gue 93-03A

Nº de catalo

FICHIERS DE MICRODONNÉES DE L'EDTR
PROPOSITION DE CONTENU
PARTIE A - VUE D'ENSEMBLE

Juin 1993

Jamie Brunet, Division des enquêtes-ménages

Philip Giles, Division des enquêtes-ménages

La série de documents de recherche de l'EDTR est conçue en vue de communiquer les résultats des études ainsi que les décisions importantes ayant trait à l'Enquête sur la dynamique du travail et du revenu. Ils sont offerts gratuitement, en français et en anglais. Pour obtenir une description sommaire des documents disponibles ou un exemplaire de ces documents, communiquez avec Philip Giles, EDTR, par la poste à Édifice Jean-Talon, 11ième étage, section D8, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6; par INTERNET: GILES@STATCAN.CA; par téléphone au (613) 951-2891; ou par télécopieur au (613) 951-3253.

SOMMAIRE

En février 1992, l'équipe chargée de l'élaboration du contenu de l'Enquête sur la dynamique du travail et du revenu (EDTR) avait diffusé auprès d'un grand nombre de lecteurs une proposition de contenu de l'enquête. Ce document a servi de base de discussion lors de consultations poussées, lesquelles ont donné lieu à des modifications du contenu de l'enquête.

Nous prévoyons procéder sensiblement de la même façon pour les produits de l'EDTR. Ce rapport (publié en deux parties, Partie A et Partie B) constitue une première proposition de contenu des fichiers de microdonnées longitudinales de l'EDTR. Les lecteurs sont priés de faire part de leurs commentaires. De plus, toute suggestion d'autres produits de l'EDTR est la bienvenue. Vos commentaires peuvent être adressés au gestionnaire de la série de documents de recherche de l'EDTR dont les coordonnées se trouvent sur la page précédente.

Le présent document, Partie A, fournit une vue d'ensemble de la stratégie proposée pour les fichiers de microdonnées longitudinales de l'EDTR. La Partie B, disponible sur demande, intéressera les personnes ayant besoin de renseignements détaillés sur des diverses variables spécifiques.

Vos suggestions et commentaires peuvent être envoyés en tout temps, mais avant la date limite du 31 octobre 1993. À ce moment, une nouvelle stratégie aura été élaborée et diffusée dans la séries de documents de recherche de l'EDTR.

TABLE DES MATIÈRES

1. APERÇU DE LA DIFFUSION DES RÉSULTATS DE l'EDTR

À ce jour, l'équipe de l'EDTR (Enquête sur la dynamique du travail et du revenu) s'est surtout intéressée au contenu de l'enquête et à la collecte des données. Bien qu'il s'agisse d'aspects importants, plusieurs utilisateurs sont aussi intéressés par nos plans de diffusion. Pour répondre aux besoins des utilisateurs de données, l'équipe de l'EDTR prévoit un large éventail de produits et sollicite des conseils au sujet de leur élaboration. Le présent document représente une première démarche en ce sens. Les premières données de l'EDTR seront diffusées au début de 1995. Tout au long de 1994, les plans de diffusion seront élaborés et perfectionnés.

En attendant que les données soient disponibles, la ligne de produits de l'EDTR comporte deux éléments. Un bulletin d'information, *La Dynamique*, fournit des renseignements à jour sur l'évolution de l'enquête et les questions qui s'y rapportent. Distribué quatre fois par année, ce bulletin vise à tenir les lecteurs informés sur les aspects généraux de l'enquête.

Une série de documents de recherche de l'EDTR existe pour ceux qui sont intéressés à suivre les développements de plus près. Ces documents portent sur des questions concernant la conception de l'enquête, sur l'évaluation de la qualité des données et sur l'étude exploratoire. Le bulletin *La Dynamique* ainsi que les documents de recherche sont offerts gratuitement. Une personne peut s'abonner à l'ensemble de la série ou peut commander un exemplaire des documents qui l'intéressent plus particulièrement. Chaque document de recherche est brièvement décrits dans *La Dynamique*.

Aperçu de la ligne de produits

Les principaux avantages d'une enquête longitudinale sont les possibilités d'analyse des microdonnées, en particulier dans le cas des changements dans le temps de certaines caractéristiques individuelles. Par conséquent, l'établissement de microdonnées et d'une bonne documentation se situe au coeur de nos préoccupations et constitue le principal objet du présent document.

Divers types de fichiers de microdonnées (pour plus de détails, se reporter à la section 4) seront offerts. Si la demande est suffisante, des ateliers à l'intention des utilisateurs des données de l'EDTR seront tenus pour aider à la compréhension et à l'utilisation des données.

