforcera de diffuser les fichiers de microdonnées dans un délai de 6 à 12 mois à compter de la première publication des données de l'enquête concernée. La publication des données d'enquête s'effectue parfois en plusieurs phases, de façon à ce qu'elles puissent être contrôlées et analysées par l'ONS avant d'être anonymisées en vue d'une utilisation ultérieure par un autre établissement. » (Voir http://www.surveynetwork.org/home/index.php?q=tools/dissemination/policy, traduction de l'anglais)

- Les principales fonctions et responsabilités concernant la définition et la mise en œuvre de la politique de diffusion et des procédures correspondantes. Exemples :

Directeurs d'enquête, avec les attributions suivantes :

- Identifier les besoins des principaux intéressés et veiller à la création d'un fichier anonymisé permettant de répondre aux besoins de la communauté d'utilisateurs tout en respectant la législation sur les statistiques ; et

- Initier un premier contrôle du fichier de microdonnées, identifier les problèmes éventuels à résoudre et soumettre une version préliminaire au comité de divulgation des microdonnées.

Comité de divulgation des microdonnées, avec les attributions suivantes :

- Examiner toutes les demandes de diffusion de fichiers de microdonnées anonymisés déposées par les directeurs d'enquête, suivant les critères préétablis ;

- Valider la diffusion de tous les fichiers ou indiquer aux directeurs d'enquête comment ils peuvent être améliorés ;

- Superviser l'octroi de licences et résoudre les risques liés à d'éventuelles infractions ; et

- Mettre à jour les lignes directrices à l'attention des directeurs d'enquête, relatives à la création de fichiers de microdonnées anonymisés.

Commission de diffusion, avec les attributions suivantes :

- Examiner les demandes d'accès des chercheurs portant sur des fichiers de microdonnées sous licence ;

- Fournir aux utilisateurs autorisés l'accès aux fichiers de données ; et

- Répondre aux demandes d'assistance et d'informations complémentaires des utilisateurs.

Le **Directeur général de l'établissement** valide toute communication de fichiers de microdonnées anonymisés aux utilisateurs, sur la base des conseils et des recommandations du Comité de divulgation.

- Une description générale de la politique tarifaire de l'établissement. Idéalement, cette politique favorisera l'utilisation la plus large possible des données produites en les rendant abordables. Il convient donc que les producteurs de données veillent à ce que les coûts liés à la création de fichiers de microdonnées anonymisés soient intégrés au budget des enquêtes. Parallèlement, l'ONS pourra légitimement tenter de récupérer les coûts liés à la fourniture de services spéciaux destinés uniquement à une catégorie d'utilisateurs en particulier.

Des informations plus détaillées sur les procédures et les conditions relatives à la diffusion des fichiers de microdonnées sont contenues dans les protocoles et les notes de procédure, notamment :

- Modalités de demande d'accès aux données (demandes en ligne, formulaires à utiliser, etc.)

- Autorisations et restrictions pour les différents types d'ensembles de données (voir section 6.2)

- Entité responsable de l'octroi des autorisations d'accès et autres renseignements pratiques sur les procédures de contrôle

- Méthodes de contrôle de la divulgation statistique retenues

- Informations à fournir par les chercheurs et destination possible de ces informations

- Politique tarifaire détaillée

- Forme et étendue de l'assistance technique dont bénéficient les utilisateurs

- Autres renseignements pratiques

La *UK Statistics Authority* a publié un guide des bonnes pratiques en matière de statistiques officielles [22]. Bien que non spécifique à la diffusion des fichiers de microdonnées, ce guide constitue un bonne exemple de principes et de protocoles clairement formulés.

6.1 Fondement législatif

Un fondement législatif est essentiel à plusieurs égards:

« (i) pour gagner la confiance du public dans les dispositions établies - en d'autres termes qu'ils sache que des contraintes juridiques déterminent ce qui est ou non autorisé ;

(ii) pour qu'une compréhension mutuelle s'instaure entre les ONS et les chercheurs en ce qui concerne ces dispositions ;

(iii) pour qu'une plus grande cohérence soit assurée dans le traitement des projets de recherche ; et

(iv) pour que soient mis en place les fondements d'un système visant à sanctionner les infractions. » [24]

Certains ONS ne peuvent se référer à aucune disposition légale sur les statistiques pour la diffusion des fichiers de microdonnées. Celle-ci est même parfois explicitement interdite. Dans d'autres lois, le sujet n'est pas explicitement abordé et laissé à l'appréciation des établissements concernés. C'est le cas notamment quand la date de création de la législation remonte à une période où une telle disposition n'était pas envisageable. La diffusion des fichiers de microdonnées peut alors être soumise à une révision préalable de la législation

en vigueur. La Loi sur la statistique canadienne, par exemple, date du tout début du XX^{e} siècle, mais fut révisée en 1971 - entre autres pour habiliter les ONS à diffuser des fichiers de microdonnées.

Cependant, « il n'est pas indispensable que les dispositions en question soient énoncées dans une loi. Les détails .../... peuvent plus aisément être précisés par le biais de règlements, d'arrêtés, etc., qui n'en ont pas moins un effet juridique. En l'absence de dispositions légales, l'une ou l'autre forme d'autorisation est essentielle. La réputation de l'ONS sera compromise s'il n'existe pas une forme quelconque d'autorité s'exerçant sur la divulgation des microdonnées, même anonymisées. Il importe que la législation (ou l'autorisation) prenne en compte les aspects suivants :

(i) Ce qui est ou non autorisé, et à quelles fins ;

(ii) Les conditions de la divulgation ; et

(iii) Les conséquences d'un non-respect de ces conditions. » [24]

6.2 Conditions applicables aux FMGD

A priori, les données considérées comme relevant du domaine public peuvent être consultées par toute personne ayant accès au site Web d'un ONS. Il est néanmoins recommandé de préciser les conditions d'utilisation de ces données, ainsi que les mesures de précaution qu'il convient de prendre. Si ces dispositions n'ont pas toujours de force légale, elles permettent néanmoins de sensibiliser l'utilisateur aux questions abordées. Il peut notamment être stipulé dans les « conditions d'utilisation » que toute tentative de recoupement des informations avec d'autres sources est interdite. L'utilisateur devra accepter ces conditions en ligne pour pouvoir télécharger les données voulues.

Un exemple de fichier à grande diffusion est fourni sur le site de Statistique Canada. Il s'agit d'un fichier de microdonnées issu de l'*Enquête conjointe Canada/EtatsUnis sur la santé*[7] (voir également *European Social Survey*[8]). Pour accéder aux fichiers de microdonnées, il suffit aux utilisateurs de s'inscrire. Le but est qu'ils puissent être informés des éventuelles modifications apportées aux fichiers existants et de la mise à disposition de nouveaux fichiers.

La diffusion des fichiers de microdonnées implique le respect de certaines règles et de certains principes. L'Encadré 10 précise les conditions de base applicables aux FMGD.

6.3 Conditions applicables aux fichiers sous licence

Les conditions générales relatives aux **fichiers de microdonnées sous licence** doivent reprendre les grands principes de base, complétés par des dispositions applicables à l'organisme d'appartenance des chercheurs. Deux variantes sont possibles : premièrement, les données sont fournies à un chercheur ou à une équipe de chercheurs à des fins particulières ; deuxièmement, elles sont mises à disposition d'un établissement pour son usage interne sur la base d'un contrat-cadre (organisation internationale ou centre de recherche par exemple). Dans les deux cas de figure, l'organisme d'appartenance du chercheur doit être mentionné et la licence d'utilisation signée par les représentants appropriés.

Accès accordé à un chercheur ou à une équipe de chercheurs à des fins particulières

Quand les données sont mises à disposition en vue d'un projet de recherche précis, l'équipe de chercheurs doit être identifiée. Cela passe par le dépôt d'une demande formelle d'accès aux données (voir modèle présenté à l'Annexe 1). Les conditions d'obtention des données (voir exemple de l'Encadré 12) stipuleront que les fichiers ne doivent en aucun cas être communiqués à des personnes extérieures à l'établissement et doivent être stockés de façon sûre. Dans la mesure du possible, les fins d'utilisation des données seront indiquées – avec la liste des résultats escomptés et la politique de diffusion de l'organisation. L'accès à des ensembles de données sous licence est réservé aux établissements légalement reconnus comme commanditaires (ministère de tutelle, université, centre de recherche, organisation nationale ou internationale).

Contrat-cadre conclu avec un établissement

Dans ce cas de figure, il est convenu que les données sont réservées à l'usage interne de l'établissement, aux fins qu'il jugera utiles, et que les mesures de sécurité qui conviennent doivent être prises. La conformité aux principes en vigueur doit être garantie, une personne

7 Statistique Canada, voir http://www.statcan.gc.ca/pub/8230022x/2003001/4069119-fra.htm, site consulté le 10 janvier 2011.

8 *European Social Survey*, voir http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=78&Itemid=190, site consulté le 22 août 2007 (uniquement en anglais).

nommément citée en assumant formellement la responsabilité. Chaque utilisateur doit être informé des conditions générales applicables aux fichiers de données : une déclaration de confidentialité peut être exigée. En présence d'un tel accord et d'un système de sécurité approprié, la destruction des données après utilisation n'est pas requise. L'Encadré 13 contient un exemple de formulation de contrat-cadre.

Encadré 10 **Conditions d'accès et d'utilisation applicables aux FMGD**

1. Les données et autres documents fournis par l'ONS ne peuvent en aucun cas être revendus ou cédés à d'autres personnes, institutions ou organisations sans l'accord écrit de l'ONS.
2. Les données seront utilisées exclusivement à des fins statistiques ou dans le cadre de travaux de recherche scientifique. Elles serviront uniquement à constituer des données agrégées, y compris à la modélisation d'informations, toute recherche portant sur des particuliers ou des organisations individuelles étant proscrite.
3. Toute tentative de ré-identification des répondants est exclue. Qui plus est, il ne sera fait usage de l'identité d'aucune personne ni organisation découverte fortuitement. Toute découverte de ce type devra être signalée sans délai à l'ONS.
4. Il ne sera fait aucune tentative de recoupement entre des ensembles de données mis à disposition par l'ONS, ni entre des données de l'ONS et d'autres ensembles de données, qui soit susceptible de permettre l'identification de particuliers ou d'organisations.
5. Tous les livres, articles, documents de travail, thèses, dissertations, rapports ou autres publications fondés sur des données obtenues auprès de l'ONS contiendront des références à leur source, conformément à l'exigence de citation associée à l'ensemble de données fourni.
6. Une copie au format électronique de toutes les publications fondées sur les données demandées sera adressée à l'ONS.
7. Le collecteur initial des données, l'ONS et les organismes de financement concernés rejettent toute responsabilité relative à l'utilisation, à l'interprétation ou à des conclusions tirées de ces données.

Remarques :

- Les points 3 et 6 impliquent que les utilisateurs aient la possibilité d'entrer aisément en contact avec le producteur de données. Il est de bon usage d'indiquer un numéro de téléphone, une adresse électronique et, si possible, de mettre à disposition un système de feed-back en ligne.
- Pour le point 5, voir Encadré 11.

6.4 Conditions spécifiques aux centres de données sécurisés

Les **centres de données sécurisés** sont créés pour les données particulièrement sensibles ou très détaillées ne pouvant être suffisamment anonymisées pour permettre une diffusion générale de celles-ci.

On rencontre également les termes « laboratoire de données » ou « centre de données de recherche ». Un centre de données sécurisé peut être aménagé au siège de l'ONS ou sur des sites majeurs tels que des universités proches de la communauté de chercheurs. Ils permettent d'ouvrir aux chercheurs l'accès à des fichiers de données complets sans risque de divulgation d'informations confidentielles. En règle générale, du personnel de l'ONS supervise l'accès et l'utilisation des données, les ordinateurs ne doivent pas être connectés à un réseau externe, et les résultats des chercheurs sont impérativement contrôlés par un analyste de l'ONS en vue de garantir le respect de la confidentialité des données. Un modèle d'une politique d'accès adaptée à un centre de données sécurisé est fourni à l'Annexe 2. L'Annexe 3 décrit en outre le formulaire type de demande d'accès à un centre de données sécurisé.

Les centres de données sécurisés présentent l'avantage d'offrir l'accès à des microdonnées détaillées, mais l'inconvénient d'imposer aux chercheurs un lieu de travail différent du leur.

De plus, les coûts de mise en place et d'exploitation d'un tel centre sont élevés. De nombreux pays ont néanmoins choisi d'ouvrir l'accès aux microdonnées aux chercheurs sur site. Ces chercheurs sont assermentés conformément aux lois sur les statistiques, au même titre que les employés permanents des ONS. Ce système tend à favoriser les chercheurs qui vivent à proximité de l'ONS.

6.5 Gestion des infractions des chercheurs

L'expérience des pays rompus à la pratique de la diffusion de microdonnées montre que les cas de violation de la confidentialité des fichiers de données sont très rares. Ces infractions desservent les intérêts des chercheurs. Elles mettent en péril leur réputation personnelle, ainsi que celle de leur organisme

Au vu de l'importance de disposer d'une politique de diffusion de microdonnées viable, les ONS ont intérêt à prévoir des procédures d'exécution des dispositions correspondantes :

- Il convient de faire cesser immédiatement toute infraction, ce qui est primordial pour préserver la confiance des répondants et la crédibilité de l'établissement.

- En cas d'infraction à la loi, des poursuites judiciaires doivent être envisagées.

- Si les chercheurs ne respectent pas leur engagement, l'ONS peut être amené à suspendre les droits d'accès individuels et de l'organisme.

- Si l'engagement contractuel est pris par une organisation représentée par un chercheur, les sanctions à prendre vis-à-vis de l'un de ses membres seront envisagées par celle-ci plutôt que par l'ONS. La perte d'un droit d'accès individuel peut entraîner la même sanction pour l'ensemble de l'organisation.

Encadré 11 **Comment citer un fichier de données électronique**

Il n'existe pas de normes universelles pour la citation des ensembles de microdonnées. Voici les éléments essentiels que la référence rédigée doit contenir : nom du producteur des données, intitulé de l'ensemble de données et année de référence (suivis d'une indication précisant qu'il s'agit de microdonnées), numéro de référence correspondant (contenant idéalement le numéro de version), nom du distributeur des données et date d'obtention des fichiers (voir également [18] pour plus d'informations sur la citation de produits statistiques).

Exemple 1

U.S. Dept. of Commerce, Bureau of the Census. AMERICAN COMMUNITY SURVEY (ACS) : ECHANTILLON DE MICRODONNEES A GRANDE DIFFUSION (PUMS), 2005 [fichier informatique]. ICPSR04587-v1. Washington, DC : U.S. Dept. of Commerce, Bureau of the Census [producteur], 2005. Ann Arbor, Institut du Massachussets : Consortium interuniversitaire pour la recherche en sciences politiques et sociales (ICPSR) [distributeur], 08-08-2007.

Exemple 2

Fichier de microdonnées à grande diffusion (FMGD) : Enquête sur la dynamique du travail et du revenu (EDTR), vague 8, 2000 : Familles économiques [fichier informatique] / Canada, Statistique Canada, Division des enquêtes-ménages, version 2, Ottawa, Ont. : Statistique Canada [producteur] ; Statistique Canada. Initiative de démocratisation des données [distributeur], 28-08-2003.