La ligne de produits inclura également des publications analytiques sur des sujets permettant d'exploiter les données de l'EDTR. Voici nos prévisions quant au rôle que l'équipe de l'EDTR sera appelée à jouer sur les analyses :

- L'équipe de l'EDTR effectuera des analyses qui seront intégrées au processus de diffusion des données. Bien que la diffusion des données devrait être aussi convenable que possible, elle devrait être accompagnée d'une analyse substantielle. En faire moins, ce serait sous-estimer les données.

- Selon les fonds disponibles, la réalisation d'études axées sur l'application des données ainsi que sur les techniques d'analyse particulièrement adaptées à l'EDTR se fera à contrat. Les résultats seront publiés par Statistique Canada, mais seront pleinement reconnus aux auteurs.

- Nous cherchons des occasions de participer à des projets d'analyse conjointe avec des chercheurs, tant de Statistique Canada que de l'extérieur, car cela semble de plus en plus être un excellent moyen d'obtenir le meilleur des deux mondes.

- L'équipe de l'EDTR participera, peut-être aussi financièrement, à des projets d'analyse utilisant les données de l'EDTR conjointement avec des données d'autres enquêtes longitudinales.

Bien qu'il n'y ait pas encore de plan précis à ce sujet, des publications régulières de données seront produites si certains ensembles de totalisations, présentant un intérêt commun, peuvent être identifiés. La difficulté est de savoir comment mettre en évidence les aspects longitudinaux des données. En effet, le contenu de l'enquête est sensiblement identique à celui d'enquêtes transversales courantes (comme l'Enquête sur la population active et l'Enquête sur les finances des consommateurs), qui ont des programmes de publication des données bien élaborés. L'EDTR a pour objectif de compléter ces enquêtes existantes, et la façon de publier les données pour atteindre ce but n'est pas clairement définie. De plus, le temps consacré à la production de publications réduirait le temps que l'on pourrait consacrer à la production de fichiers de données et de documentation pertinente.

2. OBJET ET PORTÉE DU DOCUMENT

Un document décrivant le contenu proposé de l'EDTR a été diffusé à un vaste public en février 1992 dans le but d'obtenir les commentaires des utilisateurs éventuels de données avant même de procéder à l'élaboration du questionnaire. Le présent document est conçu dans le même esprit : solliciter des conseils sur la façon de parfaire la structure des fichiers de microdonnées avant de mettre au

point des systèmes de production. La partie A fournit une description générale du contenu proposé pour les fichiers de microdonnées longitudinales de l'EDTR; la partie B donne plus de détails sur chacune des variables.

Ce rapport doit être considéré comme un point de départ. Cela signifie, premièrement, que peu de choses sont définitives. Deuxièmement, le document ne traite pas de toute la ligne de produits; les versions futures contiendront les modifications proposées lors de la consultation et soulèveront de nouvelles questions. En raison de l'énormité de la tâche de définition des produits, nous avons jugé essentiel d'adopter ce type de consultation «progressive», plutôt que d'attendre que tous les aspects aient été étudiés avant de les soumettre aux utilisateurs de données.

Le rapport comprend :

- une explication de la structure générale des variables;
- la fréquence de diffusion proposée;
- une description des variables calculées (dans la partie B).

Le rapport ne comprend pas :

- les algorithmes détaillés des variables calculées, qui seront produits ultérieurement;
- le contenu définitif de l'enquête -- le rapport énumère les éléments de contenu actuellement connus;

- un tour complet de la question de la confidentialité, bien qu'une brève discussion soit incluse;

- la méthode à utiliser pour indiquer les changements découlant du traitement des données. Les valeurs imputées seront identifiées dans les fichiers de microdonnées de l'EDTR, mais aucune méthode pour le faire n'a encore été proposée.

3. CONFIDENTIALITÉ

Dans la partie B du document, toutes les variables sont présentées sans avoir été «filtrées». Or, il est évident qu'avant la diffusion de tout fichier de microdonnées à grande diffusion, on procède à la suppression ou au regroupement des valeurs relatives à certaines variables. Statistique Canada reconnaît cependant que certaines analyses ne peuvent être réalisées avec les seules microdonnées «filtrées». Un mécanisme sera conçu pour permettre aux chercheurs d'utiliser le champ complet de microdonnées pour des analyses statistiques, sans compromettre la confidentialité assurée aux répondants. (Un futur document de recherche de l'EDTR abordera cette question.) Bien que les détails sont loin d'être finaux, le scénario envisagé est le suivant :

- Les fichiers de microdonnées à grande diffusion contiendront toutes les variables. Lorsqu'un «filtrage» sera nécessaire pour assurer la confidentialité, une autre technique que la suppression sera utilisé -- par exemple, l'attribution aux personnes de valeurs possibles, en gardant les moyennes dans certaines classes de la population.

- Lorsque les chercheurs le jugeraient utile pour certains travaux de recherches initiaux, il serait également possible d'inclure toutes les données

(c'est-à-dire qu'il n'y aurait pas de suppression ou de regroupement des valeurs) pour une centaine d'individus seulement, qui seraient représentatifs de l'éventail possible des valeurs. Évidemment, le choix de ces personnes devrait se faire avec soin afin d'éviter que les données ne permettent de les identifier.

- Pour les études requérant des données «non filtrées», le chercheur écrirait le code permettant l'extraction et l'analyse des données et le télécommuniquerait à Statistique Canada. Le programme serait exécuté avec les données du fichier interne et les résultats seraient communiqués au chercheur une fois qu'on se serait assuré qu'ils ne contiennent aucune information contrevenant aux dispositions en matière de confidentialité. Autant que possible, le processus entier serait informatisé; plus il serait informatisé, plus les délais de traitement seraient courts et moins le chercheur aurait à défrayer de coûts.

4. FRÉQUENCE DE DIFFUSION DES FICHIERS

L'équipe de l'EDTR entend diffuser deux types de fichiers : celui de données transversales et celui de données longitudinales. Le fichier de données transversales contient de l'information pour une seule année de référence et ressemblera à un fichier «traditionnel» de microdonnées d'enquête. En raison de sa combinaison unique de données sur le travail et le revenu, le fichier de données transversales de l'EDTR sera un produit exceptionnel en soi, car il fournira des données qui ne peuvent actuellement être obtenues d'aucune autre source. Les fichiers de données longitudinales, l'objet principal du présent document, portent sur plus d'une année de référence.

Le tableau 1 illustre les liens qui existent entre les périodes de référence de collecte des données et les périodes de diffusion proposées des fichiers. Un fichier de données transversales sera produit chaque année, fournissant toute l'information de l'EDTR relative à l'année de référence en question. En raison de l'introduction décalée de l'échantillon, les fichiers de données transversales des trois premières années de collecte contiendront des données recueillies auprès d'un seul panel de répondants. À compter de l'année de référence 1996, les données transversales proviendront de deux panels de répondants. (L'échantillon initial de chaque panel comprendra environ 20 000 ménages.)

TABLEAU 1 - Diffusion des fichiers de microdonnées, selon le type

	Année de référence											
	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Panel 1	X	X	X	X	X	X						
Panel 2				X	X	X	X	X	X			
Panel 3							X	X	X	X	X	X
Fichier X	●	●	●	●	●	●	●	●	●	●	●	●
Fichier L3	<=	==	=>	<=	==	=>	<=	==	=>	<=	==	=>
Fichier L6	<=	=*	==	=*	=*	=>						
Fichier L6				<=	==	==	==	==	=>			
Fichier L6							<=	==	==	==	==	=>

Nota : Fichier X = fichier de données transversales; Fichier L3 = fichier de données longitudinales pour trois ans; Fichier L6 = fichier de données longitudinales pour six ans; * = exception à la proposition

À cause de la nature du plan d'échantillonnage (chevauchement des panels) et du renouvellement des panels, les fichiers de données longitudinales sont censés couvrir deux périodes de référence : des fichiers pour trois ans et des fichiers pour six ans. Les fichiers pour trois ans contiendront des données recueillies auprès d'un échantillon commun de deux panels (une exception : le premier portera sur un seul panel en raison de l'introduction graduelle de l'échantillon). Ce large échantillon conviendra pour les plus petits sous-groupes de la population. Cependant, certaines analyses exigent une période de référence plus longue; dans de tels cas, un fichier portant sur une période de six ans serait plus approprié. Le chercheur pourra donc choisir entre un nombre plus grand d'observations et une période de référence plus longue.

Le premier fichier de données de l'EDTR à être disponible sera le fichier de données transversales pour l'année de référence 1993. La période cible qui a été fixée pour la diffusion de ce fichier est le début de 1995.

Les enquêtes longitudinales sont fondamentalement difficiles à traiter -- la cohérence chronologique n'est pas facile à obtenir. Voilà principalement pourquoi nous n'avons pas proposé la diffusion d'un fichier de données longitudinales chaque année. Cependant, comme de nombreux utilisateurs de données ont déjà hâte que les fichiers de données longitudinales de l'EDTR soient diffusées, nous envisageons des diffusions annuelles pour le premier panel seulement. Au delà des plans dont on vient de donner les grandes lignes, cela signifie que des fichiers de données longitudinales pour deux ans, quatre ans et cinq ans seront disponibles. (Ces exceptions sont signalées par un astérisque (*) dans le tableau.)