Ces références, bien que valables, ne répondent pas à toutes les exigences et ne seront pas satisfaisantes pour tous les centres de données universitaires. « Des URL peuvent être fournis, mais leur validité est souvent éphémère. (…) Les versions révisées des données sont habituellement diffusées sous le même nom sans indication standard du numéro de version. Les réviseurs ne suivent pas de règles fixes, voire aucune règle. Selon les cas de figure, les sources sont citées dans les références bibliographiques, intégrées au texte ou ne sont pas mentionnées du tout. Quoi qu'il en soit, les références contiennent rarement assez d'informations pour garantir l'accès à un ensemble de données identiques dans le futur. » [1] (Traduction de l'anglais)

Pour résoudre ce problème (et d'autres), M. Altman et G. King (Centre de données Harvard-MIT) proposent la solution suivante : « Les références aux données électroniques doivent comprendre six éléments essentiels. Les trois premiers sont classiques ; on retrouve les mêmes pour les documents au format papier. Il s'agit du ou des auteur(s), de la date de publication et de l'intitulé de l'ensemble de données. La présentation de ces éléments doit être adaptée au type d'article ou d'ouvrage dans lequel la référence est citée. L'auteur, la date et le titre permettent de saisir rapidement le type de données dont il est question et quand elles ont été consultées. Cependant, ils ne suffisent pas à identifier de manière univoque un ensemble de données, ni à le localiser, le récupérer ou le vérifier de manière fiable. Par conséquent, nous avons choisi d'ajouter trois autres éléments faisant appel aux technologies modernes. Chacun de ces éléments est conçu de façon à ne pas perdre sa validité en cas d'évolution technologique. Le but est aussi de tirer le meilleur parti du format électronique des données quantitatives.

La quatrième élément est un identifiant unique. Il s'agit d'une désignation synthétique ou d'une chaîne de caractères absolument unique permettant d'identifier durablement l'ensemble de données indépendamment de son lieu de stockage. (…) Les identifiants uniques garantissent la persistance du lien entre la référence et l'objet en question. Il faut toutefois garantir également et pouvoir s'assurer que l'objet en question n'est pas modifié de façon significative en cas de changement de format de stockage. Nous ajoutons donc une signature électronique universelle (*universal numeric fingerprint*, UNF).

L'UNF est une petite chaîne de caractères alphanumériques de longueur fixe résumant le contenu de l'ensemble de données. Toute modification des données, aussi minime soit-elle, entraînerait une nouvelle signature électronique universelle. (…) Le dernier élément d'une référence standard est un gestionnaire de supports. (…) Voici un exemple de référence complète intégrant ces éléments essentiels standard :

Sidney Verba. 1998. *U.S. and Russian Social and Political Participation Data*. hdl:1902.4/00754 ; UNF :4:ZNQRI14053UZq 389x0Bffg?==;http://id.thedata.org/hdl%3A1902.4%2F00754. » [1] (Traduction de l'anglais)

- L'ONS prendra les mesures qui s'imposent pour éviter d'autres infractions ; et,

- Pour les infractions mineures, un avertissement peut suffire.

Informations complémentaires sur la gestion des infractions aux contrats d'utilisation :

> « Il est de bonne règle que ce genre d'engagement s'appuie sur un certain fondement juridique, par exemple qu'il soit incorporé dans une législation habilitante. Ainsi, des mesures juridiques pourraient être prises à l'égard d'auteurs d'infractions. Cela n'empêche pas l'adoption d'autres mesures s'agissant des infractions, telles que le refus de fournir d'autres services au chercheur concerné et/ou éventuellement à l'organisme dont il relève. » [24] (Traduction de l'anglais)

Si l'ONS souhaite recueillir le feed-back des utilisateurs, il convient qu'il mette en place des procédures de suivi régulier des chercheurs. L'occasion peut être saisie pour leur rappeler leur obligation de communiquer leurs résultats et leur demander des suggestions d'amélioration du programme d'enquête.

Encadré 12 **Conditions d'accès et d'utilisation des fichiers de données sous licence**

Remarque : les points 1 à 8 ci-après sont identiques aux conditions applicables aux fichiers à grande diffusion. Les points 9 et 10 sont à adapter pour un contrat-cadre.

1. Les données et autres documents fournis par l'ONS ne peuvent en aucun cas être revendus ou cédés à d'autres personnes, institutions ou organisations sans l'accord écrit de l'ONS.

2. Les données seront utilisées exclusivement à des fins statistiques ou dans le cadre de travaux de recherche scientifique. Elles serviront uniquement à constituer des données agrégées, y compris à la modélisation d'informations, toute recherche portant sur des particuliers ou des organisations individuelles étant proscrite.

3. Toute tentative de ré-identification des répondants est exclue. Qui plus est, il ne sera fait usage de l'identité d'aucune personne ni organisation découverte fortuitement. Toute découverte de ce type devra être signalée sans délai à l'ONS.

4. Il ne sera fait aucune tentative de recoupement entre des ensembles de données mis à disposition par l'ONS, ni entre des données de l'ONS et d'autres ensembles de données, qui soit susceptible de permettre l'identification de particuliers ou d'organisations.

5. Tous les livres, articles, documents de travail, thèses, dissertations, rapports ou autres publications fondés sur des données obtenues auprès de l'ONS contiendront des références à leur source, conformément à l'exigence de citation associée à l'ensemble de données fourni.

6. Une copie au format électronique de toutes les publications fondées sur les données demandées sera adressée à l'ONS.

7. L'ONS et les organismes de financement concernés rejettent toute responsabilité relative à l'utilisation, à l'interprétation ou à des conclusions tirées de ces données.

8. Le nom de l'établissement concerné, du chercheur et des autres chercheurs qui utiliseront les données doivent être indiqués. Le chercheur principal devra signer un contrat de licence au nom de l'établissement. S'il n'est pas habilité à signer au nom de l'établissement récepteur, indiquer le nom du représentant de ce dernier.

9. Les fins d'utilisation auxquelles les données sont destinées doivent être précisées, avec la liste des résultats escomptés et la politique de diffusion de l'établissement.

(Les conditions 8 et 9 peuvent être différentes pour les établissements d'enseignement supérieur)

Encadré13 **Contrat-cadre**

Contrat entre [l'organisme fournisseur] et [l'organisme bénéficiaire] relatif au dépôt et à l'utilisation de microdonnées

A. **Le présent contrat concerne les ensembles de microdonnées suivants:**

1. _____
2. _____
3. _____
4. _____
5. _____

B. **Dispositions du contrat:**

En qualité de propriétaire des droits d'auteur relatifs aux produits mentionnés à la section A, ou en vertu de l'autorisation en bonne et due forme accordée par celui-ci, le représentant de [l'organisme fournisseur] accepte de mettre à disposition de [l'organisme bénéficiaire] les ensembles de données mentionnés à la section A, leur utilisation étant réservée aux employés de [l'organisme bénéficiaire], sous réserve du respect des conditions suivantes:

1. Les microdonnées (sous-ensembles de données compris) et autres éléments protégés par droits d'auteur mis à disposition par [l'organisme fournisseur] ne peuvent en aucun cas être revendus ou cédés à d'autres personnes, institutions ou organisations sans l'accord écrit de [l'organisme fournisseur]. Les éléments non protégés par droits d'auteurs ne contenant pas de microdonnées (questionnaires d'enquête, manuels et dictionnaires de codes ou de données p. ex.) pourront en revanche être communiqués sans son accord préalable. [L'organisme fournisseur] reste propriétaire de l'ensemble des produits mis à disposition par ses soins.
2. Les données seront utilisées exclusivement à des fins statistiques ou dans le cadre de travaux de recherche scientifique. Elles serviront uniquement à constituer des données agrégées, y compris à la modélisation d'informations, toute recherche portant sur des particuliers ou des organisations individuelles étant proscrite.
3. Toute tentative de ré-identification des répondants est exclue. Qui plus est, il ne sera fait usage de l'identité d'aucune personne ni organisation découverte fortuitement. Toute découverte de ce type devra être signalée sans délai à [l'organisme fournisseur].
4. Il ne sera fait aucune tentative de recoupement entre des ensembles de données mis à disposition par [l'organisme fournisseur], ni entre des données de [l'organisme fournisseur] et d'autres ensembles de données, qui soit susceptible de permettre l'identification de particuliers ou d'organisations.
5. Tous les livres, articles, documents de travail, thèses, dissertations, rapports ou autres publications fondés sur des données obtenues auprès de l'ONS contiendront des références à leur source, conformément à l'exigence de citation associée à l'ensemble de données fourni.
6. Une copie au format électronique de toutes les publications fondées sur les données demandées sera adressée à [l'organisme fournisseur].
7. [L'organisme fournisseur] et les organismes de financement concernés rejettent toute responsabilité relative à l'utilisation, à l'interprétation ou à des conclusions tirées de ces données.
8. Les données seront stockées dans un environnement sûr et l'accès à celles-ci restreint de façon appropriée. [L'organisme fournisseur] se réserve le droit de demander à tout moment des informations sur les modalités de stockage et sur les systèmes de diffusion en place.
9. [L'organisme bénéficiaire] s'engage à présenter à [l'organisme fournisseur] un rapport annuel sur l'utilisation et sur les utilisateurs des ensembles de données susmentionnés, précisant le nombre de chercheurs ayant accès à chaque ensemble de données et les résultats des travaux de recherche effectués.
10. L'accès aux données est accordé pour une période de [période déterminée ou contrat à durée illimitée].

C. **Correspondance:**

[L'organisme bénéficiaire] désignera un interlocuteur unique pour le présent contrat. Si cette personne venait à être remplacée, [l'organisme bénéficiaire] s'engage à communiquer sans délai à [l'organisme fournisseur] le nom et les coordonnées d'un nouvel interlocuteur. Toute correspondance administrative ou relative à une procédure particulière peut être envoyée par e-mail, par fax ou par courrier postal en indiquant les informations suivantes :

Correspondances envoyées par [l'organisme fournisseur] à [l'organisme bénéficiaire]:

Nom de la personne à contacter: _____
Fonction de la personne à contacter: _____
Adresse du destinataire: _____
_____
Email: _____
Fax: _____

Correspondances envoyées par [l'organisme bénéficiaire] au [dépositaire] :

Nom de la personne à contacter: _____
Fonction de la personne à contacter: _____
Adresse du destinataire: _____
_____
Email: _____
Fax: _____

D. **Signataires**

Les signataires ci-dessous ont lu et approuvé les dispositions du présent contrat :

Représentant de [l'organisme fournisseur]

Nom _____

Signature _____ Date _____

Représentant de [l'organisme bénéficiaire]

Nom _____

Signature _____ Date _____

7. Qu'entend-on par « anonymisation » des microdonnées ?

Ce chapitre se fonde très largement sur le manuel publié par le CENEX sur le contrôle de la divulgation statistique [3], qui peut être téléchargé gratuitement à l'adresse suivante : http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf. (en anglais uniquement)9

La diffusion de fichiers de microdonnées au public, aux chercheurs ou à d'autres établissements est une mission délicate pour les ONS. Il s'agit, d'une part, de mettre à disposition des fichiers de microdonnées appuyant un grand nombre d'analyses statistiques et, d'autre part, de préserver l'identité des répondants. Les processus relatifs à ce deuxième aspect sont communément désignés par la notion de contrôle de la divulgation statistique (CDS) ou d'anonymisation.

« La confidentialité ne pouvant pas être garantie à 100 %, les risques et les avantages liés à l'accès aux données doivent être pesés en fonction des risques de divulgation et de la sensibilité des données. » [14] (Traduction de l'anglais) La protection des données est une question parfaitement légitime pour les offices de statistique et les autres producteurs de données. Il existe d'ailleurs une abondante littérature sur ce sujet. La littérature sur la gestion de la confidentialité des données contient également de nombreuses références à ce que Peter Madsen [13] a appelé le « paradoxe de la confidentialité ». En 2003, lors d'un atelier consacré à l'étude de la confidentialité, organisé sous l'égide de la *National Science Foundation* des Etats-Unis, Peter Madsen a fait valoir que « le zèle déployé pour sécuriser entièrement les données dans le contexte de la recherche aboutit paradoxalement à un moindre avantage pour la société au lieu de le maximiser[10] ».

7.1 Concepts liés au contrôle de la divulgation statistique (CDS)

Il y a divulgation quand une personne ou une organisation constate ou apprend une information inédite sur une autre personne ou un autre établissement. On distingue deux types de risque de divulgation : la **divulgation d'identité** et la **divulgation d'attributs[11]**. Il y a divulgation d'identité quand cette dernière est directement associée à un enregistrement du fichier de données diffusé. Cela arrive quand l'enregistrement en question contient des variables permettant d'identifier sans ambiguïté un répondant (nom, numéro de passeport / d'identification ou coordonnées téléphoniques p. ex.). Il est essentiel de supprimer ces variables de tout fichier de microdonnées avant sa diffusion. Il y divulgation d'attributs quand des valeurs (précises ou estimées) présentes dans les données diffusées peuvent être associées à une personne en particulier.

On appelle « clé » (ou variable clé) une combinaison de variables d'un enregistrement de microdonnées pouvant permettre de remonter jusqu'à un répondant. Cette « réidentification » est possible quand un répondant (a) appartient à une classe de population rare pour une valeur clé donnée ; et quand (b) la clé peut servir à recouper un fichier de microdonnées avec d'autres fichiers de données susceptibles de contenir des identifiants directs ou autres (listes de votants, registres fonciers ou dossiers scolaires), voire en utilisant des moteurs de recherche publics sur Internet).

Dans la majorité des pays en développement, le risque de divulgation lié à la possibilité de recouper un fichier de microdonnées d'enquête avec d'autres fichiers de données est actuellement limité, soit parce qu'ils n'existent pas encore, soit parce qu'ils ne sont pas largement diffusés. En réalité, le risque de divulgation inhérent à la diffusion de fichiers de microdonnées d'enquête peut être minimisé de façon satisfaisante, tout simplement en supprimant les identifiants directs des enregistrements, ainsi que les renseignements géographiques trop précis par rapport à la strate d'échantillonnage. Il convient toutefois d'évaluer le risque de divulgation au cas par cas. Il existe en effet des exceptions notables à la règle générale susmentionnée, qui justifient des mesures de contrôle supplémentaires. Les fichiers de microdonnées d'enquête contenant des enregistrements relatifs à des groupes cibles restreints – une entreprise par exemple – sont beaucoup plus difficiles à anonymiser.

Enfin, autre considération importante pour les ONS : la nécessité de protéger leurs bases d'échantillonnage –notamment la base d'échantillonnage intercensitaire

9 Le site Web de l'*American Statistical Society,* régulièrement mis à jour, contient également des informations et des documents sur le contrôle de la divulgation statistique. Voir www.amstat.org/committees/cmtepc/index.cfm?fuse action=main (en anglais uniquement).

10 Un résumé des résultats de cet atelier de travail est fourni sur http:// www.nsf.gov/sbe/ses/mms/nsfworkshop_summary1.pdf. (en anglais uniquement)

11 Terminologie proposée par D. Lambert [12].

générale utilisée pour la conception des différentes enquêtes réalisées auprès des ménages. La diffusion des bases d'échantillonnage détaillées est incompatible avec les mesures visant à protéger l'identité des répondants. Ces informations non codées peuvent servir à faire des recoupements – ainsi, il est possible de déterminer des coordonnées géographiques (nom et localisation d'UPE) en recoupant les facteurs de pondération et d'échantillonnage (qui doivent apparaître dans les fichiers de microdonnées pour que les utilisateurs puissent procéder à des déductions statistiques justes). La protection des bases d'échantillonnage s'explique en particulier par la volonté de limiter la probabilité que les répondants d'une unité primaire d'échantillonnage (UPE) soient sélectionnés pour d'autres enquêtes non réalisées par l'ONS. Dans les pays où les UPE de base ne sont mises à jour que tous les 10 ans, cela peut se traduire par une augmentation du taux de non-réponse (répondants lassés) et être source de distorsions.