Diversité des fichiers

Bien que ce document ne fait mention que d'un seul type de fichier de données longitudinales, une gamme d'autres produits pourraient être offerts. Par exemple, un fichier sommaire de données longitudinales, conçu pour des totalisations simples, constitue une possibilité de produit. Ce fichier pourrait faire ressortir les variables calculées qui se caractérisent par des changements dans le temps, mais supprimer les détails dans les longues explications et la synchronisation. Un produit de ce type pourrait être un point de départ pour un chercheur qui commence à utiliser les données de l'EDTR. Un prochain document de recherche de l'EDTR explorera cette possibilité plus en détail.

5. ORGANISATION DES FICHIERS DE MICRODONNÉES LONGITUDINALES DE L'EDTR

L'équipe de l'EDTR a posé comme hypothèse que chaque utilisateur aimerait créer des fichiers de travail qui contiendrait uniquement les variables et les observations présentant un intérêt pour son projet d'analyse. Pour cette raison, un logiciel d'extraction de données sera joint à tous les fichiers de microdonnées. Cette façon de procéder garantira un stockage plus efficace des données, et il sera facile pour les utilisateurs de sélectionner les variables et les sous-ensembles d'observations qui les intéressent. Cela permettra également aux utilisateurs de choisir facilement l'unité d'observation la plus appropriée : personne, groupes de personnes habitant ensemble (famille, ménage), employeurs (l'EDTR demandera aux répondants certains renseignements au sujet de chacun de leurs employeurs).

Il reste encore à décider du format de sortie des résultats du logiciel d'extraction de données. Un format rectangulaire comportant des enregistrements de longueur fixe, et chaque variable dans une position donnée, pourrait être le format par

défaut. Il s'agit d'un format courant de présentation des données, applicable à la plupart des logiciels d'analyse. À plus long terme, il devrait être possible de créer directement des fichiers dans le format propre au logiciel utilisé. Par exemple, deux logiciels d'analyses statistiques bien connus, le SAS et le PSSS, requièrent des fichiers de données dans un format spécifique à leur logiciel. Il serait donc utile d'offrir l'option de créer directement un fichier SAS ou un fichier SPSS. À noter que ce ne sont que des exemples et que les logiciels pour lesquels il serait possible de créer directement des fichiers seraient déterminés en fonction des besoins exprimés par les utilisateurs.

Les fichiers de sortie seront fournis sur le type de support requis : CD-ROM, disquette, bande magnétique. Il est possible cependant que les coûts varient selon le type de support d'information.

6. STRUCTURE DES VARIABLES

Bien que le fait d'offrir un logiciel d'extraction de données rende inutile un cliché d'enregistrement détaillé, les utilisateurs doivent comprendre comment les données sont présentées. Le principe de base à retenir est de présenter les données comme si elles avaient été recueillies une fois, à la fin de la période de référence couverte par le fichier (par exemple, trois ans ou six ans). Un processus conceptuel (pour ne pas dire réel) en vue d'atteindre cet objectif est expliqué dans les paragraphes qui suivent.

Les données de l'EDTR sont recueillies deux fois par année. L'interview de janvier traite des activités sur le marché du travail de l'année précédente. L'interview de mai recueille les données sur le revenu pour l'année précédente. Si on regroupait simplement toutes les données telles qu'elles sont recueillies, on obtiendrait un résultat semblable à celui qui est présenté dans la figure 1. Une observation

correspondrait à toutes les données recueillies au sujet d'une personne en particulier pendant une année donnée.

Figure 1 : Présentation des données selon la collecte

PERSONNE 1 DONNÉES POUR L'ANNÉE 1
PERSONNE 2 DONNÉES POUR L'ANNÉE 1
●
●
●
PERSONNE N DONNÉES POUR L'ANNÉE 1
PERSONNE 1 DONNÉES POUR L'ANNÉE 2
PERSONNE 2 DONNÉES POUR L'ANNÉE 2
●
●
●
PERSONNE N DONNÉES POUR L'ANNÉE 2
●
●
●
PERSONNE N DONNÉES POUR L'ANNÉE 6

Étant donné l'importance de l'analyse longitudinale, il serait peu commode d'utiliser les données conformément à la présentation de la figure 1 pour analyser les changements dans le temps à un niveau individuel. L'étape suivante consiste donc (encore une fois en terme d'une évolution conceptuelle) à regrouper toutes les données qui concernent une personne, de façon telle qu'illustré dans la figure 2.

Figure 2 : Présentation des données - une observation par personne

PERSONNE 1 DONNÉES POUR L'ANNÉE 1 ... POUR L'ANNÉE 6
PERSONNE 2 DONNÉES POUR L'ANNÉE 1 ... POUR L'ANNÉE 6
●
●
●
PERSONNE N DONNÉES POUR L'ANNÉE 1 ... POUR L'ANNÉE 6

Il faut ensuite procéder à une restructuration des variables relatives à chacune des observations de la figure 2. Le nouveau classement des données, regroupées par variable, est illustré à la figure 3.

Figure 3 : Présentation des données selon l'ordre des variables plutôt qu'en ordre chronologique

PERSONNE 1 (VAR 1 A1)...(VAR 1 A6)...(VAR M A1)...(VAR M A6)
PERSONNE 2 (VAR 1 A1)...(VAR 1 A6)...(VAR M A1)...(VAR M A6)
⋮
PERSONNE N (VAR 1 A1)...(VAR 1 A6)...(VAR M A1)...(VAR M A6)

À de nombreux égards, il n'existe pas de différences significatives entre les figures 1, 2 et 3, et en particulier entre les figures 2 et 3, pour ce qui est de la présentation des données. Il ne s'agit que d'un mode de classement différent. Certaines redondances existeraient et pourraient être éliminées. Par exemple, la date de naissance n'aurait pas à être répétée pour chacune des années de référence dans le fichier de sortie. Un tel classement affecterait le degré de facilité ou de difficulté pour un utilisateur lors de la manipulation du fichier.

La prochaine étape consiste à supprimer les «interruptions de données artificielles» dues à la façon dont les données sont recueillies. Par exemple, dans l'EDTR, les répondants doivent indiquer les dates de début et de fin d'emploi pour chacun de leurs employeurs. Supposons qu'un répondant ait travaillé pour un employeur donné du 12 juillet 1993 au 5 novembre 1995, les données telles que recueillies se présentent de la manière suivante :

Période de référence	Date de début	Date de fin
1993	12 juillet	En cours à la fin de l'année
1994	Même emploi que l'année précédente	En cours à la fin de l'année
1995	Même emploi que l'année précédente	5 novembre

Dans un fichier longitudinal portant sur les années de référence 1993, 1994 et 1995, l'information sur les dates de début et de fin d'emploi pour cet employeur peut être plus efficace lorsqu'elle est présentée de la façon suivante : date de début = 12 juillet 1993 et date de fin = 5 novembre 1995. Il n'y a aucune perte d'information.

Variables fixes et variables dynamiques

Les variables de l'EDTR seront ou bien recueillies directement auprès du répondant ou bien calculées. Qu'elles soient directes ou calculées, les variables peuvent être de nature «fixe» ou «dynamique». Une variable fixe est une variable qui ne change pas avec le temps, par exemple la date de naissance. Une variable dynamique peut changer pendant la période où le répondant fait partie de l'échantillon de l'EDTR, par exemple son état matrimonial.

Bien sûr, la valeur des variables dynamiques ne changera pas pour tous les répondants. Ainsi, l'état matrimonial de certains répondants de l'EDTR changera au cours de la période de référence et ne changera pas pour d'autres.

Pour les variables fixes, des valeurs numériques seront (habituellement) attribuées aux diverses réponses possibles.

Les valeurs des variables dynamiques seront présentées d'une manière identique -- la difficulté qui se pose est de trouver une façon simple mais efficace de représenter les changements de valeur. Nous proposons la façon suivante :

$$X_1 \ D_1 \ X_2 \ D_2 \ \ldots \ X_T \ D_T \ X_{T+1}$$

où T représente le nombre de changements de valeur (défini séparément pour chaque variable dynamique et chaque répondant),
$X_1, \ldots, X_{T+1}$ sont les valeurs définies de la variable,
D_i est la date du changement de la valeur X_i à la valeur X_{i+1}
$(i = 1, \ldots, T)$.

Compte tenu de la précision attendue de la part des répondants, la date du changement reflétera le niveau de détails amassés pour l'enquête. Les possibilités sont :

- Année/mois
- Année/semaine
- Année/mois/journée

La deuxième possibilité (année/semaine) dérive de la dernière (année/mois/jour). On appliquera un concept selon lequel l'année est divisée en 53 semaines. Une semaine est définie comme allant du dimanche au samedi, inclusivement. La semaine 1, qui peut comprendre moins de sept jours, est la période qui va du 1er janvier au premier samedi du mois de janvier. La semaine 53 peut aussi compter moins de sept jours et s'étend du dernier dimanche de décembre au 31 décembre. Selon cette définition, toutes les années ont exactement 53 semaines. Le concept de l'année de 53 semaines sera appliqué pour les variables définies pour chacune des semaines de la période de référence et pour les données relatives à des dates présentées selon le modèle année/mois/jour, lorsqu'on estimera que le jour déclaré représente une estimation fiable de la semaine du mois en question, mais non du jour exact.