7.2 Scénarios de divulgation

La première étape du processus d'anonymisation d'un fichier de microdonnées en vue de sa diffusion consiste à identifier les parties du fichier qui présentent un risque de divulgation. Par conséquent, avant de mettre en œuvre des mesures de CDS, il est intéressant d'examiner les différents cas de figure dans lesquels l'utilisateur d'un fichier de microdonnées est susceptible d'identifier un répondant. On en distingue principalement deux :

- **Connaissance de la réponse :** l'utilisateur dispose, à titre personnel, de suffisamment d'informations sur les attributs associés à un ou plusieurs enregistrements. Autrement dit, il appartient au cercle de connaissances d'une unité statistique. Plus le groupe cible est restreint (enquêtes auprès d'une entreprise ou d'un organisme, ou enquêtes auprès des ménages dans de petits pays ou dans des régions à faible densité de population p. ex.).

- **Recoupement avec des archives externes :** l'utilisateur associe certains enregistrements du fichier de microdonnées diffusé à d'autres ensembles de données (ou registres) publics contenant des identifiants directs, malgré l'interdiction exprès stipulée dans le contrat d'utilisation des données. Pour ce faire, il utilise les variables clés disponibles dans les deux ensembles de données pour former des clés d'identification (rapprochement de données).

Avant d'anonymiser des données, il convient de définir un scénario de divulgation décrivant les informations éventuellement communiquées, ainsi que les méthodes potentielles d'identification de particuliers. Ce scénario est souvent fondé sur des hypothèses prudentes, afin d'envisager le pire cas de figure possible. Il est parfois nécessaire de définir plusieurs scénarios, car différentes sources d'informations peuvent être proposées à l'utilisateur, simultanément ou alternativement.

7.3 Evaluation du risque de divulgation

La confidentialité est rompue en cas de ré-identification d'un répondant et quand le contrevenant (un utilisateur non autorisé ou ne respectant pas les conditions stipulées dans le contrat d'utilisation et d'accès aux données) arrive à associer des variables sensibles à un individu.

Avant de diffuser des microdonnées, les ONS doivent évaluer les données « afin de déterminer si une publication mettrait en péril l'identité de certains particuliers ou établissements. Plusieurs facteurs sont pris en compte dans ce cadre : 1) le niveau de détails présenté par les données à diffuser (notamment précisions géographiques et variables réputées communes à des sources de données externes, susceptibles de servir de clés et augmentant ainsi le risque d'identification) ; 2) certaines variables ou combinaisons de variables permettant d'isoler des répondants au sein de l'échantillon et pouvant favoriser leur identification par des personnes extérieures ; et 3) l'existence d'autres données accessibles à l'extérieur de l'[ONS], telles que celles déjà publiées sur la même enquête ou sur une enquête associée, ou encore les informations détenues par d'autres établissements sur le même répondant. » [14] (Traduction de l'anglais)

« La détermination du risque de divulgation de données identifiables est une tâche complexe qui nécessite à la fois une analyse statistique empirique et du discernement. » [14] Il existe plusieurs méthodes, qui présentent toutes des atouts. Aucune d'entre elles ne s'est cependant imposée comme la « meilleure ». Certes, le risque de divulgation ne peut pas être entièrement éliminé, mais des mesures sont envisageables pour le compenser en réduisant l'utilité des données. Ces mesures impliquent l'établissement d'un seuil de risque (*threshold rule* en anglais) permettant de déterminer si la diffusion de l'ensemble de données est sûre ou non. Il existe essentiellement deux méthodes de mesure mathématique du risque de ré-identification :

- **Mesures individuelles :** évaluation du risque pour chaque enregistrement. Ces mesures sont généralement exprimées sous forme de probabilité de ré-identification d'un répondant ou en termes d'unicité et de rareté au sein de l'échantillon.

- **Mesures générales :** évaluation du risque pour l'ensemble du fichier. Ces mesures aboutissent à un nombre estimé de ré-identifications et peuvent être dérivées de mesures individuelles agrégées.

L'avantage des mesures individuelles réside dans le fait que seuls les enregistrements apparaissant comme sensibles au regard du seuil de risque fixé doivent être protégés. Cela limite la perte d'informations et d'utilité des données. Se limiter à une mesure générale du risque de ré-identification peut inciter à recourir à des techniques de CDS pour chaque enregistrement du fichier de microdonnées. La perte d'informations risque alors d'être plus importante et l'intérêt des données pour l'analyse statistique moindre.

Les méthodes de mesure du risque de ré-identification se distinguent également par l'utilisation des clés. Les mesures basées sur des clés de l'échantillon permettent d'identifier des combinaisons uniques ou rares de variables qualitatives (clés) au sein de l'échantillon. Une unité est considérée comme sensible quand sa combinaison de scores pour les variables d'identification est inférieure au seuil fixé. Par exemple, il y a fort à parier qu'une personne de sexe masculin âgée de 30 ans, médecin de profession et père de quatre enfants de sexe féminin sera unique au sein de l'échantillon d'enquête. L'enregistrement correspondant sera donc considéré comme sensible, même si la population totale compte probablement beaucoup d'autres individus présentant les mêmes caractéristiques. Le risque pour une unité donnée peut être déterminé aussi par la combinaison de scores pour les variables d'identification qualitatives au sein de la population ou par sa probabilité de ré-identification. La fréquence d'apparition au sein de la population étant généralement inconnue, ces probabilités doivent être établies par modélisation. Avec des variables d'identification continues, il est impossible d'exploiter le critère de rareté des clés, dont la majorité, sinon toutes, sont généralement uniques. Le risque de divulgation via des variables continues est estimé en évaluant la probabilité de ré-identification par rapprochement de variables de deux ensembles de données distincts, sur la base de la « proximité » des valeurs correspondantes.

Plusieurs modes d'identification de cas et de variables dans un fichier peuvent conduire à la divulgation d'informations. L'une des méthodes courantes consiste à générer des distributions de fréquence et des tableaux multidimensionnels, afin d'identifier les cellules correspondant à un faible nombre de cas. Les données géographiques détaillées comptent parmi les principales sources d'identification de particuliers, notamment par des utilisateurs de la même région et qui connaissent bien les caractéristiques de certains répondants.

Encadré 14 **Liste de points à vérifier en vue d'évaluer les différents cas de figure et les risques de divulgation**

En vue de l'examen préalable à la diffusion de fichiers de microdonnées d'enquête et de recensement, pour lesquels le *Census Bureau* américain est tenu de protéger la confidentialité des répondants, une liste de points à vérifier est disponible (*Checklist on Disclosure Potential of Proposed Data Releases*, www.census.gov/srd/sdc/). Cette liste facilite la procédure de contrôle de divulgation. Elle est destinée à aider les personnes chargées de ce contrôle à déterminer s'il faut opter pour la mise à disposition de fichiers de microdonnées à grande diffusion ou de données tabulaires. La section 3 de cette liste porte sur les fichiers de microdonnées. Les points à vérifier concernent des éléments importants tels que les coordonnées géographiques, les variables présentant un risque inhabituel de divulgation, les informations contextuelles et écologiques, l'éventualité d'un rapprochement ou d'une comparaison avec d'autres données et .

7.4 Techniques de contrôle de divulgation spécifiques aux fichiers de microdonnées

La première grande étape du CDS d'un fichier de microdonnées est la suppression de tous les identifiants directs – variables permettant d'identifier un répondant sans équivoque. Ensuite, l'anonymisation du fichier de microdonnées se poursuit. Elle fait appel à des **méthodes de « masquage »** ou passe par la création de **microdonnées synthétiques**. Un fichier de microdonnées synthétique est créé aléatoirement via un processus permettant de conserver certaines corrélations statistiques ou internes au fichier de microdonnées brutes initial.

Les méthodes de masquage désignent des techniques permettant de créer une version modifiée du fichier de microdonnées brutes initial. On distingue deux types de méthodes de CDS pas masquage. Les **méthodes basées sur la perturbation de l'information**

consistent à modifier les données avant leur publication en y introduisant délibérément un élément erroné pour des raisons de confidentialité. **Les méthodes basées sur la restriction d'informations** consistent à réduire le contenu informatif des données fournies en en supprimant une partie ou par agrégation.

Vous trouverez dans ce qui suit une présentation générale des techniques de CDS les plus répandues, avec le type de données auxquelles elles s'appliquent : variables continues ou qualitatives, ou les deux. Une variable est dite continue lorsqu'elle est numérique et peut permettre d'effectuer des opérations arithmétiques (revenu, âge et taille du ménage p. ex.). Les variables définies uniquement pour un ensemble fini et ne pouvant pas servir à des calculs mathématiques sont dites qualitatives ou catégorielles (variables classées sur une échelle ordinale, telles que le plus haut niveau de scolarité atteint, ou variables mesurées sur une échelle nominale, telles que le statut marital, où l'ordre des valeurs est insignifiant).

Techniques basées sur la restriction d'informations

Les techniques basées sur la restriction des données fournies ou de leur contenu informatif consistent à modifier les fichiers de microdonnées dans le but d'éliminer les variables ou les enregistrements susceptibles d'être associé(e)s sans équivoque à un individu en particulier. L'autre solution consiste à créer des catégories, de façon à accroître le nombre potentiel de répondants pour une catégorie donnée. L'ONS peut ainsi décider de fixer un seuil imposant un nombre minimum de réponses par cellule. Voici les six grandes techniques de masquage basées sur la restriction d'informations :

1. **Echantillonnage** : remplace la diffusion de l'intégralité du fichier de microdonnées ; recommandé pour la publication des données de recensement de population. A condition que l'échantillon soit suffisamment restreint (5 % de la population p. ex.) et de supprimer toutes les variables d'identification directes, cette technique permet de réduire le risque de divulgation de façon satisfaisante pour les données qualitatives. Toutefois, si le fichier de microdonnées contient des variables continues, susceptibles d'être recoupées plus facilement avec un fichier de données accessible ailleurs, la mise en œuvre de techniques de CDS supplémentaires s'impose.

2. **Recodage global :** consiste à regrouper certaines valeurs en fonction de classifications prédéfinies, de façon à ce que les réponses individuelles n'apparaissent plus. Cette méthode se prête aussi bien aux variables continues ou discrètes qu'aux codes géographiques. A titre d'exemple, l'âge peut être tronqué et remplacé par des classes d'âge, la profession et les codes industriels pouvant se fondre dans de vastes catégories. De même, les renseignements géographiques plus précis que la strate d'échantillonnage peuvent être supprimés. La méthode du recodage global est adaptée à la fois aux données continues et qualitatives.

3. **Regroupement des valeurs extrêmes supérieures et inférieures :** opération indiquée en présence de variables numériques ou ordinales dont les valeurs supérieures et inférieures sont extrêmement rares et susceptibles de révéler l'identité des répondants. Le regroupement des valeurs extrêmes supérieures et inférieures consiste à créer des catégories « grossières » (pour les valeurs supérieures, personnes âgées de plus de x ans ou revenu supérieur à y, p. ex.). Cette technique se prête aux variables continues comme aux variables qualitatives mesurées sur des échelles ordinales, pour lesquelles une logique de classement peut être définie.

4. **Suppression locale** : technique très simple utilisée quand le rapprochement de deux variables individuelles peut conduire à l'identification d'une personne bien précise. Autrement dit, une combinaison des variables peut fournir une clé de ré-identification pour un enregistrement donné. Prenons un enregistrement contenant les informations suivantes : âge = 85 et situation scolaire = actuellement inscrit en école primaire. Un tel enregistrement sera probablement unique, dans la mesure où rares sont les octogénaires inscrits à l'école primaire[12]. Dans ce cas précis, la suppression de l'une des deux informations (situation scolaire ou âge) suffirait sans doute à éliminer le problème. La suppression locale est utile surtout pour les données qualitatives.

5. **Suppression de certaines variables** : méthode requise pour les informations considérées comme impropres à la diffusion parce que trop sensibles, (appartenance ethnique ou religion p. ex.).

12 Voir p. ex. http://news.bbc.co.uk/2/hi/africa/4244520.stm

6. **Suppression de certains enregistrements**: parfois nécessaire pour préserver l'anonymat des répondants avec un ensemble de variables unique. Quand un enregistrement est entièrement supprimé du fichier de microdonnées, il est impératif de calculer et de prévoir des facteurs de pondération ajustés. Cette méthode doit être utilisée avec parcimonie, car la suppression d'un nombre significatif d'enregistrements entraîne une distorsion des données.

Techniques de perturbation

Les techniques de perturbation consistent à modifier les données de façon à les rendre moins précises et plus difficiles à recouper. En cas de tentative de ré-identification, les valeurs ainsi modifiées rendent la corrélation incertaine. Voici un bref descriptif des sept grandes techniques de perturbation :

1. **Ajout de bruit (ou perturbation aléatoire)** : technique consistant à ajouter des valeurs aléatoires à celles transmises par un répondant. Les méthodes varient selon que l'on ajoute du bruit à une ou à plusieurs variables ou de façon à préserver les moyennes, variances et covariances. Des techniques de programmation linéaires permettent en outre de minimiser les écarts entre les valeurs modifiées et les valeurs réelles.

2. **Permutation de données** : méthode consistant à modifier un fichier de microdonnées en remplaçant les valeurs correspondant à des variables confidentielles par celles d'un autre enregistrement. Les enregistrements sont ainsi « permutés » de façon à conserver les valeurs marginales ou les cellules correspondant à un faible nombre de cas. Cette technique peut être utilisée pour des variables continues ou qualitatives.

3. **Permutation de rang :** système permettant de trier les variables à protéger par ordre croissant, puis d'en regrouper certaines. Des paires d'enregistrements sont sélectionnées de façon aléatoire dans chaque groupe et les valeurs correspondantes permutées avec celles d'autres paires au sein d'un rang prédéfini. La création de groupes de différentes tailles conduit à différents modèles de données.

4. **Micro-agrégation** : consiste à remplacer une valeur observée dans l'échantillon par la moyenne calculée pour un groupe restreint d'unités (petit agrégat ou micro-agrégat), dont celui qui est contrôlé. Les unités d'un même groupe sont représentées par la même valeur dans le fichier diffusé. Chaque groupe contient un nombre minimum d'unités prédéfini (k). La valeur minimale admissible de k est 3. Pour une valeur k donnée, le but est de scinder l'ensemble des unités en groupes d'au moins k unités en réduisant au minimum la perte d'informations, généralement exprimée sous forme de perte de variabilité. Les unités sont donc regroupées en fonction du nombre de similitudes qu'elles présentent. Le mécanisme de micro-agrégation protège les données en faisant en sorte que le fichier contienne au moins k unités associées à la même valeur. Cette technique est parfois appliquée à des variables continues.

5. **Arrondi** : diverses méthodes d'arrondi sont utilisées selon le but recherché : arrondi contrôlé visant à conserver les totaux et certaines propriétés générales ou arrondi aléatoire destiné à faire en sorte que le nombre de cas apparaissant dans la cellule d'un tableau de données agrégées ne trahisse pas le nombre de cas d'une ou deux observations.

6. **Ré-échantillonnage**: consiste à tirer différents échantillons indépendants des valeurs relatives aux variables à masquer. Le critère de classification retenu est le même. Les variables masquées sont créées comme suit : la première valeur est la moyenne de toutes les premières valeurs de chaque échantillon, etc.

7. **Post-randomisation (ou randomisation *a posteriori*) :** version randomisée de la permutation de données. Cette technique ajoute une composante incertaine aux valeurs relatives à certaines variables en les remplaçant suivant un mécanisme probabiliste. Comme dans le cas de la permutation, la protection des données est assurée, car les utilisateurs sont dans l'impossibilité d'établir avec certitude l'exactitude d'une valeur diffusée. Par conséquent, toute tentative de recoupement de l'enregistrement avec des identifiants externes peut facilement conduire à des erreurs de classification ou d'interprétation. Cette méthode est employée essentiellement pour des variables qualitatives, mais peut également être appliquées pour des variables numériques continues.