Nous pourrions apporter quelques modifications à cette façon de procéder, comme :

- Au lieu de numéroter de 1 à 53 les semaines de chacune des années, nous pourrions recourir à une numérotation consécutive. Ainsi, la première semaine de la deuxième année serait la semaine 54 et ainsi de suite. Dans ce scénario, la seule différence est la convention utilisée pour la numérotation.
- En supposant que la séparation des années civiles n'est pas de la première importance, le même concept pourrait être appliqué à l'ensemble de la période de référence. Cette fois-ci, la première et la dernière semaine de la période de référence seraient les seules à pouvoir compter moins de sept jours, et certaines semaines empiéteraient sur une autre année civile. Cette méthode simplifierait les calculs de durée, puisque les semaines seraient numérotées de 1 à 313 pour un fichier de six ans.

Malgré les difficultés que pose la dernière option en raison du chevauchement des années civiles, si on accorde beaucoup d'importance aux valeurs de durée, il serait possible de fournir l'équation permettant de déterminer de telles valeurs.

La figure 4 donne un exemple de présentation d'une variable dynamique. Cet exemple est fondé sur la variable «situation vis-à-vis de l'activité» à laquelle cinq valeurs peuvent être associées. L'inclusion de la valeur 9 (personne hors de l'échantillon de l'EDTR) illustre un autre aspect de la présentation des données. L'EDTR recueillera de l'information sur les «nouveaux membres», c'est-à-dire toutes les personnes habitant avec un répondant de l'EDTR qui ne faisaient pas partie du ménage au moment de l'introduction du panel des répondants. Donnons l'exemple d'une personne qui vivait chez ses parents au moment où le ménage a été choisi pour faire partie de l'échantillon de l'EDTR. Par la suite, cette personne a

quitté la maison familiale et s'est mariée. Son conjoint est ce que nous appelons un «nouveau membre».

Figure 4 - Exemple de présentation d'une variable dynamique

Situation vis-à-vis de l'activité : les valeurs valides sont 1 = occupé(e) /
2 = chômeur(euse) / 3 = inactif(ive) / 8 = non connue /
9 = personne hors de l'échantillon de l'EDTR
Fichier longitudinal pour six ans couvrant la période de référence 1993-1998

9	**9**	**4**	**0**	**1**	**1**	**9**	**6**	**3**	**1**	**2**	**9**	**8**	**1**	**7**	**3**	
X_1	A_1		S_1		X_2	A_2		S_2		X_3	A_3		S_3		X_4	

Dans l'exemple illustré à la figure 4, cette personne est un nouveau membre, dont des données ont été recueillies pour la première fois dans l'EDTR en janvier 1995. Il n'y a pas d'information pour l'année de référence 1993. Dans la première semaine de 1994 (A_1 = 94, S_1 = 01), la personne était occupée (X_2 = 1). Une entrevue subséquente a permis de déterminer que dans la semaine 31 de 1996 (A_2 = 96, S_2 = 31), la personne est devenue chômeuse (X_3 = 2). Un autre changement a été noté dans la semaine 17 de 1998 (A_3 = 98, S_3 = 17), la personne étant devenue inactive (X_4 = 3). Aucun autre changement n'est signalé, ce qui signifie que la personne était toujours inactive à la fin de la période de référence de l'enquête (dans notre exemple, le 31 décembre 1998).

Bien qu'elles soient dynamiques de par leur nature, la plupart des valeurs relatives au revenu seront représentées par :

$$X_1 \ X_2 \ X_3 \ X_4 \ X_5 \ X_6$$

Cela tient au fait que nous ne recueillons pas de données sur les dates des changements associées à ces variables. De fait, pour des variables comme les sources de revenus, la période de référence est l'année, et aucune information n'est demandée à un niveau infra-annuel.

7. CONTENU DES FICHIERS DE MICRODONNÉES LONGITUDINALES DE L'EDTR

Une description détaillée de toutes les variables est fournie dans la partie B de ce document. En principe, toutes les données recueillies au sujet de tous les répondants seront accessibles à l'utilisateur de données grâce à un mécanisme quelconque permettant un accès total, mais protégeant la confidentialité des individus, comme mentionné dans la section 3. Les variables calculées seront calculées et fournies pour les cas «typiques».