Fichiers de microdonnées synthétiques

Les fichiers de microdonnées synthétiques sont créés selon un processus aléatoire, la contrainte à respecter étant que les corrélations statistiques ou internes au fichier initial soient préservées. Il est tentant de fournir des fichiers entièrement constitués à l'aide de simulations et ne présentant aucun risque de divulgation à la place des fichiers de microdonnées initiaux. Plusieurs techniques ont d'ailleurs été mises au point à cet effet. Cependant, par rapport aux méthodes de masquage des données, les différentes techniques de création de fichiers de microdonnées synthétiques sont toutes extrêmement laborieuses et complexes à mettre en application. Les données d'enquête comprennent généralement des centaines de variables, dont la distribution et les corrélations ne sont pas simples à modéliser avec des outils de paramétrage standard. Les travaux de recherche portant sur l'amélioration et la mise en œuvre de ces techniques se poursuivent.

7.5 Trouver le bon compromis entre risque de divulgation et perte d'informations

La modification d'un fichier de microdonnées via des techniques de contrôle de divulgation statistique se traduit par une perte d'informations. Il appartient aux ONS de trouver le juste équilibre entre cette perte d'informations et le risque de divulgation. Tout comme le risque de divulgation, la perte d'informations liée aux différentes techniques de CDS peut faire l'objet d'une évaluation par l'ONS. Pour ce qui est des données qualitatives, elle peut être évaluée à l'aide de méthodes telles que la comparaison directe, la comparaison de tableaux de contingence ou la mesure entropique. Dans le cas de variables continues, elle peut être mesurée par comparaison des moyennes quadratiques, des moyennes absolues ou de la variation des moyennes[13]. Le potentiel d'utilisation des fichiers de microdonnées est cependant si étendu qu'il est pour ainsi dire impossible d'estimer avec précision la perte d'informations. La meilleure solution consiste à identifier la catégorie d'utilisateurs la plus affectée par le recours aux mesures de CDS. Il s'agit habituellement des chercheurs habitués à mener des analyses statistiques approfondies basées sur des fichiers de microdonnées. Cette catégorie d'utilisateurs est généralement restreinte et ses travaux peuvent contribuer de façon significative et importante à l'intérêt général. Cela souligne la nécessité de ne pas se contenter de diffuser des FMGD et d'y adjoindre des fichiers sous licence anonymisés dans une moindre mesure. Par ailleurs les ONS devraient se faire une idée assez précise de la perte d'informations liée au CDS en examinant les formulaires de demande d'accès aux fichiers sous licence, où figurent les raisons pour lesquelles la version à grande diffusion d'un fichier de microdonnées ne sera pas utile dans le cadre des projets de recherche concernés.

7.6 Documentation du processus de divulgation des données statistiques

Les méthodes de CDS employées doivent permettre d'équilibrer la perte d'informations et la probabilité de divulgation d'informations personnelles. Il convient que les utilisateurs soient informés, le cas échéant, du fait que le risque de divulgation de l'ensemble de données a été évalué, ainsi que des méthodes de protection appliquées. Il est également souhaitable qu'ils connaissent la nature et l'étendue de toute modification liée au contrôle de la divulgation.

La ou les technique(s) de CDS mise(s) en œuvre peut ou peuvent être précisée(s), mais le niveau de détails disponible ne devrait pas permettre à l'utilisateur de reconstituer les fichiers de microdonnées initiaux de quelque manière que ce soit.

13 Consulter le site Web de l'IHSN (www.ihsn.org) pour obtenir des sources sur les techniques d'évaluation de perte d'informations.

Encadré 15 **Documentation du CDS par le Census Bureau américain**
Mesures appliquées aux échantillons de microdonnées à grande diffusion du recensement 2000

« La confidentialité sera protégée via les processus suivants: permutation de données, regroupement des valeurs extrêmes supérieures, seuils géographiques, perturbation des données relatives à l'âge pour les ménages de grande taille, et restriction des détails fournis pour certaines variables qualitatives.

La *permutation de données* est une méthode de limitation de la divulgation d'informations conçue pour protéger la confidentialité des données figurant dans des tableaux de fréquence (nombre de personnes ou pourcentage de la population présentant certaines caractéristiques). Elle consiste à modifier les données initiales ou à permuter des enregistrements concernant un échantillon de cas. La permutation concerne des enregistrements individuels et protège donc également les microdonnées.

Le *regroupement des valeurs extrêmes supérieures* désigne une méthode de limitation de la divulgation d'informations consistant à classer tous les cas correspondant ou supérieurs à une distribution de fréquence donnée dans une seule et même catégorie.

Les *seuils géographiques* empêchent la divulgation des données relatives à des particuliers ou à des ménages associés à des unités géographiques avec des dénombrements de population inférieurs à un certain niveau (voir description des unités de microdonnées à grande diffusion (PUMA) et des super-PUMA fournie à la section III).

La *perturbation des données* relatives à l'âge, autrement dit la modification de l'âge des membres d'un foyer, est requise pour les ménages de grande taille (10 personnes ou plus), pour des raisons de confidentialité.

Le *détail des variables qualitatives* sera tronqué si la catégorie concernée ne respecte pas le seuil de population national minimum fixé. »

Source: http://www.census.gov/population/www/cen2000/pums/index.html, site consulté le 10 août 2010 (traduction de l'anglais).

8. L'accès aux microdonnées doit-il être payant ou gratuit ?

Les producteurs de données peuvent envisager la vente de microdonnées comme un moyen de rentrer dans leurs frais de production - qui ne sont pas toujours budgétés. Les offices de statistique s'étant émancipés, ils sont souvent contraints de trouver et de mettre en place des activités rémunératrices. A cet égard, les microdonnées constituent un produit très précieux.

Si la mise à disposition des fichiers de microdonnées augmente la valeur d'une enquête, elle induit également des coûts liés à la documentation et à l'anonymisation. Quand la production de microdonnées à diffuser n'est pas intégrée au budget de l'enquête, l'ONS n'a pas d'autre choix que de tenter de récupérer sa mise.

Un service de diffusion de microdonnées complet génère également d'autres coûts, liés notamment au contrôle des fichiers préalable à la diffusion, à l'octroi des licences d'utilisation, à l'exploitation d'un centre de données sécurisé, à l'assistance fournie aux utilisateurs et à la gestion de l'infrastructure. En règle générale, ces tâches requièrent du personnel dédié et une infrastructure spéciale. Si l'établissement ne dispose pas du budget nécessaire ou s'il ne peut pas assumer la responsabilité de ces tâches en l'état, il sera probablement difficile d'assurer un service viable de qualité.

La vente des données est une façon de faire porter aux bénéficiaires du service de données une partie des coûts à supporter.

8.1 Exemple de deux pays

Le recouvrement des coûts engagés par les ONS est un fait de longue date partout dans le monde. Il s'est généralisé dans les années 1980, souvent en réponse aux coupes budgétaires et aux pressions des organismes de tutelle en faveur d'un transfert du contribuable à l'utilisateur des coûts liés aux statistiques. L'examen complet des activités des ONS axées sur le recouvrement des coûts n'est pas l'objet de ce document. Voici néanmoins quelques remarques :

- L'institut de statistique néo-zélandais (SNZ) a tenté de récupérer 25 % de son budget total à travers la vente de divers produits et services. Cette décision fut prise pendant une période de rigueur administrative dans le pays. L'initiative de SNZ a permis d'éviter la suppression de certains programmes. L'établissement a toutefois échoué à atteindre son objectif et l'initiative a finalement été jugée impropre à le réaliser. Renonçant à rentrer dans ses frais, NSZ a donc entamé une stratégie de réduction des coûts. Selon le site Web de l'établissement, près de 90 % des informations recueillies par NSZ seraient accessibles gratuitement. Des frais sont prélevés pour les totalisations personnalisées et les tableaux détaillés. L'accès aux fichiers de microdonnées est payant, mais les tarifs ne sont pas précisés sur le site Web[14].

- Statistique Canada a mis en place un programme complet de recouvrement des coûts dans les années 1980 pour faire face à la pression budgétaire et politique. L'établissement s'est attelé à la tâche malgré de sévères critiques des utilisateurs. Ce programme a eu un succès mitigé et a été en grande partie supprimé, ou pour le moins considérablement réduit. Le prix de l'accès aux microdonnées a considérablement augmenté, poussant les chercheurs canadiens à utiliser les données recueillies aux Etats-Unis, facilement accessibles auprès du Consortium interuniversitaire de recherche en sciences politiques et sociales (ICPSR). L'Initiative de démocratisation des données (IDD) a été lancée pour pallier à ce problème, qui desservait les intérêts canadiens. L'analyse suivante a servi d'impulsion à l'IDD :

  > « ...le véritable exercice de la démocratie implique la possibilité pour les citoyens d'accéder à des informations complexes et d'acquérir les compétences requises pour comprendre ces informations. » Si Paul Bernard est conscient des pressions exercées sur Statistique Canada pour que l'organisme réduise ses coûts et augmente ses revenus, il a toutefois le sentiment qu'elles se sont traduites par un « accès aux données réservé aux catégories d'utilisateurs qui ont les moyens de payer ». Selon Paul Bernard, cette situation

14 Voir http://www.stats.govt.nz/about-us/making-more-informationfree/default.htm, site consulté le 18 octobre 2007 (en anglais uniquement)

> risque fort « ...d'entraver la participation au débat public des catégories disposant de ressources limitées, [tout comme de] celles qui ont peu de chance de dégager un profit ou de tirer un bénéfice tangible et relativement immédiat de l'utilisation des données. » Il ajoute qu'elle « ... est susceptible, à long terme, d'aboutir à un développement sous-optimal et à un semblant de démocratie. » [31] (Traduction de l'anglais)

Un réseau de 74 établissements d'enseignement donnant accès à toute la collection de données publiques de Statistique Canada fut ainsi créé. Cette collection comprend environ 300 FMGD et plusieurs milliers d'autres fichiers, bases de données et fichiers géographiques.

Les frais d'abonnement couvrent les coûts d'assistance ainsi que l'élaboration d'une infrastructure technique optimisée pour les abonnés et pour l'organisme. Ils n'ont pas vocation à couvrir les coûts de production des données.

Le portail d'accès aux ressources de l'IDD n'est pas ouvert aux ministères, dont beaucoup ont signé des accords avec Statistique Canada portant sur la participation au financement de certaines enquêtes. De manière générale, ces accords englobent l'accès aux microdonnées. Le service proposé ne répond donc pas à un besoin des ministères.

8.2 Accès payant ou gratuit ?

Il n'y a pas de réponse absolue à cette question. De nombreux arguments plaident en faveur d'une minimisation du prix – pour ne pas dire de la gratuité totale – de l'accès aux fichiers de microdonnées.

Accès gratuit

Le principal argument en faveur de la gratuité est que de nombreux ONS sont entrain de renforcer leurs pratiques de diffusion de fichiers de microdonnées. Ces établissements craignent que des tarifs trop élevés rebutent les utilisateurs.

- L'accès payant réduit considérablement le nombre d'utilisateurs potentiels et, par là-même, la valeur des données.
- Dans les pays en développement, il peut être un obstacle pour les principaux intéressés : étudiants, centres de recherche locaux, universités, etc.

L'autre argument en faveur de la gratuité est le coût lié à la perception des droits d'accès. Par ailleurs, selon que ces droits vont à l'ONS ou à un organisme de tutelle, la motivation du personnel à les percevoir ne sera pas la même : pour fonctionner correctement, le système doit être efficace.

- Un accès payant crée une exigence de qualité et de service.
- Il est peu rémunérateur : la majorité des demandes émanent de la communauté universitaire, dont les ressources sont limitées et qui peut aisément aller ailleurs et utiliser d'autres données.

L'expérience de certains pays – principalement des pays industrialisés – montrent que le recouvrement des frais engagés est possible jusqu'à un certain point. Il paraît toutefois difficile de récupérer tous les coûts supplémentaires liés à la diffusion des microdonnées. Il peut également être démontré qu'une stratégie de recouvrement des coûts agressive ne favorise pas l'utilisation des fichiers de microdonnées et, à long terme, diminue la valeur potentielle d'une enquête.

La solution idéale pour l'ONS est que tous les coûts soient inscrits au budget de l'enquête, ce qui maximise l'accessibilité des données. Ces coûts peuvent être supportés par les commanditaires, ce qui maximise les bénéfices de l'enquête. Cet aspect est déterminant dans les pays où les chercheurs disposent de moyens limités et où les producteurs ont peu de ressources à affecter à l'analyse de données.

Accès payant

Il est fort probable qu'un système d'accès payant génère des revenus. D'autres facteurs doivent cependant être pris en compte avant de prendre ou non la décision de mettre en place un tel système, notamment :

- L'ONS est-il légalement en droit de facturer des droits d'accès à ses produits ?
- Quels coûts l'ONS souhaite-t-il récupérer ?
- Ces coûts sont-ils clairement identifiables : serontils compris et acceptés par les utilisateurs ?

- Les utilisateurs ont-ils les moyens de payer ? Est-il envisageable qu'une association d'utilisateurs soit créée pour recouvrir les coûts en amont ? Cela implique une identification claire des coûts à recouvrir et la répartition de ces derniers entre les organismes utilisateurs.

- Les droits d'accès peuvent-ils être perçus efficacement?

- Les fichiers complexes disponibles gratuitement sur un site Web sont susceptibles d'être consultés par des personnes ne disposant pas des compétences requises pour utiliser des microdonnées. Cela peut conduire à une augmentation des besoins d'assistance.

Autre facteur important à prendre en compte pour l'élaboration d'une stratégie tarifaire : la cohérence avec la politique tarifaire appliquée à d'autres produits et services, tels que les publications au format papier et l'accès à Internet. La majorité des sites Web des ONS sont accessibles gratuitement, dans la mesure où les coûts marginaux sont faibles, voire inexistants. Il n'en va pas de même pour les produits au format papier. La tarification de ces publications est fixée en fonction des coûts marginaux de production et d'expédition des copies supplémentaires. Le même principe pourrait s'appliquer aux fichiers de microdonnées et au surcoût généré par l'aide ou l'assistance fournie aux utilisateurs supplémentaires.

9. A quel moment du cycle de diffusion les microdonnées doivent-elles être rendues publiques ?

La production et la publication des fichiers de microdonnées doivent s'inscrire dans un cycle de diffusion. Il est fortement souhaitable que les informations destinées à un large public soient diffusées en priorité, de façon à ce que l'ONS puisse remplir ses objectifs immédiats et mettre en place un système de rétroaction du public. Il s'agit notamment des rapports descriptifs d'enquête et des analyses du producteur des données. Il est important que les producteurs de données officiels établissent ces documents officiels et les communiquent / publient dès le début du cycle de diffusion.

La production d'un fichier de microdonnées réclame du temps, mobilise des ressources spécialisées et exige une procédure de validation. De plus, les ONS doivent parfois répondre à des objectifs analytiques / scientifiques internes ou externes précis. Ils peuvent ainsi être amenés à reporter la création d'un fichier de microdonnées à quelques mois après la publication des résultats de l'enquête concernée. Quel que soit le calendrier prévu, les chercheurs aiment être informés des dates de publication prévisionnelles, de façon à pouvoir planifier leurs propres travaux. Les délais doivent être raisonnables. S'ils sont de plusieurs années, les résultats seront beaucoup moins pertinents.

« Même quand la diffusion intervient assez rapidement, il est parfois souhaitable de publier une partie des microdonnées ou des données agrégées avant la date de mise à disposition de l'ensemble complet de microdonnées. [Il est recommandé] de prévoir ces questions dès les phases de planification ou de les aborder dès que le besoin s'en fera sentir. » [14] (Traduction de l'anglais)

Encadré 16 **Politique relative aux délais de publication des données du NCHS (Etats-Unis)**

« Il est dans l'intérêt général et de celui de la science de pratiquer un échange ouvert d'informations et de points de vue. A cet égard, la politique du NCHS est de diffuser les microdonnées *le plus tôt possible* après la collecte des données, suivant les seules restrictions imposées par les ressources disponibles, par les contraintes techniques et par les exigences qualitatives. Le NCHS n'entravera pas une diffusion rapide des microdonnées aux fins de réserver à ses collaborateurs, à ses partenaires ou au personnel d'autres organisations les droits de publication y afférents.

1. Les fichiers de données à grande diffusion seront mis à disposition aussitôt après les étapes nécessaires de préparation, de contrôle et de validation par les instances compétentes, notamment par le Comité de contrôle de divulgation du NCHS.

   Suivant l'intérêt que présente les données et le domaine de compétences, le NCHS peut être amené à mettre à contribution certains partenaires (dont les bailleurs de fonds) dans le cadre du processus de préparation des données précédant leur diffusion, y compris pour la modification, le recodage et la définition de la structure définitive du fichier. Si ces étapes se révèlent nécessaires, le NCHS peut être amené à communiquer au partenaire désigné des données qui n'ont pas encore été rendues publiques. La communication dans un tel cadre de fichiers qui ne sont pas encore prêts à être diffusés doit être prévue par les dispositions de la politique de confidentialité du NCHS. Elle doit en outre s'effectuer en accord avec les dispositions légales applicables par le NCHS, avec le consentement éclairé et sur les conseils d'un comité d'éthique de la recherche sur des sujets humains. La communication de données dans un tel cadre est en principe régie par un accord stipulant les mesures de protection adéquates à prendre par le partenaire concerné.

2. Le NCHS ne « bloquera » pas les données prêtes à être diffusées. De plus, elle ne donnera pas à ses partenaires d'accès privilégié et anticipé à des fichiers de données ou à des tableaux prêts à être diffusés. Enfin, elle ne fournira pas d'accès privilégié à des tableaux basés sur des fichiers de données qui n'ont pas encore été rendus publics.

   Lorsque des données non publiées sont communiquées à un utilisateur, il convient d'en envisager la mise à disposition à d'autres demandeurs, dans la limite des clauses de confidentialité. Les fichiers ou tableaux dont la publication n'a pas encore été approuvée compte tenu du risque de divulgation d'informations confidentielles qu'ils présentent peuvent éventuellement être consultés au Centre de données du NCHS (ou, en accord avec la politique du NCHS en matière de confidentialité, faire l'objet de contrats d'utilisation spéciaux), ceci afin de garantir l'utilisation la plus large possible des données produites. Dans de rares cas (publications ministérielles dont la date de parution est lointaine p. ex.) des données tabulaires peuvent être mises à disposition avant leur diffusion générale.

Les exceptions à cette politique générale seront rares et justifiées au cas par cas. Toute demande doit être soumise avant le lancement de la collecte des données au responsable des questions de confidentialité désigné par le NCHS, et approuvée par le Directeur général. » (Traduction de l'anglais)

Source: http://www.cdc.gov/nchs/about/policy/data_release.htm (site consulté le 7 juin 2010, en anglais uniquement).

10. Quelles sont les exigences à remplir en termes d'infrastructure technique ?

« L'accès aux [microdonnées] et l'exploitation optimale de ces données nécessitent une infrastructure technologique adaptée, une large convergence de vues au plan international en matière d'interopérabilité, ainsi que des mécanismes efficaces de contrôle de la qualité des données. (…) La pérennité de l'infrastructure requise pour l'accès aux données revêt une importance particulière. Les établissements de recherche et les organismes publics devraient assumer officiellement la responsabilité de faire en sorte que les données (…) soient efficacement préservées, gérées et rendues accessibles de façon à pouvoir être exploitées de manière efficiente et adéquate sur le long terme. (…) Il convient [en outre] de s'attacher en particulier à encourager l'utilisation de techniques et d'instruments destinés à garantir l'intégrité et la sécurité des données de recherche. En ce qui concerne la garantie de l'intégrité d'un ensemble de données, tout devrait être mis en œuvre pour s'assurer du caractère complet des données et de l'absence d'erreurs. En ce qui concerne la sécurité, les données, de même que les métadonnées et descriptions correspondantes, devraient être protégées contre la perte, la destruction, la modification et l'accès non autorisé, intentionnels ou non, en conformité avec des protocoles de sécurité explicites. » [17]

Encadré 17 **Microdata Management Toolkit (IHSN)**

Le « Microdata Management Toolkit » (outils de gestion de microdonnées, ci-après le Toolkit), développé par le NSD norvégien et par le Groupe de gestion des données sur le développement de la Banque Mondiale dans le cadre du Réseau International pour les Enquêtes auprès des Ménages (IHSN), vise à promouvoir l'adoption de normes internationales et de bonnes pratiques en matière de documentation, de diffusion et de conservation de microdonnées.

Le Toolkit comprend deux modules. **Metadata Editor** (éditeur de métadonnées) est utilisé pour documenter les données conformément aux normes internationales en vigueur pour les métadonnées (DDI et Dublin Core). Grâce au programme gratuit **Explorer** (explorateur), les utilisateurs peuvent lire les fichiers créés avec Metadata Editor. Ce programme permet aux utilisateurs de visualiser les métadonnées et de réexporter les données vers divers formats communs (Stata, SPSS, etc.). Metadata Editor et Explorer font tous deux appel à la technologie Nesstar et ont été mis au point par le NSD norvégien. Enfin, **CD ROM Builder** (créateur de CD-ROM) est utilisé pour générer des produits faciles d'utilisation (CD-ROM, site Web) en vue de la diffusion et de l'archivage de données.

Voir http://www.ihsn.org/toolkit

Une infrastructure technologique adaptée doit être mise en place pour couvrir les différents aspects de l'archivage de microdonnées (documentation, catalogage et diffusion, anonymisation et conservation).

Documentation des microdonnées

Des normes de métadonnées internationales ont été élaborées pour formaliser la documentation des microdonnées et des ressources connexes. L'Initiative DDI et les normes du Dublin Core décrites au chapitre 2 constituent des solutions pratiques. La réalisation d'une documentation conforme à ces normes est par ailleurs facilitée par l'existence d'éditeurs de métadonnées spécialisés tels que le Microdata Management Toolkit de IHSN (Encadré 17) et le logiciel de gestion Nesstar Publisher du NSD norvégien.

Catalogage et diffusion des microdonnées

Les utilisateurs éventuels doivent être correctement informés de l'existence et des caractéristiques des ensembles de données mis à disposition. Nombre d'entre eux disposent souvent de très peu d'informations sur les ensembles de données disponibles – et encore ! Des métadonnées de qualité doivent être fournies, de préférence sous la forme d'un catalogue interrogeable en ligne.

L'objectif d'un catalogue de microdonnées est de permettre aux utilisateurs d'accéder facilement aux données et à la documentation via un format extrêmement pratique pour les utilisateurs. Un catalogue d'enquêtes contient divers outils pour :

- trouver le fichier de données qui répond le mieux aux besoins de l'utilisateur. Ce type d'outil paraît peu utile quand la quantité de fichiers de microdonnées est restreinte. Il prend toutefois toute son importance quand le nombre de fichiers augmente. Dans ce cas, un outil permettant de rechercher des fichiers de données par variables est très précieux.

- évaluer l'adéquation des informations trouvées avec les besoins du chercheur concerné (p. ex. univers de l'enquête, concepts et définitions). Ce rôle est assuré par les métadonnées qui constituent la documentation du fichier.

- accéder aux données. Cela fait appel à un système d'extraction et/ou de chargement des données. Généralement, les fichiers concernés peuvent être chargés via un site ou un portail Web et un serveur FTP. L'ONS peut également recourir à ce type d'outil en interne pour fournir des données sur CD/DVD.

- exploiter les données. Il n'existe pas d'outil unique pour l'exécution des travaux analytiques des chercheurs. Ces derniers préfèrent disposer de divers formats, ce qui leur permet d'utiliser les outils de leur choix. Les formats courants sont SPSS, STATA, SAS et ASCII.

Voici les caractéristiques d'un système de catalogage de microdonnées performant :

- **Du point de vue de l'utilisateur**, un bon catalogue :

- est conforme à une norme de métadonnées internationale.

- Les normes internationales de métadonnées en XML facilitent considérablement la production et la maintenance de ces catalogues.

- est basé sur le Web pour faciliter les recherches.

- est riche en métadonnées, y compris sur les variables.

- Les catalogues d'enquête sont d'autant plus pertinents et utiles quand les métadonnées contiennent une description détaillée de l'enquête proprement dite (informations sur l'intitulé, l'enquêteur principal, l'échantillonnage, la date de la collecte des données, les sujets abordés, la couverture géographique, etc.), mais aussi de chaque variable (désignation et étiquette des variables, catégories, formulation des questions, instructions aux enquêteurs, définitions).

- La norme de métadonnées DDI et les outils logiciels mis à disposition par IHSN, notamment le Microdata Management Toolkit et la plate-forme NADA (disponibles sur www.ihsn.org).

- est interrogeable sur tous les champs de l'étude. Dans le cadre de la norme DDI, cela signifie que le catalogue doit permettre d'effectuer des recherches sur une étude (intitulé, année, pays, organisation) et sur les variables (désignation de la variable, étiquette de la variable, étiquette de la valeur de la variable). Il est souhaitable que le catalogue offre des fonctionnalités de recherche en texte intégral conviviales.

- contient des informations claires sur la politique et sur les procédures d'accès aux données.

- comprend une liste et des liens directs vers les documents de référence (questionnaires, manuels, rapports).

- offre une fonction de « recherche par mot-clé » suivant une taxonomie standard.

  Afin de faciliter l'échange d'informations entre les catalogues, la communauté des archives de données a élaboré un thésaurus décrivant les thèmes couverts par les ensembles de données répertoriés dans leurs catalogues respectifs. Il s'agit d'une liste de termes ou de concepts utilisés pour décrire des éléments précis (ensembles de données, variables, livres, etc.). Les termes d'un thésaurus sont généralement organisés en arborescence ou sous forme de listes hiérarchisées (partant de termes génériques pour aller vers des notions plus spécifiques). En règle générale, le thésaurus indique des synonymes et des termes connexes pour faire en sorte que les recherches des utilisateurs aboutissent même s'ils n'emploient pas les termes consacrés.

  De nombreuses archives s'appuient sur un thésaurus pour l'ajout de mots-clés concernant une étude ou de concepts relatifs à des variables. L'utilisation d'un thésaurus favorise la cohérence en garantissant qu'un même objet soit toujours décrit par le même terme. De plus, la mise à disposition du thésaurus aux utilisateurs pour la recherche de données augmente leurs chances d'employer les termes et les concepts qui permettront d'obtenir la liste de résultats la plus pertinente.

A titre d'exemple, le catalogue géré par le CESSDA, organisation faîtière européenne des archives de données en sciences sociales, s'appuie sur un thésaurus.

- doit permettre d'afficher les résultats des recherches rapidement, même s'il est volumineux. Cela implique un système d'indexage efficace.

- propose un outil de comparaison des produits du catalogue. Cette fonction est utile pour comparer des variables entre plusieurs enquêtes standard ou quand plusieurs versions d'une enquête ont été téléchargées.

- fournit des informations bien visibles sur les politiques d'accès applicables à chaque étude (p. ex. disponibilité des microdonnées et, le cas échéant, description claire des modalités d'obtention de cellesci).

- met à disposition des fonctions d'aide en ligne performantes.

- fournit des liens entre les produits du catalogue et des ressources disponibles sur un site Web externe, et permet d'ajouter des informations complémentaires, telles que des références bibliographiques aux publications basées sur une étude.

- ***Du point de vue de l'administrateur du catalogue, un bon système de catalogage :***

  - offre un environnement sécurisé pour le stockage et le partage de données et de métadonnées.

  - met à disposition des outils de gestion des processus d'accès aux microdonnées (de l'accord automatique pour les microdonnées accessibles sans restriction à des systèmes de gestion et de traitement de procédures d'accès soumises à autorisation).

  - fournit une solution pour la communication de fichiers à grande diffusion et de fichiers sous licence.

  - constitue un moyen sûr de communiquer des microdonnées et de la documentation, accroissant ainsi le taux d'accès des utilisateurs finaux.

  - recueille des informations sur les utilisateurs du catalogue, sur les téléchargements et, le cas échéant, sur les fins auxquelles les données sont utilisées. De tels enregistrements présentent un intérêt pour les commanditaires des études, dans la mesure où ils leurs permettent d'évaluer l'utilisation qui est faite des microdonnées. Ils sont également utiles pour les utilisateurs finaux, qui ont ainsi la possibilité d'être informés de la publication de nouvelles versions des données ou de la révision des études qu'ils ont téléchargées.

Anonymisation des microdonnées

L'anonymisation des données réclame du personnel compétent en statistiques, qui soit familiarisé avec l'utilisation de logiciels tels que Stata ou SPSS. Certains logiciels spécialisés permettent de mesurer ou de diminuer le risque de divulgation. Aucune de ces applications n'offre néanmoins de solution intégrée satisfaisante pour les fichiers de données hiérarchisés complexes. Sur le plan pratique, l'anonymisation reste en grande partie un processus ad-hoc. Des travaux visant à élaborer des outils et des lignes directrices susceptibles de faciliter l'anonymisation des microdonnées sont en cours (parmi d'autres au sein du réseau IHSN).

Comme nous l'avons vu, l'anonymisation d'un fichier d'enquête se déroule en deux temps : il s'agit d'abord d'identifier les éléments présentant potentiellement un risque de divulgation, puis de mettre en œuvre des techniques de restriction ou de perturbation des données en vue de réduire ce risque. Cette dernière étape implique l'intervention d'une personne disposant des compétences nécessaires pour pouvoir recommander les restrictions d'informations qui nuiront le moins aux futurs utilisateurs des fichiers.

Conservation des microdonnées (et des métadonnées)

Les données et les métadonnées numériques sont vulnérables à l'obsolescence des logiciels, des composants matériels et des supports de stockage, ainsi qu'aux menaces physiques et aux erreurs humaines. La conservation des données et des métadonnées à long terme nécessite donc la mise en place de procédures et d'une infrastructure adéquates. Les principes et les bonnes pratiques en matière de conservation de données sont décrits en détails dans un document de travail IHSN élaboré par le Consortium interuniversitaire pour la recherche en sciences politiques et sociales (ICPSR) [8].