Les renseignements qui suivent seront accessibles à partir du fichier ou du logiciel d'extraction pour permettre la «généralisation» de programmes d'utilisateurs pour tout fichier longitudinal de l'EDTR :

- Nombre d'années de référence couvertes par le fichier
- Première année de la période de référence du fichier
- Dernière année de la période de référence du fichier
- Dates de la première et dernière semaine de chaque année de référence
- Nombre de jours dans les semaines 1 et 53 de chaque année de référence
- Nombre total de personnes visées par le fichier
- Nombre de personnes âgées de 15 ans et moins, de 16 ans à 69 ans, ainsi que de 70 ans et plus visées par le fichier

- Nombre d'employeurs-personnes compris dans le fichier (c.-à-d. la somme du nombre d'employeurs de toutes les personnes pendant la période de référence)
- Nombre de ménages au 1er janvier de chaque année
- Nombre de familles économiques au 1er janvier de chaque année
- Nombre de familles du recensement au 1er janvier de chaque année
- Pour chaque variable dynamique, le nombre maximal de changements enregistrés pour toute personne (ou employeur, ou ménage, ou famille, le cas échéant) -- il s'agit de la valeur maximale que peut prendre T selon le modèle défini dans la présentation des variables dynamiques.

Les catégories de variables suivantes sont décrites en détail dans la partie B :

- Renseignements généraux au niveau de la personne
- Données démographiques
- Niveau d'instruction
- Antécédents professionnels
- Renseignements généraux relatifs à l'activité sur le marché du travail
- Caractéristiques des emplois pour chaque employeur
- Épisodes sans emploi
- Incapacité
- Soins
- Revenu
- Patrimoine
- Renseignements relatifs aux autres membres du ménage
- Renseignements relatifs à la famille et au ménage

On trouve aussi des renseignements sur le contenu de l'enquête dans les documents de recherche de l'EDTR nº 92-01A «Contenu de l'Enquête sur la dynamique du travail et du revenu : Partie A - Données démographiques et données relatives à l'activité sur le marché du travail» et nº 92-01B «Contenu de l'Enquête sur la dynamique du travail et du revenu : Partie B - Données sur le revenu et le patrimoine». Pour plus de détails sur les questions proprement dites de l'enquête, élaborées en vue de l'essai sur le terrain de 1993, se reporter aux documents de recherche de l'EDTR nº 93-02 «Le "questionnaire" de la collecte des données sur le travail de l'EDTR - janvier 1993» et nº 93-04 «Le questionnaire et les procédures de collecte des données sur le revenu de l'EDTR - mai 1993».

Comme elles se situent hors du sujet tant du contenu de l'enquête que des variables dérivées, nous décrivons brièvement ici deux des catégories de la liste qui précède :

- Renseignements généraux au niveau de la personne
  - Identificateur de la personne
  - Données géographiques sur le lieu de résidence
  - Date de l'interview préliminaire (c'est-à-dire date à laquelle la personne est entrée dans l'échantillon de l'EDTR)
  - Poids d'échantillonnage (pour que les valeurs fournies par l'échantillon soient représentatives de l'ensemble de la population)
  - Renseignements recueillis ou non par personne interposée (pour chaque interview, identification de la personne ayant fourni les renseignements et, le cas échéant, lien entre le répondant substitut et la personne)
  - langue d'interview (pour chaque interview)

- Renseignements sur les autres membres du ménage

  Cette section fournit des renseignements sur toutes les autres personnes qui faisant partie du même ménage que le répondant à un moment ou à un autre pendant la période de référence du fichier. Les variables suivantes sont fournies au sujet de ces autres membres (certaines de ces variables sont dynamiques) :

  - Identificateur de l'autre membre
  - Lien entre le membre du ménage et le répondant
  - Situation des particuliers dans la famille
  - Date de naissance
  - Sexe

Les fichiers de données d'enquête contiennent des «poids d'échantillonnage» à utiliser pour que les valeurs associées aux personnes faisant partie de l'échantillon soient représentatives de l'ensemble de la population. La pondération est une opération plus complexe dans le cas des enquêtes longitudinales, comparativement aux enquêtes transversales, étant donné que la population évolue dans le temps : des personnes y entrent et en sortent (lorsque leur lieu de résidence change), des personnes naissent et d'autres meurent.

On prévoit inclure dans les fichiers longitudinaux toutes les personnes auprès desquelles des données ont été recueillies au cours de la période de référence (même si elles n'ont pas fait partie de l'échantillon pendant toute la période). Un poids longitudinal et un poids transversal sera attribué à chaque personne pour chaque année de référence. Les poids d'échantillonnage d'une année donnée seront fondés sur la population dénombrée au 1er janvier de l'année en question. Par conséquent, certains répondants auront un poids égal à zéro, s'ils ne faisaient pas partie de l'échantillon pendant une année donnée. Les utilisateurs devront donc posséder une connaissance de base des méthodes de pondération pour pouvoir

utiliser les poids de manière adéquate dans leurs analyses. La question de la pondération sera traitée en détail dans un futur document de recherche de l'EDTR.