11. Quelles sont les exigences institutionnelles relatives à la diffusion de fichiers de microdonnées ?

Pour de nombreux ONS, la diffusion de fichiers de microdonnées est une activité très récente, susceptible de créer des configurations d'utilisation de leurs données inédites et importantes. Dans un document intitulé *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financ*ée sur fonds publics, L'Organisation de coopération et de développement économique (OCDE) définit une série de principes fondamentaux destinés aux fournisseurs de données :

> ***« Ouverture***
> Par ouverture, on entend l'accès dans des conditions d'égalité de la communauté scientifique internationale, à un coût le plus bas possible, de préférence ne dépassant pas le coût marginal de la diffusion. (…)
>
> ***Flexibilité***
> La flexibilité suppose de prendre en compte les évolutions rapides et souvent imprévisibles des technologies de l'information, (…), des systèmes juridiques et des cultures de chaque pays (…). (…)
>
> ***Transparence***
> L'information sur les données de la recherche et les organisations productrices de données, la documentation sur les données ainsi que les spécifications des conditions qui régissent leur utilisation devraient être accessibles au plan international, en toute transparence, dans l'idéal via l'internet. (…)
>
> ***Conformité au droit***
> Les dispositifs d'accès aux données devraient respecter les droits et intérêts légitimes de tous les acteurs de l'activité de recherche publique (…).
>
> ***Protection de la propriété intellectuelle***
> Les dispositifs d'accès aux données devraient tenir compte de l'applicabilité du droit d'auteur ou des autres législations sur la propriété intellectuelle pouvant concerner les bases de données de la recherche financée sur fonds publics. (…)
>
> ***Responsabilité formelle***
> Les dispositifs d'accès devraient promouvoir des pratiques institutionnelles explicites et formalisées, telles que l'élaboration de règles et de réglementations, sur les responsabilités des diverses parties intervenant dans les activités liées aux données. Ces pratiques devraient concerner la paternité des données, la mention des producteurs, la propriété, la diffusion, les restrictions d'utilisation, les modalités financières, les règles éthiques, les conditions de licence, la responsabilité civile et l'archivage durable. (…)
>
> ***Professionnalisme***
> Les dispositifs institutionnels pour la gestion des données de la recherche devraient être fondés sur les normes professionnelles applicables et sur les valeurs inscrites dans les codes de conduite des milieux scientifiques concernés. (…)
>
> ***Interopérabilité***
> L'interopérabilité technologique et sémantique est essentielle pour faciliter et encourager l'accessibilité et l'utilisation des données de la recherche dans un contexte international et interdisciplinaire. Les dispositifs d'accès devraient tenir dûment compte des normes internationales pertinentes applicables en matière de documentation des données (…).
>
> ***Qualité***
> La valeur et l'utilité des données de recherche dépendent pour une large part de la qualité des données elles-mêmes. Les gestionnaires de données et les organisations de collecte de données devraient particulièrement veiller au respect de normes de qualité explicites. (…)
>
> ***Sécurité***
> Il convient de s'attacher en particulier à encourager l'utilisation de techniques et d'instruments destinés à garantir l'intégrité et la sécurité des données (…). (…)

Efficience
L'un des buts essentiels poursuivis en s'attachant à promouvoir l'accès et le partage des données est d'améliorer l'efficience globale de la (collecte de données) financée sur fonds publics afin d'éviter une duplication inutile et coûteuse des efforts de collecte de données. (…)

Responsabilité de rendre compte
Le fonctionnement des dispositifs d'accès aux données devrait faire l'objet d'une évaluation périodique par les groupes d'utilisateurs, les institutions responsables et les organismes de financement (…). (…) [17]

Le respect de ces principes passera sans doute par de nouvelles procédures et par de nouvelles mentalités. Dans ses lignes directrices [17], l'OCDE définit les grandes questions liées à l'octroi de l'accès aux données (ces questions sont valables également pour les microdonnées créées par les producteurs de données officiels à des fins statistiques) :

- « Questions institutionnelles et de gestion : bien qu'une meilleure accessibilité soit importante pour toutes les communautés scientifiques, la diversité des activités scientifiques donne à penser que des approches sur mesure et des modèles institutionnels variés de la gestion des données seront sans doute plus efficaces pour répondre aux besoins des chercheurs.

- Questions financières et budgétaires : l'infrastructure des données scientifiques nécessite une planification budgétaire spécifique continue et un soutien financier adéquat. L'exploitation des données de la recherche ne pourra être optimisée si les coûts d'accès, de gestion et de conservation doivent être imputés après coup aux projets de recherche. Il faut néanmoins souligner que les coûts de stockage et de gestion des données ont fortement diminué ces dernières années, et le manque d'informations sur ces changements peut en soi représenter un obstacle au progrès.
- Questions juridiques et politiques : les législations nationales et les accords internationaux, notamment dans des domaines comme les droits de propriété intellectuelle et la protection de la vie privée, influent directement sur les pratiques en matière d'accès aux données et de partage des données, et doivent être pleinement pris en compte dans l'élaboration des dispositifs d'accès aux données.
- Questions de culture et de comportement : la mise en place de mécanismes de formation et de rétribution adaptés est une composante nécessaire de la promotion des pratiques en matière d'accès et de partage des données. Ces considérations valent pour ceux qui financent, produisent, gèrent et exploitent les données (…). »

(…) La responsabilité des divers aspects de l'accès aux données et de leur gestion devrait être établie dans les documents concernant les tâches officielles des instituts, les demandes de subventions, les contrats de recherche, les accords de publication et les licences, par exemple. » [17]

Les référentiels numériques fiables – une alternative

Pour certains producteurs de données, la mise en place et la gestion d'une archive de données et d'un service de diffusion est un objectif peu réaliste – entre autres pour des raisons budgétaires et juridiques. L'une des alternatives consiste à déléguer cette tâche. Prenons l'exemple de l'UKDA[15] hébergée par l'Université d'Essex, qui gère et diffuse les données des instituts de statistique, des établissements de recherche et des chercheurs eux-mêmes. Autre exemple : le Consortium interuniversitaire pour la recherche en sciences politiques et sociales (ICPSR), installé à l'Université du Michigan, qui remplit une fonction similaire aux EtatsUnis[16].

Ces archives de données assurent non seulement une gestion efficace de l'octroi des licences, mais jouent également un rôle de chef de file en matière de valorisation des données et d'innovation. A titre d'exemple, citons les nouvelles pages Web de l'UKDA, qui servent d'orientation pour la gestion et le partage des données. L'objectif est de proposer aux auteurs, aux gestionnaires et aux curateurs de données les meilleures stratégies et méthodes de création, de préparation et de stockage d'ensembles de données à vocation publique[17].

15 United Kingdom Data Archive ; voir http://www.data archive.ac.uk/

16 Voir http://www.esds.ac.uk/aandp/create/research.asp ou http://www.nsf.gov/pubs/2005/nsb0540/

17 Voir http://www.data archive.ac.uk/sharing/

12. Comment promouvoir l'utilisation des fichiers de microdonnées ?

L'existence de fichiers de microdonnées à grande diffusion ou de fichiers sous licence adaptés est-elle suffisante pour voir se développer une base d'utilisateurs de ces produits ? Pas nécessairement malheureusement. Il faut convaincre les utilisateurs de s'impliquer, donc les former en ce sens.

Les enquêtes nationales par sondage et les fichiers de données à grande diffusion issus de recensements présentent un intérêt pour une vaste communauté de chercheurs et d'analystes. Une documentation et une publicité appropriées devraient entraîner une très large utilisation de ces ensembles de données. Le programme mondial des enquêtes démographiques et de santé (*Demographic and Health Survey,* DHS) démontre de façon criante l'ampleur de la demande d'ensembles de données de ce type[18]. Les ensembles de données du programme DHS sont accessibles gratuitement et aisément. Téléchargés par un nombre considérable d'utilisateurs, ils ont contribué à une série très riche et diversifiée d'études et de publications.

Dans les pays habitués à produire ce type de fichier, la richesse du débat d'orientation sur leur utilisation est une évidence. Ailleurs, l'utilité des fichiers de microdonnées n'est cependant pas toujours bien comprise. Les établissements qui commencent tout juste à diffuser des fichiers de microdonnées devront sans doute recourir à divers moyens pour en promouvoir l'utilisation et informer les utilisateurs potentiels de la valeur qu'ils représentent – ainsi que de leurs limites.

Une culture du partage des données et de la coopération devrait produire une abondance de nouvelles connaissances. Les ONS et leurs partenaires sont vivement encouragés à promouvoir l'utilisation des fichiers de microdonnées au niveau national et international, notamment à travers des séminaires et des formations. Les opportunités ne manquent pas.

Les fichiers de microdonnées revêtent une grande importance pour la recherche et l'éducation. Ils jouent un rôle majeur dans l'élaboration des politiques et des programmes des gouvernements nationaux et des organisations internationales. Ces instances doivent être considérées comme les alliés naturels des ONS pour la promotion d'une utilisation appropriée des fichiers de microdonnées. Les universités ont en outre un rôle clé à jouer dans la formation des nouveaux utilisateurs.

Il est indispensable de faire en sorte que les bonnes personnes prennent conscience de l'existence du produit et des avantages qu'il présente, et d'organiser des campagnes de sensibilisation pour garantir le succès de la diffusion des microdonnées des ONS. Cela passe par l'identification et par un rapprochement des organisations et des membres de celles-ci susceptibles de devenir des utilisateurs potentiels. Dans de nombreux cas, les ONS ont déjà conscience de cette nécessité et ont déjà été sollicités par ces utilisateurs. Les liens vers des sites Web, les brochures et les invitations à des séminaires comptent parmi les mesures envisageables.

Comme nous l'avons évoqué, il est impératif de former les utilisateurs potentiels. L'Initiative de démocratisation des données canadienne (IDD) organise ainsi des sessions de formation visant à promouvoir et à appuyer l'utilisation des fichiers de microdonnées ainsi que d'autres produits pour la recherche et à des fins d'enseignement[19]. Au Canada, de nombreux chercheurs étaient déjà familiarisés avec les fichiers de microdonnées. Ce n'était cependant pas le cas des établissements chargés d'en assurer la promotion. La situation a été améliorée grâce à un système de mailing électronique permettant aux chercheurs, au personnel des ONS, aux intermédiaires et aux autres personnes intéressées de poser des questions et de partager leur expérience. L'archive des réponses fournies constitue une source d'informations inestimable et soulage les ONS, qui ainsi ne sont pas tenus de répondre à toutes les questions qu'ils reçoivent.

L'une des clés du succès de la diffusion de fichiers de données aux chercheurs est l'interaction du personnel des ONS avec le réseau de la communauté de chercheurs et des archives. Cela aide les deux parties à cerner mutuellement leurs besoins et leurs problématiques et crée les bases d'une communication bidirectionnelle.

18 http://www.measuredhs.com

19 Nous invitons les lecteurs à visiter le site du Dépôt des documents de formation de l'IDD : https://ospace.scholarsportal.info/handle/1873/69 (site consulté le 18 octobre 2007).

Annexe 1

Demande d'accès à un ensemble de données sous licence dans le cadre d'un projet de recherche précis

Le présent modèle de formulaire doit être adapté aux exigences spécifiques applicables.

Les informations fournies par vos soins dans le présent formulaire resteront confidentielles, sauf en cas de violation de l'accord légal conclu, auquel cas l'ONS peut être amené à en informer des instituts de statistique partenaires situés dans d'autres pays.

Merci de renvoyer le présent formulaire par courrier ou par fax à l'adresse ou au numéro suivant, accompagné d'une lettre à l'en-tête de l'organisme responsable :

Adresse : [adresse]
Fax : [numéro de fax]
E-mail (copie scannée) : [adresse e-mail]

Intitulé et numéro de référence du / des ensemble(s) de données demandés (titre exact, année et numéro de référence figurant dans le catalogue d'enquêtes):

_____

_____

Dispositions applicables

Dans le présent accord,

1. Le « Chercheur principal » désigne le principal interlocuteur désigné pour toute communication concernant le présent accord. Le Chercheur principal assume la responsabilité du respect des dispositions du présent Accord régissant l'accès aux données. Le chercheur principal doit être habilité à représenter l'organisme bénéficiaire dans le cadre de ce type d'accord

2. Les « Autres chercheurs » désignent les personnes différentes du Chercheur principal, y compris les assistants, qui auront accès aux données confidentielles.

3. L'« Organisme bénéficiaire » désigne l'organisation / l'université / l'établissement qui emploie le Chercheur principal.

Section A. Chercheur principal

- Prénom _____
- Nom _____
- Titre _____
- Organisation _____
- Fonction dans l'organisation _____
- Adresse postale _____
- Téléphone (avec code pays) _____
- Fax (avec code pays) _____
- E-mail _____

Section B. Autres chercheurs

Indiquez les noms, titres et organismes d'appartenance de tous les autres membres de l'équipe de chercheurs qui accèderont aux données confidentielles.

- Nom et prénom _____
- Fonction _____
- Organisme d'appartenance _____

Section C. Organisme bénéficiaire

Nom de l'organisation _____

Type d'organisation (cochez la bonne réponse)

- ☐ Administration publique et assimilée
- ☐ Université
- ☐ Centre de recherche
- ☐ Entreprise privée
- ☐ Organisation internationale
- ☐ Organisation non-gouvernementale (nationale)
- ☐ Organisation non-gouvernementale (internationale)
- ☐ Autre (précisez) _____

Site Web de l'organisation (URL) _____
Adresse postale _____

Section D. Description de l'utilisation envisagée des données

Prière de décrire le projet de recherche (questions, objectifs, méthodes, résultats escomptés, partenaires). Si les informations fournies sont insuffisantes, vous risquez de voir votre demande rejetée ou de recevoir une demande d'informations complémentaires. Ces informations peuvent être jointes en annexe à la présente demande.

Liste du/des résultat(s) attendu(s) et politique de diffusion

_____
_____
_____

Section E. Identification des fichiers de données et des variables nécessaires

Des métadonnées détaillées sont disponibles sur le site Web de l'ONS. Elles comprennent une description des fichiers de données et des variables constituant chaque ensemble de données. Les chercheurs qui n'ont pas besoin de consulter l'ensemble de données complet sont priés d'indiquer le sous-ensemble de variables ou de cas qui les intéressent. Dans une telle configuration, le risque de divulgation est réduit. La probabilité d'obtenir les données demandées est donc plus grande.

La demande vise à obtenir l'accès (cochez la bonne réponse) :

- ☐ à l'ensemble de données complet (tous les fichiers et cas).
- ☐ à un sous-ensemble de variables et/ou de cas, conformément aux indications suivantes (remarque : les variables telles que les coefficients de pondération de l'échantillon et les identifiants des enregistrements sont systématiquement contenues dans les sous-ensembles).

Section F. Accord régissant l'accès aux données

Le Chercheur principal et les Autres chercheurs acceptent de se conformer aux dispositions suivantes :

1. Les données confidentielles pourront être consultées uniquement par le Chercheur principal et par les Autres chercheurs désignés dans le présent Accord.
2. Toute reproduction ou communication de copies des données confidentielles ou de toutes données fondées sur les données initiales à des personnes différentes de celles mentionnées dans le présent Accord est interdite, sauf autorisation exprès de l'ONS.
3. Les données seront utilisées exclusivement à des fins statistiques ou dans le cadre de travaux de recherche. Elles serviront uniquement à constituer des données agrégées, toute recherche portant sur des particuliers ou des organisations individuelles étant proscrite. Les données ne seront pas utilisées, de quelque façon que ce soit, à des fins administratives, propriétaires ou pour faire appliquer la loi.
4. Le Chercheur principal s'engage à ce qu'aucune personne n'utilise les données fournies dans le but d'essayer d'identifier un individu, une famille, une entreprise ou une organisation. Qui plus est, il ne sera fait usage de l'identité d'aucune personne ni organisation découverte fortuitement. Toute découverte de ce type devra être signalée sans délai à l'ONS. Une telle découverte ne saurait être révélée à quiconque ne figurerait pas dans le présent Accord régissant l'accès aux données.
5. Le Chercheur principal prendra les mesures de sécurité qui s'imposent pour éviter tout accès non autorisé aux microdonnées sous licence fournies par l'ONS. Les microdonnées devront impérativement être détruites une fois les travaux de recherche terminés, à moins de fournir à l'ONS une garantie suffisante de la sécurisation des données et sous réserve que l'ONS n'autorise leur conservation par écrit. La destruction des microdonnées sera confirmée par écrit à l'ONS par le Chercheur principal.
6. Tous les livres, articles, documents de travail, thèses, dissertations, rapports ou autres publications fondés sur des données obtenues auprès de l'ONS contiendront des références à leur source, conformément à l'exigence de citation associée à l'ensemble de données fourni.
7. Une copie au format électronique de toutes les publications fondées sur les données demandées sera adressée à l'ONS.
8. L'ONS et les organismes de financement concernés rejettent toute responsabilité relative à l'utilisation, à l'interprétation ou à des conclusions tirées des données fournies.
9. Le présent Accord entrera en vigueur à la date d'octroi des droits d'accès à l'ensemble de données confidentiel et restera applicable jusqu'à la date de fin du projet, ou jusqu'à une date antérieure si le projet se termine de manière anticipée.
10. Il incombe au Chercheur principal de requérir l'accord préalable de l'ONS pour toute modification des spécifications du projet, des dispositifs de sécurité, ainsi que pour tout changement affectant le personnel ou l'organisation décrite dans le présent formulaire de demande d'accès. Tout changement intervenant au sein de l'organisation employant le Chercheur principal implique le dépôt d'une nouvelle demande et met fin au projet initial.
11. Toute violation des dispositions du présent Accord sera prise très au sérieux par l'ONS, qui entamera des poursuites contre les personnes jugées responsables de l'infraction, qu'elle soit ou non intentionnelle. Un non-respect des consignes de l'ONS est considéré comme une violation majeure du présent Accord et peut entraîner des poursuites judiciaires. L'ONS tiendra et partagera avec ses archives de données partenaires une liste des personnes et des organisations coupables d'une violation des dispositions de l'Accord régissant l'accès aux données, qui se verront interdire la diffusion de données à l'avenir.

Signataires

Le Chercheur principal ou un représentant habilité de l'organisme bénéficiaire a lu et approuvé les dispositions de l'Accord régissant l'accès aux données, présentées à la section F ci-dessus :

Nom _____

Signature _____
Date _____

Demande examinée par … le [date]
Décision du comité :

- ☐ Demande acceptée
- ☐ Demande rejetée [motif] _____
- ☐ Demande de complément d'information : _____
  _____

Annexe 2

Modèle de politique d'accès à un centre de données sécurisé

Il s'agit d'une proposition de formulation à adapter en fonction du pays concerné.

Objectifs

Le Centre national de données sécurisé (« Centre ») a été créé par l'Archive de données nationale pour permettre aux chercheurs disposant de certaines qualifications d'accéder à des fichiers de microdonnées statistiques confidentiels, sous étroite surveillance. Le Centre met à disposition un mécanisme permettant aux chercheurs de consulter des fichiers de données détaillés en toute sécurité, sans mettre en péril l'anonymat des répondants.

Lieu

Le Centre est situé [adresse, tél., fax, e-mail et site Web]

Activités du Centre

Les chercheurs peuvent accéder aux données sur place sous la surveillance du personnel du Centre. Un ordinateur est mis à leur disposition, ainsi que des logiciels et un espace de bureau.

Données

- Le personnel du Centre constitue les fichiers de données nécessaires avant l'arrivée du chercheur et veille à ce qu'aucune donnée confidentielle ne quitte le Centre.
- Les chercheurs souhaitant effectuer des analyses multiples faisant appel à plusieurs ensembles de données n'auront accès qu'à un ensemble de données à la fois. En aucun cas les chercheurs ne seront autorisés à fusionner des ensembles de données de leur propre initiative.
- Le Centre autorise les chercheurs à apporter leurs propres données anonymisées en vue de les associer avec les ensembles de données du Centre et de créer des jeux de données fusionnés destinés à être stockés au Centre. Les données ainsi fournies par les chercheurs doivent être des données propriétaires recueillies et « détenues » par les chercheurs concernés, ou d'autres données publiques légalement obtenues par les chercheurs. Les chercheurs devront IMPERATIVEMENT fournir au Centre une documentation complète de toutes les données destinées à être fusionnées avec des données du Centre. Il incombe aux chercheurs qui souhaitent recourir à des fichiers fusionnés de solliciter le personnel du Centre pour s'assurer que leurs données peuvent être fusionnés avec celles du Centre. Le Centre prend en charge les fichiers de données au format SAS, SPSS et Stata.
- Le Centre crée des copies de sauvegarde de tous les fichiers informatisés à intervalles réguliers. Les fichiers de sauvegarde sont stockés dans un endroit sûr, dont l'accès est réservé au personnel du Centre. Ils peuvent néanmoins être mis à disposition des chercheurs ayant besoin d'effectuer des analyses complémentaires. Ces fichiers de sauvegarde contiennent les données fournies par les utilisateurs, ainsi que les fichiers fusionnés. Ils seront détruits sur demande écrite de l'utilisateur.

Postes de travail

- Le Centre dispose de [N] postes de travail réservés aux utilisateurs et d'une imprimante laser en noir et blanc installée dans une pièce sécurisée. Les ordinateurs du Centre ne sont pas connectés à Internet et sont configurés de façon à ce qu'aucun support amovible (cédérom, DVD, disquette ou clé USB) ne puisse être employé par l'utilisateur.
- Les postes de travail du Centre sont équipés d'un processeur [Pentium X XXX MHz] et fonctionnent sous [Windows NT / Autre].

Logiciels

- Les logiciels CSPro, EPI-Info, SAS, SPS et Stata sont installés sur les postes de travail en plus du pack MSOffice. Des langages analytiques / de programmation supplémen-taires peuvent être pris en charge sur demande. Pour plus d'informations sur les versions des logiciels disponibles au Centre, prière de nous contacter.
- Les chercheurs doivent disposer des compétences nécessaires pour effectuer leurs propres analyses à l'aide de l'une des applications logicielles fournies. Le Centre n'assure pas de service d'assistance technique pour ces logiciels.

Espace de bureau

- Les chercheurs travaillent sous la surveillance du personnel du Centre et uniquement aux heures d'ouvertures de bureau habituelles (du lundi au vendredi, 8h30 – 17h00).
- L'accès au Centre est réservé aux chercheurs mentionnés dans le formulaire de demande. Une pièce d'identité avec photo leur sera demandée à l'entrée.
- Un maximum de trois chercheurs travaillant sur le même projet peuvent occuper un poste de travail.
- Le Centre applique la règle du « premier arrivé, premier servi ».

Surveillance du personnel du Centre (en vue du contrôle de divulgation)

- Les chercheurs externes s'abstiendront d'apporter tout document, manuel, ouvrage, etc., susceptible de permettre l'identification et la divulgation d'informations personnelles mises à disposition au Centre. De même, les téléphones portables, pagers et autres outils de communication avec l'extérieur sont interdits.

- Les chercheurs n'enregistreront aucun résultat, fichier ou programme sur un support de données électronique amovible. Le personnel du Centre s'en chargera sur demande.

- Les chercheurs ne seront autorisés à emporter les résultats de leurs analyses qu'après un contrôle de divulgation effectué par le personnel du Centre. Les contrôles de divulgation consistent à rechercher les cellules correspondant à moins de cinq cas, les tableaux avec variables géographiques, les modèles avec variables géographiques (ou des variables équivalentes à des variables géographiques), ou encore des listes de cas.

- Tous les fichiers journal doivent être imprimés ou archivés au format électronique. Ils seront conservés par le Centre, qui gardera uniquement les programmes et les procédures exécutés par des chercheurs externes. Les fichiers journal concernant leurs propres recherches ne seront pas conservés.

- Tous les résultats générés par des programmes statistiques et toutes les notes manuscrites y relatives feront l'objet d'un contrôle de divulgation par le personnel du Centre avant de quitter le Centre. Seuls les tableaux synthétiques peuvent être emportés. En aucun cas un tableau ne doit contenir des cellules correspondant à moins de cinq cas observés. Le cas échéant, de telles cellules seront supprimées, généralement par effacement. Pour veiller à ce qu'elles ne puissent être reconstituées à partir d'autres cellules de la même ligne ou colonne, le personnel supprime les totaux pour les lignes et les colonnes correspondant à ces cellules. Une fois le contrôle de divulgation terminé, les chercheurs obtiennent une photocopie des tableaux définitifs. Le personnel du Centre s'appuie sur les meilleures pratiques en vigueur pour déterminer si les données tabulaires sont identifiables et prennent des décisions prudentes. Les décisions du Centre sont irrévocables et ne sauraient être remises en cause par les chercheurs.

Frais d'admission

Les utilisateurs du Centre devront s'acquitter de frais correspondant à la location d'espace et d'équipement, au temps de surveillance par le personnel, au contrôle de divulgation, à la maintenance des équipements informatiques (composants matériels et logiciels), ainsi qu'à la création et à la gestion des fichiers de données demandés par le chercheur. Frais d'accès au Centre :

Organisme d'appartenance du chercheur principal	Frais de préparation et de création des fichiers (coûts fixes)	Utilisation des équipements (par jour et par poste de travail)
Utilisateurs nationaux		
Personnel du agence membre du Centre	Gratuit	Gratuit
Autre organisme public	[Coût/monnaie]	[Coût/monnaie]
Université / centre de recherche	[Coût/monnaie]	[Coût/monnaie]
ONG	[Coût/monnaie]	[Coût/monnaie]
Utilisateurs internationaux		
Recherche en partenariat avec le Centre	[Coût/monnaie]	[Coût/monnaie]
Organisation internationale	[Coût/monnaie]	[Coût/monnaie]
Université / centre de recherche	[Coût/monnaie]	[Coût/monnaie]
ONG	[Coût/monnaie]	[Coût/monnaie]

Un montant supplémentaire pourra être facturé en cas d'opération spéciale – fusion de données supplémentaires, création de formats de fichier personnalisés ou encore acquisition et installation d'un logiciel spécifique non standard. Ce montant sera convenu entre le chercheur et le personnel du Centre. Les paiements s'effectuent à l'avance, préalablement à l'utilisation du Centre.
Les paiements sont à effectuer à l'ordre de : [informations relatives au mode de paiement]

Soumission des projets de recherche

Les chercheurs utiliseront le formulaire de la page suivante pour soumettre leurs projets. Les chercheurs potentiels sont invités à vérifier auprès du personnel du Centre que les données qui les intéressent sont effectivement disponibles avant de rédiger leur projet. Ce dernier doit être décrit de façon à aider le personnel du Centre à créer les fichiers analytiques requis pour le projet. Les variables nécessaires et la sélection de cas éventuellement requise devront être clairement précisées. Seuls les éléments de données indispensables aux analyses envisagées seront contenus dans le fichier de données analytique. Le demandeur devra indiquer en quoi les données souhaitées sont nécessaires. Les projets trop volumineux et complexes ou, à l'inverse, trop minces nécessiteront une communication approfondie entre le personnel du Centre et les demandeurs, ce qui est susceptible de ralentir le processus. Le travail de préparation des fichiers de données pourra être réalisé dans un délai raisonnable si les projets volumineux et complexes sont scindés en plusieurs parties et si les données nécessaires sont clairement définies.

Les chercheurs qui souhaitent combiner des données du Centre avec des données externes devront fournir ces dernières au personnel du Centre avant de venir.

Dès réception, le projet de recherche sera évalué par un comité de contrôle réuni à cet effet.

Les critères suivants président à l'examen d'un projet :

- Faisabilité technique et scientifique du projet
- Disponibilité des ressources du Centre
- Risque de divulgation d'informations confidentielles

A noter qu'en acceptant la demande formulée, le Centre ne valide pas la pertinence générale ou la méthodologie du projet, ni les théories sous-jacentes, et ne lui reconnaît pas de valeur particulière.

L'approbation du Centre constitue une simple appréciation de la légalité de l'utilisation du fichier de données dans le cadre des recherches décrites et indique que le projet pourra probablement être mené à bien au Centre.

Annexe 3

Demande d'accès à un centre de données sécurisé

Le présent modèle de formulaire doit être adapté aux exigences spécifiques applicables.

Les informations fournies par vos soins dans le présent formulaire resteront confidentielles, sauf en cas de violation de l'accord légal conclu, auquel cas le Centre peut être amené à en informer des instituts de statistique partenaires situés dans d'autres pays.

Merci de renvoyer le présent formulaire par courrier ou par fax à l'adresse ou au numéro suivant, accompagné d'une lettre à l'en-tête de l'organisme responsable :

Adresse : [adresse]
Fax : [numéro de fax]
E-mail (copie scannée) : [adresse e-mail]

Intitulé et numéro de référence du / des ensemble(s) de données demandés (titre exact, année et numéro de référence figurant dans le catalogue d'enquêtes):

_____

_____

Dispositions applicables

Dans le présent accord,

1. Le « Chercheur principal » désigne le principal interlocuteur désigné pour toute communication concernant le présent accord. Le Chercheur principal assume la responsabilité du respect des dispositions du présent Accord régissant l'accès aux données. Le chercheur principal doit être habilité à représenter l'organisme bénéficiaire dans le cadre de ce type d'accord

2. Les « Autres chercheurs » désignent les personnes différentes du Chercheur principal, y compris les assistants, qui auront accès aux données confidentielles.

3. Le « Organisme bénéficiaire » désigne l'organisation / l'université / l'établissement qui emploie le Chercheur principal.

4. Le « Représentant de l'organisme bénéficiaire » désigne une personne habilitée à représenter l'Organisme bénéficiaire dans le cadre d'un tel accord.

Section A. Chercheur principal

- Prénom _____
- Nom _____
- Titre _____
- Organisation _____
- Fonction dans l'organisation _____
- Adresse postale _____
- Téléphone (avec code pays) _____
- Fax (avec code pays) _____
- E-mail _____

Section B. Autres chercheurs

Indiquez les noms, titres et organismes d'appartenance de tous les autres membres de l'équipe de chercheurs qui accèderont aux données confidentielles.

- Nom et prénom _____
- Fonction _____
- Organisme d'appartenance _____

Joindre à la présente demande un résumé du parcours ou le CV de chaque personne participant aux recherches, en précisant leur nationalité.

Section C. Organisme bénéficiaire

Nom de l'organisation _____

Type d'organisation (cochez la bonne réponse)

- ☐ Administration publique et assimilée
- ☐ Université
- ☐ Centre de recherche
- ☐ Entreprise privée
- ☐ Organisation internationale
- ☐ Organisation non-gouvernementale (nationale)
- ☐ Organisation non-gouvernementale (internationale)
- ☐ Autre (précisez) _____

Site Web de l'organisation (URL) _____
Adresse postale _____

Section D. Représentant de l'organisme bénéficiaire

- Prénom _____
- Nom _____
- Titre _____
- Prof./Dr/Mlle/Mme/M. _____
- Organisation _____
- Fonction dans l'organisation _____
- Adresse postale _____
- Téléphone (avec code pays) _____
- Fax (avec code pays) _____
- E-mail _____

Section E. Description de l'utilisation envisagée des données

Prière de décrire le projet de recherche (questions, objectifs, méthodes, résultats escomptés, partenaires). Indiquez pourquoi les ensembles de données publics ne répondent pas pleinement à vos besoins. Si les informations fournies sont insuffisantes, vous risquez de voir votre demande rejetée ou de recevoir une demande d'informations complémentaires. Ces informations peuvent être jointes en annexe à la présente demande.

Liste du/des résultat(s) attendu(s) et politique de diffusion

Prévoyez-vous de fusionner l'ensemble de données avec d'autres données ? ☐OUI ☐NON

Si OUI, indiquez ici la référence de tous les autres ensembles de données qui doivent être fusionnés.

_____
_____
_____
_____
_____

Section F. Identification des fichiers de données et des variables nécessaires

Des métadonnées détaillées sont disponibles sur le site Web du Centre. Elles comprennent une description des fichiers de données et des variables constituant chaque ensemble de données. Les chercheurs sont priés d'indiquer le sous-ensemble de variables ou de cas qui les intéressent, afin que le Centre puisse préparer les fichiers de données.

La demande vise à obtenir l'accès :

- ☐ à l'ensemble de données complet (tous les fichiers et cas).
- ☐ à un sous-ensemble de variables et/ou de cas, conformément aux indications suivantes (remarque : les variables telles que les coefficients de pondération de l'échantillon et les identifiants des enregistrements sont systématiquement contenues dans les sous-ensembles).