8. VARIABLES CALCULÉES

Contrairement aux variables directes, les variables calculées sont fondées sur les renseignements recueillis auprès des répondants. Les variables calculées proposées sont énumérées dans la partie B de ce rapport. Nous apprécierions que vous nous fassiez part de vos commentaires au sujet de l'utilité de telles variables ou que vous en proposiez de nouvelles.

La présente partie aborde des questions de nature plus générale, c'est-à-dire qui ne se rapportent pas à certaines variables calculées en particulier.

Le point le plus important concerne les calculs à faire dans le cas des changements de composition des membres d'une famille ou d'un ménage. Par exemple, comment devrait-on calculer le revenu familial des familles dont la composition a changé au cours de l'année?

- Le revenu familial pourrait être calculé uniquement pour les familles dont la composition est restée la même toute l'année;
- Le revenu familial pourrait être défini pour chaque personne comme la somme des revenus gagnés par tous les membres de la famille dénombrés à un moment quelconque de l'année lorsque cette personne faisait partie du ménage.

Évidemment, la première solution est la plus simple, tant pour ce qui est des calculs que de la facilité de compréhension. Il est clair également que la deuxième

possibilité offre une mesure supérieure pour les analyses au niveau des personnes où le revenu familial constitue une variable explicative.

Les calculs associés à la deuxième option posent cependant de grandes difficultés compte tenu du fait qu'on ne recueille pas directement de données infra-annuelles sur le revenu. D'autres renseignements fournis directement par les répondants peuvent faciliter les calculs, telles les dates d'emploi et la période pendant laquelle une personne a touché des prestations de programmes comme l'assurance-chômage et le bien-être social, mais on ne peut faire de calculs précis à un niveau infra-annuel.

Un dernier aspect du calcul des variables que nous aimerions souligner a trait à la non-réponse. Si une variable est calculée à partir des renseignements fournis directement de trois autres variables, et que, pour une personne en particulier, l'une des variables d'entrée est manquante, comment la variable de sortie devrait-elle être calculée?

9. QUESTIONS AUX UTILISATEURS ÉVENTUELS DES DONNÉES

Nous comptons sur vos commentaires au sujet de n'importe lequel des aspects soulevés dans ce document et la liste qui suit de questions d'orientation stratégique essentielles n'est présentée qu'à titre de guide pour les personnes qui souhaiteraient y répondre.

1. La disponibilité du fichier proposée répond-elle à vos besoins? Vous intéressez-vous surtout aux fichiers longitudinaux, aux fichiers transversaux ou aux deux?

2. Les autres fichiers longitudinaux qui pourraient être produits vous seraient-ils (potentiellement) utiles? Dans l'affirmative, que devraient-ils contenir?

3. Que pensez-vous de la forme de présentation proposée des variables dynamiques?

4. Êtes-vous d'accord avec le mode de présentation des dates et, en particulier, avec le concept de l'année de 53 semaines?

5. Êtes-vous favorable au projet de logiciel d'extraction de données permettant de créer un fichier de travail pour des applications particulières? De quels genres de fichiers de travail auriez-vous besoin?

6. Quels logiciels pensez-vous utiliser pour analyser les données de l'EDTR? S'il y en a plus d'un, lequel est le plus important?

7. Quels types d'analyses avez-vous l'intention d'effectuer avec les données de l'EDTR; par exemple, des tableaux croisés, des régressions multiples, des analyses chronologiques?

8. De quel genre de matériel vous servirez-vous pour analyser les données de l'EDTR? S'il vous est difficile de déterminer ce que vous utiliserez dans deux ou trois ans, dites-nous ce que vous utiliseriez si les données étaient accessibles maintenant?

9. Sous quelle forme préféreriez-vous recevoir les fichiers de données de l'EDTR? Comme pour la question précédente, si vous ne pouvez faire de prévisions, quelle serait votre choix aujourd'hui?

10. Que pensez-vous de la question de la confidentialité et de la solution proposée pour l'accès aux données? Si vous êtes d'accord dans l'ensemble avec le principe, quels genres de fichiers ou de données d'essai vous seraient le plus utiles?

11. En tenant compte de l'utilisation que vous prévoyez faire des données, avez-vous des suggestions concernant les variables calculées dans le cas d'un changement de composition des familles ou des ménages?

12. Cette méthode de consultation est-elle utile? Comment pourrait-elle être améliorée?

13. Dans quelle mesure des ateliers à l'intention des utilisateurs seraient utiles? Quel devrait être leur contenu? Comment devraient-ils être présentés?

14. Quelles priorités relatives devraient être accordées aux divers produits de l'EDTR?

15. Avez-vous d'autres commentaires qui n'ont pas été soulevés par les autres questions?