Section G. Logiciels nécessaires

Les chercheurs utiliseront le logiciel suivant :

☐ CSPro ☐ SAS ☐ SPSS ☐ Stata

Autre logiciel (précisez) : _____

Remarques :

- Le Centre met régulièrement à jour ses logiciels. Prière de nous contacter pour plus d'informations sur la version disponible de chaque application.
- Les chercheurs devant utiliser un logiciel ne figurant pas parmi la liste des logiciels standard fournis par le Centre devront fournir une licence d'utilisation de celui-ci en cours de validité. Le personnel du Centre installera le logiciel pour la durée des travaux de recherche (la licence d'utilisation restera la propriété du chercheur). Nous vous recommandons de contacter le Centre avant de finaliser votre demande, afin de vérifier la faisabilité technique du projet.

Section H. Accord régissant l'accès aux données

Sous réserve de son approbation par les deux parties, l'accord suivant sera signé :

Le Chercheur principal, les Autres chercheurs et le Représentant de l'organisme bénéficiaire acceptent de se conformer aux dispositions suivantes :

1. Les données confidentielles pourront être consultées uniquement par le Chercheur principal et par les Autres chercheurs désignés dans le formulaire de demande, qui signeront une Déclaration de confidentialité.
2. Les données seront utilisées exclusivement à des fins statistiques. Elles serviront uniquement à constituer des données agrégées, toute recherche portant sur des particuliers ou des organisations individuelles étant proscrite. Les données ne seront pas utilisées, de quelque façon que ce soit, à des fins administratives, propriétaires ou pour faire appliquer la loi.
3. Le Chercheur principal s'engage à ce qu'aucune personne n'utilise les données fournies dans le but d'essayer d'identifier un individu, une famille, une entreprise ou une organisation. Qui plus est, il ne sera fait usage de l'identité d'aucune personne ni organisation découverte fortuitement. Toute découverte de ce type devra être signalée sans délai au Centre. Une telle découverte ne saurait être révélée à quiconque ne figurerait pas dans le présent Accord régissant l'accès aux données.
4. Tous les livres, articles, documents de travail, thèses, dissertations, rapports ou autres publications fondés sur des données obtenues auprès du Centre contiendront des références à leur source, conformément à l'exigence de citation associée à l'ensemble de données fourni.
5. Une copie au format électronique de toutes les publications fondées sur les données demandées sera adressée au Centre.
6. Le collecteur initial des données, le Centre et les organismes de financement concernés rejettent toute responsabilité relative à l'utilisation, à l'interprétation ou à des conclusions tirées des données fournies.
7. Toute violation des dispositions du présent Accord sera prise très au sérieux par le Centre, qui entamera des poursuites contre les personnes jugées responsables de l'infraction, qu'elle soit ou non intentionnelle. Un non-respect des consignes du Centre est considéré comme une violation majeure du présent Accord et peut entraîner des poursuites judiciaires. Le Centre tiendra et partagera avec ses archives de données partenaires une liste des personnes

et des organisations coupables d'une violation des dispositions de l'Accord régissant l'accès aux données, qui se verront interdire la diffusion de données à l'avenir.

8. Le Centre se réserve le droit de mettre fin à tout projet à tout instant s'il estime que les activités d'un chercheur sont de nature à compromettre la confidentialité ou les normes déontologiques à respecter dans un environnement de recherche.
9. Aucune document imprimé, fichier électronique, autre document ou support ne quittera le Centre sans avoir été soigneusement examiné au préalable par le personnel du Centre en vue d'éliminer tout risque de divulgation.
10. Le Chercheur principal et les Autres chercheurs peuvent se voir fermer définitivement les portes du Centre pour autant que son Directeur le juge nécessaire pour protéger l'intégrité la confidentialité du Centre.

Signataires

Les signataires ci-dessous ont lu et approuvé les dispositions de l'Accord régissant l'accès aux données, présentées à la section H ci-dessus :

Le Chercheur principal

Nom _____

Signature _____
Date _____

Le Représentant de l'organisme bénéficiaire

Nom _____

Signature _____
Date _____

Le Centre attend de tous les chercheurs qu'ils appliquent les normes et les principes relatifs à la recherche statistique dans le cadre de leurs travaux. Seules les analyses ayant été approuvées pourront être effectuées. Un non-respect des consignes entraînera l'annulation du projet de recherche et un éventuel bannissement du Centre à l'avenir.

Références

[1] Altman, M. et King, G. 2006. *A Proposed Standard for the Scholarly Citation of Quantitative Data*. http://gking.harvard.edu/files/cite.pdf

[2] Boyko, E. et Watkins, W. 2003. *Sécurité des données, sécurité des environnements : les deux sont indispensables*. Actualité Eurostat. *19e séminaire du CEIES – Solutions innovantes permettant l'accès aux microdonnées - Lisbonne, 26 et 27 septembre 2002*, pp 109-118. www.cnis.fr/Agenda/CR/CR_0118.PDF

[3] CENEX-SDC. 2007. *Handbook on Statistical Disclosure Control*. http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf

[4] Dupriez, O. et Greenwell, G. 2007. *Quick Reference Guide for Data Archivists*. Document de IHSN. http://www.ihsn.org/home/download.php?file=DDI_IHSN_Checklist_OD_06152007.pdf

[5] Eurostat. 2009. *Work Session on Statistical Data Confidentiality*. Manchester 17-19 décembre 2007, *Methodologies and Working Papers*. http://www.unece.org/stats/publications/Proceedings_statistical_data_confidentiality.pdf

[6] Hamilton, E. et Humphrey, C. 2000. *Measuring the Impact of DLI: Use of the NPHS Public Use Microdata File in Academic Outcomes*.

[7] Hamilton, E, et Humphrey, C. 2002. *L'IDD et l'ENSP :* Étude *de compatibilité*. Automne 2002. http://www.statcan.gc.ca/dli-ild/doc/update-bulletin-v52-fra.pdf

[8] Consortium interuniversitaire pour la recherche en sciences politiques et sociales (ICPSR). 2009. *Principes et bonnes pratiques en matière de conservation de données*, Réseau International pour les Enquêtes auprès des Ménages (IHSN), IHSN Document de travail n°003. Décembre 2009. http://www.ihsn.org/home/index.php?q=focus/principles-and-good-practicepreserving-data

[9] ISO/CEI. 1999. *ISO/CEI 11179-1 – Technologies de l'information* – Spécification *et normalisation des éléments de données* – Part*ie 1 : Cadre pour la spécification et la normalisation des éléments de données*. http://metadata stds.org/11179-1/ISO-IEC_11179-1_1999_IS_E.pdf

[10] King, Gary. 1995. *Replication, Replication. PS: Political Science and Politics,* avec les commentaires de dix-neuf auteurs. http://gking.harvard.edu/files/replication.pdf

[11] King, Gary. 1995. *A Revised Proposal, Proposal*. Vol. XXVIII, N° 3, septembre 1995, pp. 443-499. http://gking.harvard.edu/files/abs/replicationabs.shtml

[12] Lambert, D. 1993. *Measures of Disclosure Risk and Harm*, Journal of Official Statistics, Vol 9, 407-426.

[13] Madsen, P. 2003. *The Ethics of Confidentiality: The Tension Between Confidentiality and the Integrity of Data Analysis in Social Science Research*, mimeo, Université Carnegie Mellon, Etats-Unis, juin 2003.

[14] National Center for Health Statistics (NCHS). 2002. *Policy on Micro-data Dissemination*. http://www.cdc.gov/nchs/data/NCHS%20Micro-Data%20Release%20Policy%204-02A.pdf

[15] National Center for Health Statistics (NCHS). 2004. *NCHS Staff Manual on Confidentiality*. http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

[16] National Center for Health Statistics (NCHS) – Research Data Center. 2008. *Guidelines for Proposal Submission*. http://www.cdc.gov/nchs/data/r&d/guidelines_10_14_08c.pdf

[17] Organisation de coopération et de développement économiques (OCDE). 2007. « Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics ». http://www.oecd.org/dataoecd/9/60/38500823.pdf

[18] Statistique Canada. *Comment citer les produits de Statistique Canada*. http://www.statcan.gc.ca/pub/12-591-x/12- 591-x2006001-fra.htm (site consulté le 22 juin 2010)

[19] Tambay, J. L., Goldmann, G., and White, P. 2001. *Providing Greater Access to Survey Data for Analyses at Statistics Canada*", Proceedings of the Annual Meeting of the American Statistical Association.

[20] UK Data Archive, University of Essex. 2002. *Good Practices in Data Documentation*. Version révisée. http://www.esds.ac.uk/news/goodPractice.pdf

[21] UK Data Archive, University of Essex. 2009. *Managing and Sharing Data. A Best Practice Guide to Researchers*, deuxième édition. http://www.data archive.ac.uk/news/ publications/ managingsharing.pdf

[22] UK Statistics Authority. 2009. *Code of Practice for Official Statistics*. Edition 1.0. Janvier 2009. http://www.statisticsauthority.gov.uk/ assessment/code-of-practice/code-of-practicefor-official-statistics.pdf

[23] Commission économique des Nations Unies pour l'Europe (CEE-ONU). 2000. *Terminology on Statistical Metadata*. Conférence des statisticiens européens, Normes et études statistiques, N° 53, Genève. http://www.unece.org/stats/publications/53metadaterminology.pdf

[24] Commission économique des Nations Unies pour l'Europe (CEE-ONU). Conférence des statisticiens européens. 2007. *Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice*. http://www.unece.org/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf[20]

[25] Commission économique des Nations Unies pour l'Europe (CEE-ONU). 2009. *Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes*. http://www.unece.org/stats/publications/Confidentiality_aspects_data_integration.pdf

[26] Commission économique des Nations Unies pour l'Europe (CEE-ONU) et Statistics Sweden. 2003. *Confidentialité des données statistiques et microdonnées*. Discussions menées lors du séminaire de la Conférence des statisticiens européens de 2003. http://www.unece.org/stats/publications/statistical.confidentiality.pdf

[27] Commission de statistique des Nations Unies (UNSD). 1994. Sixième *principe*, *Principes fondamentaux de la statistique officielle*. http://unstats.un.org/unsd/methods/statorg/FP-French.htm

[28] United States Bureau of the Census, Software and Standards Management Branch, Systems Support Division. 2008. *Survey Design and Statistical Methodology Metadata*. Washington DC, Section 3.4.4.

[29] US Federal Committee on Statistical Methodology. 2005. *Statistical Policy Working* Paper 22 (Revised 2005) – *Report on Statistical Disclosure Limitation Methodology*. http://www.fcsm.gov/working-papers/spwp22.html

[30] Watkins, W, et Boyko, E. 1996. *Data Liberation and Academic Freedom*, Government Information in Canada/Information gouvernementale au Canada 3, n°2. http://www.usask.ca/library/gic/v3n2/watkins2/watkins2.html

[31] Watkins, W. *The Data Liberation Initiative: A New Cooperative Model*. Unpublished paper written for the Canadian Social Science Federation. http://library2.usask.ca/gic/v1n2/watkins/watkins.html

20 NDT: document disponible en anglais uniquement. Il existe cependant une version condensée publiée en français par la Commission de statistique des Nations Unies, sous le titre Principes et lignes directrices concernant la gestion de la confidentialité et de l'accès aux microdonnées, accessible à l'adresse suivante : http://unstats.un.org/unsd/statcom/doc07/BG-Microdata-F.pdf La traduction fournie dans le présent guide s'inspire largement de cette version.

Sites Web

Association Africaine pour l'Archivage de Données Statistiques (AASDA)
http://www.aasda.net/home_dev/index.php

Australian Bureau of Statistics (ABS)
http://www.abs.gov.au

American Statistical Society, Privacy, Confidentiality, and Data Security
http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main

Central Statistics Office, Ireland (CSO)
http://www.cso.ie/

Conseil européen des archives de données en sciences sociales (CESSDA)
http://www.cessda.org/

Data.Gov (UK)
http://data.gov.uk

Data.Gov (USA)
http://www.data.gov/

Dépôt des documents de formation de l'IDD (hébergé par l'Ontario Universities Scholars Portal Economic and Social Data Service)
https://ospace.scholarsportal.info/handle/1873/69

Data Documentation Initiative Alliance (DDI)
http://www.ddialliance.org

Department of Census and Statistics, Sri Lanka
http://statistics.sltidc.lk

Dublin Core Metadata Initiative (DCMI)
http://dublincore.org/

Economic and Social Data Service (ESDS)
http://www.esds.ac.uk/aandp/create/research.asp

European Social Survey
http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=78&Itemid=190

Office statistique de l'Union européenne (Eurostat)
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

Institute for Social and Economic Research, The British Household Panel Survey
http://www.iser.essex.ac.uk/ulsc/bhps/.

Réseau International pour les Enquêtes auprès des Ménages (IHSN)
http://www.ihsn.org

Consortium interuniversitaire pour la recherche en sciences politiques et sociales (ICPSR)
http://www.icpsr.umich.edu

Luxembourg Income Study (LIS)
http://www.lisproject.org/

Measure DHS
http://www.measuredhs.com

Michigan Census Research Data Center (MCRDC)
www.isr.umich.edu/src/mcrdc/

National Center for Health Statistics (NCHS), Research Data Center (Etats-Unis)
http://www.cdc.gov/nchs

National Opinion Research Center (NORC) de l'Université de Chicago
www.norc.org/DataEnclave

National Science Foundation (Etats-Unis)
http://www.nsf.gov/index.jsp

Programme des Centres de données de recherche (CDR) de Statistique Canada
www.statcan.gc.ca/rdc-cdr/index-fra.htm

Statistique Canada, Initiative de démocratisation des données (IDD)
http://www.statcan.gc.ca/dli-ild/dli-idd-fra.htm

The Latin American and Caribbean Demographic Centre, CEPAL, Commission de statistique des Nations Unies
http://www.cepal.org.ar/software/icepa8c.html

Division de statistique des Nations Unies
http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp

UK Data Archive (UKDA)
http://www.dataarchive.ac.uk/sharing/metadata.asp
http://securedata.ukda.ac.uk/about/about.asp

UK Statistics Authority
http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html

US Census Bureau
http://www.census.gov/population/www/cen2000/pums/index.html
www.census.gov/srd/sdc/

Wikipedia
http://en.wikipedia.org

A propos de l'IHSN

En février 2004, des représentants de nombreux pays et d'agences de développement se sont réunis à Marrakech, au Maroc, pour tenir la deuxième Table ronde internationale sur la gestion axée sur les résultats en matière de développement. Il s'agissait d'examiner la façon d'améliorer la coordination de l'aide des bailleurs de fonds, afin de renforcer les systèmes statistiques et les capacités de suivi et d'évaluation nécessaires pour que les pays puissent prendre en charge la gestion de leur processus de développement. Cette table ronde a abouti, entre autres, à l'adoption du Plan d'action de Marrakech pour la statistique (MAPS).

La création d'un Réseau International pour les Enquêtes auprès des Ménages est l'une des principales recommandations du MAPS. En la concrétisant, la communauté internationale a reconnu le rôle déterminant des modèles d'enquête dans la planification, la mise en œuvre et le suivi des politiques et des programmes de développement. De plus, elle a fourni aux organisations nationales et internationales une plate-forme permettant de mieux coordonner et gérer la collecte et l'analyse de données socio-économiques et de mobiliser des moyens pour améliorer l'efficacité et l'efficience des approches de réalisation des enquêtes dans les pays en développement.

La série des documents de travail de IHSN vise à favoriser les débats d'idée autour de la conception et de la réalisation d'enquêtes auprès des ménages, ainsi que de l'analyse, de la diffusion et de l'utilisation des données recueillies via ces enquêtes. Pour soumettre un texte en vue de sa publication dans la série des documents de travail d'IHSN, contacter le secrétariat de IHSN à l'adresse suivante: info@ihsn.org.

www.ihsn.org
E-mail: info@ihsn.org